
SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
MASTER THESIS



CutOver : A Novel Joint Data Augmentation Method for Image Captioning Systems

submitted by
LAVANYA GOVINDARAJU
Saarbrücken
January 2024

Advisor:

Aliki Anagnostopoulou
German Research Center for Artificial Intelligence
Marie-Curie-Str. 1
Oldenburg, Germany

Reviewer 1: : Prof. Dr.-Ing. Daniel Sonntag

German Research Center for Artificial Intelligence
Saarland Informatics Campus
Saarbrücken, Germany

Reviewer 2: Prof. Dr. Antonio Krüger

German Research Center for Artificial Intelligence
Saarland Informatics Campus
Saarbrücken, Germany

Submitted

25, January 2024

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Declarations

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, _____

(Datum/Date)

(Unterschrift/Signature)

Acknowledgements

I extend my sincere gratitude to my advisor, Aliko Anagnostopoulou (M.Sc.), for the unwavering guidance provided throughout the course of my thesis work. Her invaluable advice, continual support, patience, and encouragement played a pivotal role from the initial stages of setbacks to the triumphant completion of my thesis. The journey has been enriched through her mentorship.

I am also grateful to Prof. Dr.-Ing. Daniel Sonntag and Prof. Dr. Antonio Krüger for affording me the opportunity to undertake this thesis project. Their support and the conducive environment provided at Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) contributed significantly to the successful completion of my research.

Lastly, I want to express my deepest appreciation to my beloved parents and my supportive friends. Their unwavering encouragement and unconditional support have been my pillars of strength throughout my academic journey. Their presence has made a profound difference, turning challenges into opportunities and accomplishments.

Saarbrücken, January 2024, Lavanya Govindaraju

Abstract

Image Captioning is a challenging task involving the generation of textual descriptions for images, requiring a seamless integration of computer vision and natural language processing. Current image captioning models heavily rely on extensive training with large-scale image-text datasets, demanding significant computational resources. However, applying data augmentation to vision-language learning, where images and captions are intricately linked, presents challenges. This study introduces "*CutOver*," a novel joint data augmentation method for image captioning systems, combining CutMix from computer vision and instance crossover augmentation from natural language processing to preserve the semantic relationship between images and text during transformation.

Acknowledging the completion of all conducted experiments and the observed outcome that the CutOver method did not yield the anticipated improvement in image captioning system performance, this thesis recognizes the misalignment with initial expectations. A detailed description, analysis, and insights into the factors contributing to the lack of success with the proposed data augmentation approach are provided. The comprehensive examination of the limitations and challenges faced in implementing CutOver contributes valuable knowledge to the understanding of its effectiveness within the specific context of image captioning systems.

Abbreviations

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

BLEU Bilingual Evaluation Understudy

BP Brevity Penalty

CV Computer Vision

CIDEr Consensus-based Image Description Evaluation

ConAug Contextual Augmentation

CNN Convolutional Neural Networks

DA Data Augmentation

DETR Data-efficient Object Transformer

EDA Easy Data Augmentation

Fast R-CNN Fast Region-Based Convolutional Neural Networks

Faster R-CNN Faster Region-Based Convolutional Neural Networks

GRU Gated Recurrent Unit

GPT Generative Pre-trained Transformer

IC Image Captioning

FNN Feedforward neural networks

LM Language Model

LSTM Long Short-Term Memory

ML Machine Learning

MSE Mean squared error

M2 Meshed-Memory Transformer

METEOR Metric for Evaluation of Translation with Explicit Ordering

NLP Natural Language Processing

ROUGE-L Recall-Oriented Understudy for Gisting Evaluation- Longest Common Sub-sequence

RNN Recurrent Neural Networks

R-CNN Region-Based Convolutional Neural Networks

ROI Region of Interest

RPN Region Proposal Network

SAT Show, Attend, and Tell

SSD Single Shot MultiBox Detector

SPICE Semantic Propositional Image Caption Evaluation

SOTA State-of-the-Art

SGD Stochastic Gradient Descent

TF-IDF Term Frequency-Inverse Document Frequency

T5 Text-to-Text Transfer Transformer

TER Translation Edit Rate

YOLO You Only Look Once

List of Figures

1	Impact of Image Augmentation on Caption Consistency	2
2	A basic architecture of Show, Attend and Tell	9
3	Meshed Memory Transformer architecture	10
4	The model architecture of mPLUG	11
5	Caption generation with augmented visual attention	13
6	Different image augmentation techniques	16
7	A visual comparison of Mixup, Cutout, CutMix, and Attentive CutMix . . .	17
8	The anatomy of a deep neural network	20
9	A convolutional neural network architecture	22
10	LSTM cell architecture	24
11	The architecture of R-CNN	28
12	The architecture of Fast R-CNN	28
13	An illustration of Faster R-CNN model	30
14	Examples of soft and hard attentions	34
15	The transformer – model architecture	36
16	Multi-head attention	38
17	Scaled dot-product attention	39
18	Joint Data Augmentation Example	44
19	CutOver pipeline	45
20	CutOver examples	48
21	Examples of MS COCO dataset	52
22	Examples of VizWiz dataset	54

List of Tables

1	Examples of text DA techniques	19
2	Performance comparison of SOTA VizWiz, VizWiz (without DA), and VizWiz (With CutOver DA)	60
3	Performance comparison of CutOver augmentation vs image augmentation methods (Blur, RandomBrightnessContrast, CoarseDropout)	60
4	Performance comparison of CutOver augmentation vs text augmentation methods (bert substitute, random swap, random delete, and synonym replacement)	61
5	Performance comparison of CutOver augmentation vs joint augmentation methods (blur + random swap)	62
6	Performance scores on n% of VizWiz dataset without augmentation	63
7	Performance Scores on n% of VizWiz Dataset with CutOver Augmentation	64
8	Comparison of groundtruth and generated captions (without augmentation and with CutOver augmentation)	67
9	Occurrences of phrases in training dataset and generated captions	68
10	Comparison of groundtruth captions and CutOver captions	69

Contents

1	Introduction	1
1.1	Motivation and Problem Statement	1
1.1.1	Challenges and opportunities in joint DA for IC	3
1.1.2	The road ahead: investigating the potential of CutOver	4
1.1.3	Unlocking the potential of joint DA	4
1.2	Approach and Contribution	5
1.2.1	Approach: Developing CutOver and assessing its impact	5
1.2.2	Contribution: Gaining insights into challenges and iterating on solutions	6
1.3	Thesis outline	6
2	Related Work	8
2.1	Image Captioning	8
2.1.1	Architectures	9
2.1.2	Interactive image captioning	12
2.2	Datasets	15
2.3	Data augmentation	15
2.3.1	Image augmentation	16
2.3.2	Text augmentation	18
3	Technical Background	20
3.1	Deep learning	20
3.1.1	Convolution neural networks	21
3.1.2	Long short-term memory	24
3.2	Image captioning	25
3.3	Data augmentation	25
3.4	Object detection	27
3.4.1	Region-based convolutional neural networks	27
3.4.2	Fast region-based convolutional neural networks	28
3.4.3	Faster region-based convolutional neural networks	29
3.5	Attention mechanism	31
3.5.1	Soft attention	33
3.5.2	Hard attention	34
3.5.3	Transformers	35
4	Methodology	42
4.1	Benchmark Architecture	42
4.2	Choice of Object Detector : Faster R-CNN	43
4.3	Datasets	43
4.4	CutOver	43

4.4.1	CutOver description	44
4.4.2	CutOver Pipeline	44
4.4.3	CutOver examples	46
4.5	Implementation details	49
4.5.1	Dataset preparation	49
4.5.2	Inputs to the model	49
4.5.3	Data pipeline	50
4.5.4	Encoder	50
4.5.5	Model hyperparameters	50
4.5.6	Attention mechanism	50
4.5.7	Decoder	51
4.5.8	Model training	51
5	Experiments and Results	52
5.1	Data description	52
5.1.1	MS COCO dataset overview	52
5.1.2	VizWiz dataset overview	54
5.2	Evaluation metrics	55
5.2.1	BLEU metric	55
5.2.2	ROUGE metric	56
5.2.3	METEOR metric	57
5.2.4	CIDEr metric	58
5.2.5	SPICE metric	58
5.3	Evaluation	59
5.3.1	Experiments	59
6	Discussion and Future Works	66
6.1	Discussion	66
6.1.1	Analysis	68
6.1.2	Insights	70
6.2	Future Works	71
7	Conclusion	74
	Bibliography	77

Chapter 1

Introduction

1.1 Motivation and Problem Statement

In recent times, there has been a significant surge in research focusing on vision-language integration. Image captioning, as a vital task within the realm of multimodal AI, has garnered substantial attention due to its potential applications in fields such as content generation, accessibility, and human-computer interaction. This task involves generating descriptive textual captions based on the visual content of images, essentially bridging the gap between the visual and textual modalities [1, 2, 3, 4, 5, 6, 7, 8].

The goal is to enable machines not only to perceive the content of images but also to express it in human-like language. This synthesis of CV and NLP has the potential to transform the way we interact with and understand visual data. At the heart of IC lies the development of sophisticated models capable of understanding the visual context of an image and generating coherent, contextually relevant textual descriptions. These models, often based on deep learning architectures, have shown remarkable performance in recent years, thanks in part to the availability of large-scale image-text datasets and advances in neural network architectures.

However, achieving such levels of performance often comes at a cost - the need for vast computational resources. Training state-of-the-art IC models, which are capable of generating high-quality, human-like captions, requires extensive access to GPUs and substantial computational power. One critical aspect that has contributed to the success of deep learning models in various domains is the use of DA techniques.

The motivation for this thesis is grounded in the inherent challenges associated with the resource-intensive nature of training IC models, which traditionally rely on large image-text pairs for effective learning. This conventional training process demands significant computational resources, presenting a hurdle in resource-limited environments where access to such resources may be constrained. This issue is particularly emphasized in the paper [9], which underscores the need for interactive models, especially in scenarios where abundant annotated data is not readily available. The conventional method of offline training for IC relies on extensive annotated data, a requirement that

becomes impractical, especially for user-specific images lacking large-scale annotations. To address the inefficiency arising from the scarcity of annotated data and the computational demands, DA emerges as a crucial technique. DA allows for the generation of diverse training instances from limited data, effectively enhancing the model’s ability to generalize and perform well on various inputs.

The primary objective of this thesis is to respond to the challenge of data efficiency in resource-constrained environments by developing precise and efficient DA techniques explicitly tailored for IC. The aim is to make IC more accessible and viable in a broader range of applications, especially those operating under constraints in terms of computational resources and annotated data availability. By advancing DA techniques, the thesis seeks to enhance the efficiency and applicability of IC models, ultimately contributing to the broader accessibility of this technology across different contexts and scenarios.

Data augmentation serves multiple purposes in deep learning, including increasing the efficiency of training and acting as a form of regularization in various domains, including CV [10, 11, 12, 13, 14, 15, 16] and NLP [17, 18, 19, 20, 21, 22, 23, 24]. In CV, DA methods such as random cropping, rotation, and flipping have been widely employed to improve the performance of image classification and object detection models [25]. These techniques help models generalize better to unseen data by introducing variability during training. However, when it comes to vision-language tasks like IC, applying conventional DA methods is not straightforward. Unlike in single modal tasks where data augmentation can be performed independently on either the images or the text, in IC, both modalities are intrinsically linked. Images and their corresponding textual descriptions provide complementary information, and any transformation applied to one modality must ensure that the semantic coherence between the two is preserved.

This unique challenge has motivated researchers to explore novel DA strategies tailored explicitly for vision-language tasks. The objective is to diversify the training data while maintaining the strong semantic relationships between images and their associated captions. These efforts are essential not only to improve the performance of IC models but also to ensure that these models can effectively assist in applications like assisting the visually impaired, generating rich textual descriptions of visual content, and enhancing human-computer interaction.



A boy is sitting on the **left** side of the **grey** sofa.



A boy is sitting on the **left** side of the **grey** sofa.

Figure 1: Impact of Image Augmentation on Caption Consistency. *Left:* ‘A boy is sitting on the **left** side of the **grey** sofa.’ *Right:* Augmentation shifts boy to the **right** and changes sofa color to **green**, yet the caption remains unchanged. Illustrates the limitations of monomodal approaches in adapting to augmented visual content.

One promising avenue for addressing this challenge is the development of joint DA techniques that operate simultaneously on both images and their textual descriptions. In monomodal approaches, IC models concentrate primarily on a single modality, either exclusively analyzing the visual content of images or solely processing the textual descriptions associated with them. An illustrative example from Figure 1 showcases a caption describing a scene where *"a boy sits on the left side of a grey sofa"*, while the applied image augmentation involves flipping and color contrast. Interestingly, despite alterations in the image, the caption remains unchanged, highlighting the limitation of monomodal approaches. Even when the boy's position shifts to the right, and the sofa color changes to green, the caption persists as *"a boy is sitting on the left side of the grey sofa."*

Monomodal strategies inherently face constraints due to the restricted integration of visual and textual information. The isolation of these modalities often leads to missed opportunities to capture a richer context and achieve a deeper understanding of the content within images. The thesis's central focus is to address these limitations and pioneer joint approaches in the field of IC.

The research initiative seeks to bridge the gap between visual and textual modalities, striving to enhance the synergy between them to provide more comprehensive and contextually relevant image captions. By exploring joint approaches, the thesis aims to overcome the inherent drawbacks of monomodal systems, ultimately working towards improved performance and usability in real-world applications. The objective is to push beyond the limitations of monomodal paradigms and pave the way for more sophisticated and integrated IC systems that better reflect the complexity and richness of visual content.

1.1.1 Challenges and opportunities in joint DA for IC

As the demand for sophisticated IC systems continues to grow, there is a pressing need for novel approaches to DA that can effectively address the unique challenges posed by multimodal tasks. The primary challenge in developing joint DA techniques for IC is to strike the right balance between introducing diversity into the training data and preserving the semantic consistency between images and text. This balance is crucial to ensure that the augmented data can effectively improve model performance while avoiding the generation of incoherent or nonsensical captions.

1. Preserving semantic consistency: One of the primary challenges in developing joint DA techniques for IC lies in preserving semantic consistency between images and text. This involves the identification and manipulation of common elements shared between modalities. Recognizing objects, entities, or concepts present in both images and captions is essential. Equally important is ensuring that any transformations applied to these common elements maintain their semantic meaning. For instance, if an image features a *"red car"*, any augmentation that changes the car's color to blue should be accompanied by a corresponding adjustment in the textual description to maintain consistency.

2. Handling diverse image-caption pairs: The diversity of image-caption pairs presents another significant challenge. Images can depict a wide range of scenes, objects, and contexts, making it challenging to create a one-size-fits-all approach to joint DA. Additionally, textual descriptions accompanying images vary significantly in terms of length and complexity. A robust joint DA strategy must be flexible enough to accommodate this variability and adapt accordingly. Addressing these challenges is crucial for ensuring the effectiveness and applicability of joint augmentation across diverse image-caption pairs.

3. Utilizing Established Metrics for Evaluating Joint Data Augmentation: The challenge extends to utilizing robust evaluation metrics. Unlike single-modal tasks, vision-language tasks such as Image Captioning (IC) demand more nuanced evaluation criteria. Common metrics such as BLEU [26], METEOR [27], CIDEr [28], ROUGE-L [29], and SPICE [30], which are used to assess the quality of generated captions. Establishing nuanced evaluation criteria is imperative for accurately gauging the success of joint augmentation methods in maintaining semantic coherence while introducing variability into image-caption pairs.

1.1.2 The road ahead: investigating the potential of CutOver

In this context, this master’s thesis aims to introduce and explore a novel joint DA method for IC systems, referred to as *CutOver*. The main idea behind *CutOver* is to intelligently combine two distinct DA techniques — one from the domain of CV (*CutMix*) and another from NLP (Instance Crossover Augmentation) [13, 31]. By integrating these techniques into a joint augmentation strategy, *CutOver* seeks to overcome the challenges of preserving semantic relationships while introducing variability into image-caption pairs. *CutMix*, initially proposed as a regularization strategy for training strong image classifiers with localizable features, involves replacing a rectangular region of an image with a corresponding region from another image. This approach encourages the model to focus on localizable features within images and has demonstrated success in improving image classification tasks [13]. Instance crossover augmentation, on the other hand, leverages concepts from NLP and aims to manipulate the textual descriptions associated with images while preserving their semantic content. This technique introduces variations in the textual modality, which can complement the visual variations introduced by *CutMix*. The combination of these techniques in *CutOver* presents a unique opportunity to tackle the challenges of joint DA for IC. By intelligently swapping and manipulating visual and textual elements, *CutOver* aims to diversify the training data while maintaining the semantic coherence of image-caption pairs. This research introduces *CutOver* as a potential solution to the challenges posed by DA in vision-language tasks. However, it is essential to acknowledge that not all novel approaches may be effective in addressing the complex requirements of joint DA for IC. While *CutOver* holds promise, it is imperative to conduct rigorous experiments and evaluations to assess its impact on model performance. Moreover, it is possible that the introduction of such complex transformations may not yield immediate improvements, and additional experiments may be necessary to understand the potential reasons behind variations in model performance.

1.1.3 Unlocking the potential of joint DA

In conclusion, the advancement of DA techniques tailored for IC is essential for expanding the frontiers of multimodal AI. The fusion of CV and NLP in IC opens up numerous possibilities for applications that can benefit society at large. However, realizing these possibilities demands overcoming the distinctive challenges posed by the joint DA. *CutOver*, as a novel technique, represents a significant step forward in addressing these challenges. By amalgamating methodologies from CV and NLP, *CutOver* endeavors to strike a delicate balance between introducing variability and preserving semantic consistency in image-caption pairs. This research aims to scrutinize the potential of *CutOver* in enhancing the robustness and performance of IC models. While *CutOver* shows promise, acknowledging the complexity of the task is crucial. Not all approaches may yield immediate improvements, necessitating additional experiments and analyses

to understand the intricacies of joint DA for IC. Through these experiments and by pushing the boundaries of innovation in DA, this master thesis aspires to contribute to the ongoing advancement of vision-language models and their potential to transform the way we interact with visual data. The exploration of CutOver and the broader landscape of joint DA aims to bring us one step closer to realizing the full potential of IC in diverse applications, from accessibility to content generation and beyond.

1.2 Approach and Contribution

The approach outlined in this master thesis centers around the development and evaluation of a novel joint DA method, known as CutOver, designed specifically for enhancing the performance and robustness of IC systems. CutOver represents a fusion of DA techniques from both the CV and NLP domains, to strike a balance between introducing diversity into the training data and preserving the essential semantic coherence between images and their associated textual descriptions. This approach is poised to address the challenges posed by DA in vision-language tasks and contribute to the ongoing evolution of IC systems.

1.2.1 Approach: Developing CutOver and assessing its impact

The initial approach aimed to develop CutOver as a novel joint DA method by drawing inspiration from existing techniques in CV and NLP. The steps in this approach included:

1. **Understanding the challenges:** Recognizing the complexities of joint DA for IC, the approach started with a comprehensive understanding of these challenges. It involved appreciating the delicate balance required between introducing variability and maintaining semantic coherence in image-caption pairs.
2. **Leveraging existing techniques:** CutOver was designed by amalgamating two established DA techniques — CutMix from CV [13] and instance crossover augmentation from NLP [31].
3. **CutMix:** Originating as a regularization strategy for image classifiers, CutMix involved replacing portions of one image with corresponding regions from another. This approach aimed to encourage the model to focus on localized features, enhancing generalization.
4. **Instance crossover augmentation:** This technique focused on textual descriptions, aiming to introduce textual variations while preserving semantic content.
5. **Intelligent fusion:** CutOver’s core innovation lies in its ability to intelligently combine these techniques to maintain meaningful transformations between visual and textual elements.
6. **Implementation:** CutOver was implemented as part of the data preprocessing pipeline for IC models, generating augmented image-caption pairs.
7. **Evaluation framework:** An extensive evaluation framework was established to assess CutOver’s impact on model performance, employing metrics such as BLEU, METEOR, CIDEr, ROUGE-L, and SPICE.
8. **Experiments and analysis:** Rigorous experiments were conducted to evaluate the performance of IC models trained with and without CutOver. The results were analyzed to understand the extent to which CutOver improved the quality of generated captions.

1.2.2 Contribution: Gaining insights into challenges and iterating on solutions

The primary contribution of this research lies in the exploration of challenges encountered during the development and evaluation of CutOver. While CutOver did not lead to the expected improvements in this specific case, the insights gained are valuable and contribute to the iterative nature of research:

1. *Understanding limitations:* The foremost contribution is a comprehensive understanding of the limitations of CutOver in the context of IC. These limitations may include unexpected interactions between visual and textual transformations or constraints specific to the datasets and models used.

2. *Iterative research:* This research underscores the iterative nature of scientific inquiry. Not all novel approaches may yield immediate success, but the process of exploration and experimentation contributes significantly to the field’s collective knowledge.

3. *Analyzing variances:* The analysis of experimental results provides insights into potential reasons behind variations in model performance resulting from the novel augmentation method. This analysis serves as a foundation for future investigations and refinements.

4. *Informing future work:* The findings of this research inform future work in the domain of joint DA for IC. Researchers can build upon these insights to develop more effective strategies and address the challenges more precisely.

5. *Comparative analysis:* While CutOver did not achieve the desired outcomes, the comparative analysis with existing augmentation techniques offers valuable guidance for selecting DA strategies in vision-language tasks.

In conclusion, this revised perspective acknowledges that CutOver did not yield the anticipated improvements in this specific case. However, it underscores the importance of exploring challenges, understanding limitations, and contributing valuable insights to the broader research community. Research is an iterative process, and even when an approach does not lead to immediate success, the journey often uncovers valuable knowledge that paves the way for future advancements in the field of vision-language models and IC. This exploration of challenges and insights gained exemplifies the resilience and adaptability of researchers, who continue to push the boundaries of knowledge, even when faced with unexpected outcomes. It is through such endeavors that breakthroughs are eventually achieved, advancing our understanding and capabilities in complex domains like multimodal AI and IC.

1.3 Thesis outline

In the **opening chapter**, we present the motivation and problem statement behind this thesis and highlight the challenges and opportunities in developing the joint DA method for IC systems. We then provide a concise overview of our approach and the major contribution, concluding with an outline of the upcoming chapters.

In **Chapter 2**, we delve into relevant literature in the field. Our exploration includes a discussion on both generic and recent IC architectures, with a focus on interactive IC. The subsequent section reviews work related to standard and domain-specific datasets. To provide a comprehensive understanding, we also outline relevant studies about both basic and advanced DA techniques, covering transformations for both images and textual

transformations.

In **Chapter 3**, we explore diverse deep learning concepts crucial to formulating our proposed methods in this thesis. These include the fundamentals of deep learning, IC, DA, object detection, attention mechanisms, and transformer architectures. Additionally, we provide insights into the mathematical formulations and mechanisms of various deep-learning model explanation techniques.

Chapter 4 shapes our proposed architecture and describes the pipeline. The initial section offers a high-level synopsis of our benchmark architecture. Subsequent sections furnish the intricate formulations for our selected object detector and corpora. Following this, we expound on the CutOver augmentation method, elucidating its pipeline through examples. The concluding section encapsulates the implementation details.

In **Chapter 5**, we delve into our experiments and results. The initial section provides a comprehensive data description of the *MS COCO* dataset and the *VizWiz* dataset, encompassing an overview, data source, and size, as well as the train/test split. The subsequent chapter elucidates various evaluation metrics, including *BLEU*, *ROUGE*, *METEOR*, *CIDEr*, and *SPICE*. Despite thorough exploration, our findings indicate that the proposed method did not outperform the SOTA method or any other augmentation methods. This comparison sheds light on the performance dynamics within the context of different augmentation techniques.

Chapter 6, the discussion and future works, delves into the analysis and insights derived from the study, providing a comprehensive discussion. We explore possible reasons for the unexpected results and the underperformance of our proposed method, *CutOver*. Additionally, the chapter outlines potential avenues for future research, providing a roadmap for continued exploration in this domain.

Chapter 2

Related Work

The Related Work chapter is organized as follows: the first section provides an overview of image captioning, delving into its architectures and interactive image captioning. Subsequently, the chapter explores datasets relevant to image captioning. The following section then introduces various data augmentation techniques, encompassing both image and text augmentation strategies.

2.1 Image Captioning

Image captioning is the computational task focused on generating descriptive textual sequences, denoted as $C = \{t_1, t_2, \dots, t_n\}$, where C represents the set of individual textual elements. Each t is a specific word or token contributing to the sequence, collectively aimed at articulating the content of a given image I . Image captioning involves the AI task of generating concise and human-like descriptions for images using natural language [32, 33, 34]. This process includes recognizing objects within the image and understanding scene details, object properties, and their interactions. The challenge lies in replicating the innate human ability to effortlessly correlate descriptions with encountered images. The generated captions aim to be both concise and comprehensive, summarizing the salient contents of the given image in a single sentence.

It represents a crucial undertaking at the intersection of the visual and linguistic domains. It addresses the intricate task of imbuing images with human-like understanding, enabling the recognition of contextual information within an image and subsequently enriching it with meaningful captions. At its core, IC tackles a formidable challenge: transforming the inherent visual content of an image, typically represented as a sequence of pixels, into a coherent sequence of words that effectively describes the image's content. This intricate process is intricately framed as an end-to-end sequence-to-sequence problem, where the input image is sequentially converted into a sequence of words. The architecture governing this framework hinges upon a fundamental design known as the encoder-decoder model.

In the early stages of IC research, traditional approaches primarily relied on rule-based systems and handcrafted features. These pioneering methods, including work by [35]

and [36], followed predefined rules and templates to identify objects, their spatial positions, and relationships within images, subsequently generating textual descriptions. Although serving as significant initial steps, these methods were constrained by their rigidity, often producing generic and unexpressive captions. The advent of deep learning introduced a transformation in IC, enabling the generation of more contextually relevant and expressive descriptions. Seminal models such as *show and tell* [37] and *show, attend, and tell* [38] marked this turning point. These models CNNs [39], including architectures like ResNet [40] and VGGNet [41], as encoders to convert input images into feature vectors. These feature vectors encapsulated the essence of the visual content, which was then passed to decoder layers. RNNs [42], such as LSTM [43] and GRU [42], within the decoder translated these feature vectors into meaningful output sequences.

2.1.1 Architectures

In this subsection, we present some of the generic and recent image captioning architectures.

The *Show, Attend, and Tell* [38] architecture is a caption generator that incorporates the attention mechanism in two variants: a hard attention mechanism and a soft attention mechanism. Soft attention indicates the relative importance of each part of the image to other parts. On the other hand, hard attention separates certain parts of the image, and only those parts were considered to generate the caption while ignoring the rest. A CNN is used as an encoder, and the feature maps are extracted from the lower convolution layer instead of the fully connected layer. These feature maps were flattened to produce annotation vectors corresponding to a part of the image. These annotation vectors were concatenated to generate a matrix which is used by the attention model to determine sections of the image more relevant to generate the next word. LSTM with an attention module is used as a decoder that can selectively focus on specific regions of an image by selecting a subset of all the feature vectors. The inputs of this LSTM model are previously generated words, hidden states, and the context vector which is a dynamic representation of the relevant part of the image input at time t . Both variants of the attentive model are trained with stochastic gradient descent on three datasets; MS COCO dataset [44], flickr30k [45], and flickr8K [46].

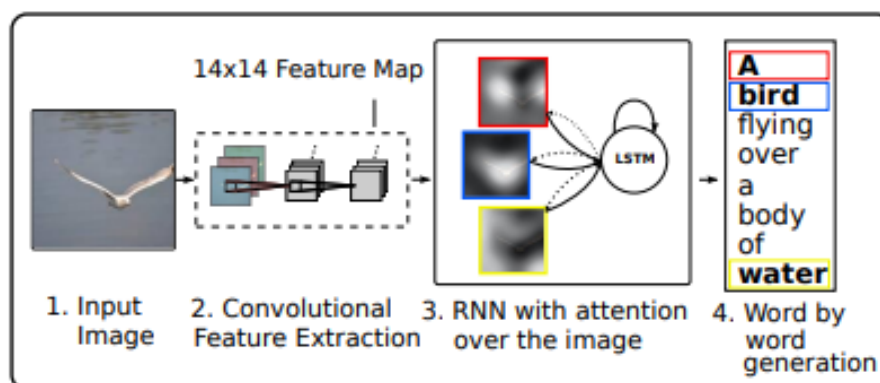


Figure 2: A basic architecture of Show, Attend and Tell [38]

Anderson et al. [47] proposed a combined *Bottom-Up and Top-Down* attention mechanism that can calculate attention at the level of objects and other salient image regions.

The bottom-up attention is implemented using Faster R-CNN [48]: an object detection model that uses bounding boxes to identify and localize instances of objects belonging to specific classes. Faster R-CNN functions as a ‘hard’ attention mechanism, as it selects only fewer image-bounding box features from a collection of possible configurations. Faster R-CNN with ResNet-101 [49] is used as an encoder, trained with a multi-task loss function to generate image feature vectors. In the decoder, two layers of LSTM are used: the first is a top-down attention model to weigh each feature during caption generation. And the second is a language model, trained with a cross-entropy loss function to generate the captions. Experiments were conducted on the MS COCO dataset and achieved a BLEU-4 score of 36.9. This method allows attention to be calculated more naturally at the level of objects and other salient locations.

Attention is all you need [50] revolutionized information processing, multimodal model construction, and sequence-to-sequence architectures, introducing transformers as a novel design to replace traditional RNN/LSTM for handling sequential data. Transformers employ self-attention, enabling the model to focus on specific parts of input sequences, facilitating better handling of long-range dependencies, and improving connections within the same sentence. This attention mechanism is crucial for tasks like translation, ensuring an accurate understanding of context and reference objects in changing scenarios.

Meshed-Memory Transformer [51] was proposed to explore the applicability of transformers in image captioning.

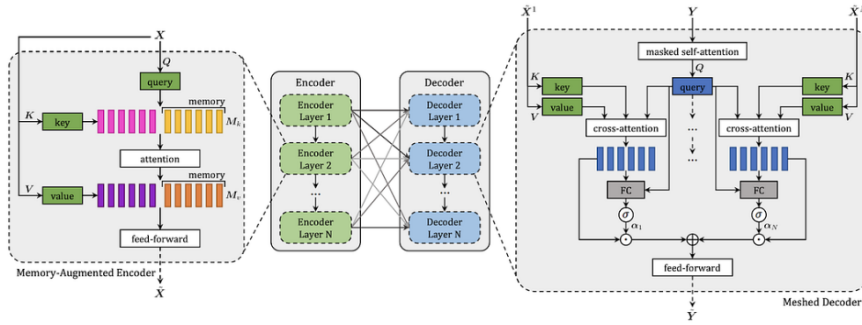


Figure 3: Meshed Memory Transformer architecture [51]

It comprises a multi-layer encoder for image regions and a multi-layer decoder that generates the output sentence. Image regions are encoded on multiple levels, considering both low-level and high-level relationships. The model can learn and encode prior knowledge when modeling these relationships using memory-augmented attention. Self-attention is based on a pairwise relationship and cannot model a priori relationships between image regions. As a solution to this limitation, a memory-augmentation encoder was proposed that extended the set of keys and values in the encoder with additional slots to extract prior information. In contrast to the original decoder block in [50] which only performs cross-attention between the last encoding layer and the decoding layers, the M2 has a meshed connection with all encoding layers. The model then summed these contributions after they had been modulated. The model was evaluated on the MS COCO dataset, and Faster R-CNN with ResNet-101 was used to represent image regions. The experiments demonstrated that the M2 transformer achieved a new state-of-art on the MS COCO dataset.

One of the recent IC models *Lemon* [52], a Large-scale iMage captiONer was introduced based on vision-language pre-training to improve the performance boost. Lemon uses the VinVL comprising an image feature extractor and a transformer model as the reference model. The multi-layer transformer model with a multi-head self-attention layer followed by a feed-forward layer on each layer, is used for multimodal fusion. The sequence-to-sequence attention mask is applied in each self-attention layer for the captioning block for text generation with the encoder layers. The output representation is either used for prediction at the end or as input to the following layer. On numerous IC benchmarks, including coco caption, nocaps, and conceptual captions, lemon has attained new state-of-the-art. Even when using a zero-shot manner, lemon has the outstanding capacity to generate captions for a diverse range of long-tail visual objects.

Another vision-language architectures *mPLUG* [53] was introduced for cross-modal understanding and generation. The goal of this architecture is to recursively exploit the effectiveness of connected cross-modal fusion and the efficiency of asymmetric co-attention for enhanced cross-modal learning.

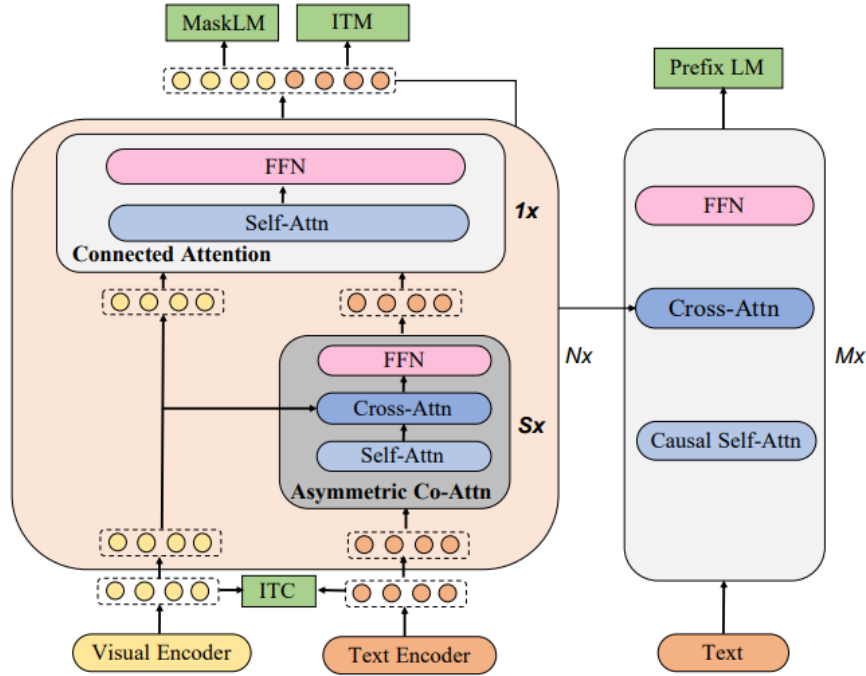


Figure 4: The model architecture of mPLUG [53]

It consists of two unimodal encoders for pictures and text separately, transformer visual encoder and text encoder are used to encode the input image patches and input text into a sequence of embeddings respectively. Following this, the representations in both the images and the languages are fed into a cross-modal skip-connected network that is made up of several skip-connected fusion blocks. The cross-modal skip-connections allow the fusion of the visual and the language representations to occur at different levels in the abstraction hierarchy across the modalities rather than at the same level. In each of the skip-connected fusion blocks, connected attention modal fusion is deployed to each of the asymmetric co-attention layers with a fixed stride value. The asymmetric co-attention block consists of a self-attention layer (SA), a cross-attention layer (CA), and

a feed-forward network (FFN). The input text feature is initially sent to the SA layer, and the visual feature is subsequently injected into the text feature by the CA layer. The results of the SA and CA are summed together and given to the FFN layer for visual-aware text representation. The connected attention model consists of the self-attention layer and the feed-forward network. The output of the asymmetric co-attention layers i.e., the image features and text features are connected and given to the SA layer and the FFN. Lastly, the result of the cross-modal representations is fed into a transformer decoder for sequence-to-sequence learning. Also, mPLUG excels in a broad range of vision-language tasks such as IC, image-text retrieval, visual grounding, and visual question answering.

2.1.2 Interactive image captioning

Automatic Judgment of Neural Network-Generated Image Captions [54] addresses the growing need for effective and efficient methods to assess the quality of captions generated by neural networks in the context of image captioning. The rise of deep learning techniques and neural captioning models has significantly advanced the field of IC, but there remains a critical challenge — evaluating the quality of generated captions. Traditionally, human annotators have been relied upon to judge the quality of captions, which is both time-consuming and potentially subject to biases. To tackle this challenge, the authors propose a novel automatic judgment system for assessing neural network-generated image captions. This system aims to provide objective and consistent evaluation of captions, thus saving time and resources while maintaining a high standard of quality assessment. The core of this system involves a multi-faceted approach to caption quality evaluation. It integrates linguistic and visual features, addressing aspects such as fluency, relevance, and diversity. Moreover, it incorporates metrics related to the captions' engagement with the content of the associated images, ensuring a comprehensive evaluation. One notable contribution is the introduction of a crowd-sourcing-based evaluation platform that provides a benchmark for comparing the quality of captions across diverse image domains. This platform allows for the systematic evaluation of neural network-generated captions, enabling researchers and developers to gauge the performance of their models more effectively. The paper leverages various automated metrics, including BLEU [26], METEOR [27], and TER [55], to systematically assess the quality of neural network-generated captions. By combining these metrics with a diverse set of features and a crowd-sourced evaluation platform, the authors provide a comprehensive framework for caption quality assessment.

Biswas et al. [56] proposed a novel IC architecture to increase the performance and explainability of [38] by augmenting their visual attention mechanism. This model can be used for various interactive machine learning and explainable artificial intelligence techniques. This paper tackles the problem of "*explanatory*" IC, an approach that seeks to provide captions that not only describe the content but also explain the underlying rationale. It introduces a novel method that leverages both top-down and bottom-up image features to enhance the descriptive and explanatory power of captions. The approach combines various elements, including beam search and re-ranking, to optimize caption generation. By doing so, it addresses the demand for interactive IC that caters to users seeking in-depth and meaningful descriptions. The central idea is to fuse top-down knowledge, such as object detection, with bottom-up features derived from the image. This fusion allows for a more comprehensive understanding of the visual context and object relationships, leading to captions that are both descriptive and explanatory. The beam search technique, in combination with re-ranking, ensures that the generated

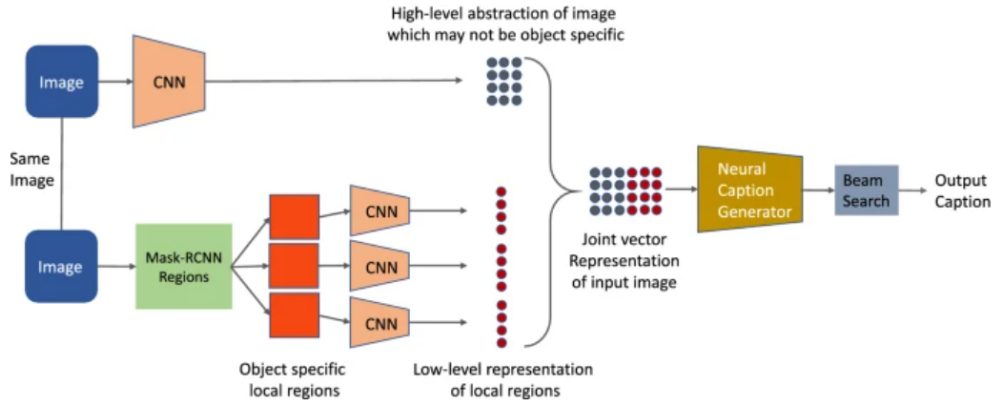


Figure 5: Caption generation with augmented visual attention builds upon Show, Attend and Tell [56]

captions not only adhere to predefined guidelines but also incorporate user-specific preferences. The goal of this paper is to create a system that aligns with user expectations, by offering not just a textual description but a deeper understanding of the image content. In the realm of interactive IC, where user feedback plays a pivotal role, the approach introduced in this work positions itself as a bridge between automated caption generation and meaningful human interaction. By employing top-down and bottom-up features, beam search, and re-ranking, the authors aim to enable a more fine-grained and interactive captioning system that supports user preferences and serves as a model for future developments in user-centric IC.

In the IC domain, most of the work is done in the English language compared to other languages. [57] addresses the challenge of enhancing German image captions using a combination of machine translation and transfer learning techniques. The primary focus is on improving the quality and fluency of image captions generated in the German language, which is particularly crucial for various applications, including accessibility and content localization. The paper leverages machine translation to translate German captions into English and then back into German. This process aims to refine the captions by aligning them with the natural language patterns in English, which has a more extensive and diverse dataset available for training captioning models. The iterative translation process helps correct language issues, improve word choice, and enhance overall caption fluency. Additionally, the authors explore transfer learning, where models pre-trained on English IC datasets are fine-tuned for the specific task of generating German captions. This approach capitalizes on the wealth of pre-existing English captioning models and fine-tunes them for German, taking advantage of the knowledge transfer. The paper showcases how this combination of machine translation and transfer learning effectively enhances German image captions. The proposed approach contributes to making image captions more accessible, natural, and culturally appropriate for German-speaking audiences. The study provides insights into the potential of cross-lingual and transfer learning methods for improving the quality of generated captions in languages with limited training data.

Another method to improve an IC model was proposed in [9] that represents a significant milestone. Image captioning, a task involving generating natural language descriptions for images, is traditionally reliant on large-scale annotated training data for the development of state-of-the-art models [58, 59, 60]. However, the impracticality of this

approach in resource-limited environments and the increasing demand for personalized image captions in an interactive ML setting have paved the way for innovative solutions. This paper introduces a transformative approach that aligns with the growing need for data-efficient and adaptable IC models. The central premise revolves around integrating human feedback into the training process of IC models. It recognizes the inherent challenges posed by the resource limitations in creating large-scale annotated datasets, especially when the goal is to cater to user-specific images. In essence, the research establishes a bridge between user interactions and the enhancement of IC capabilities. This user-centric approach not only reduces the reliance on vast annotated datasets but also builds user trust in AI/ML-based systems, a vital consideration in the age of human-AI collaboration [61, 62, 63, 64]. The key components of this approach involve starting with a base IC model, which has been pre-trained on the widely recognized MS COCO dataset. This model serves as a foundation for generating captions for previously unseen images. However, the innovative aspect comes into play when users engage with the model by providing feedback on the generated captions and the corresponding images. This feedback is harnessed for data augmentation, creating a wealth of additional training instances to facilitate the model’s gradual adaptation. To mitigate the challenges of catastrophic forgetting, a critical problem when adapting models to new data, the authors introduce a sparse memory replay component. The ultimate vision is the development of highly customizable IC models that can seamlessly adapt to new and diverse user-specific data, even in resource-constrained environments.

Putting Humans in the Image Captioning Loop [65] presents an innovative approach to enhance IC models through interactive ML. Traditional training of IC models often relies on large amounts of annotated data, which can be impractical in resource-limited settings. To address this challenge, the authors propose an interactive system that leverages human feedback to adapt the model to user-specific data efficiently. The proposed approach consists of three key components: feedback collection, data augmentation, and model update. In the feedback collection phase, the paper explores various methods to gather user feedback. Users can provide different types of feedback, such as corrected captions, marked objects or regions, and explicit alignment between corrected words and images. The goal is to strike a balance between collecting rich feedback and maintaining user engagement. The authors also consider the use of deep active learning acquisition functions to select specific examples for feedback, potentially improving the efficiency of feedback collection. Data augmentation plays a crucial role in maximizing the impact of user feedback. The paper discusses different augmentation strategies, including both caption-based and image-based techniques. Caption-based augmentation involves methods like synonym replacement, back-translation, and paraphrasing. Image-based augmentation includes various image transformations, such as cropping, warping, and flipping. The authors also explore multi-modal augmentation, where both captions and images are modified simultaneously to ensure they remain coherent. Finally, the model update phase focuses on efficiently updating the model based on the augmented training data. Instead of retraining the model from scratch, the paper explores batch-wise model updates, allowing for more efficient adaptation to new information. The authors address challenges such as avoiding catastrophic forgetting, expanding the decoder for user-specific vocabulary, and integrating information about novel objects not previously observed.

2.2 Datasets

In this section, we highlight some of the frequently utilized datasets for IC tasks, encompassing both standard and domain-specific datasets.

1. Standard captioning datasets: Standard captioning datasets encompass a diverse range of images and cover a wide array of topics, providing a general-purpose dataset for captioning models.

The Microsoft common objects in context (*MS COCO*) [44] dataset was introduced for the object recognition tasks in the context of scene understanding. This was done by collecting many non-iconic images focusing on the common objects in their natural everyday environment and different viewpoints. This dataset contains 91 object categories, 328,000 images having 2,500,000 labeled instances, and 5 captions for each of the images. One of the major advantages of this dataset is that it has larger instances per category, thereby, enhancing the 2D localization and improving the contextual information learning. The *flickr30k* [45] dataset was introduced for the IC task. The pictures were collected from the Flickr website, describing daily activities, events, and locations, annotated with five captions each. This dataset comprises 31,783 images, 29,783 training images, 1000 testing images, and 1000 validation images. This dataset is often regarded as a popular benchmark for the sentence-based IC.

2. Domain-specific datasets: Domain-specific datasets focus on particular domains or industries, tailoring the data to specific contexts or applications.

The *VizWiz* dataset [66] is a challenging dataset designed for the visual question answering tasks to facilitate the visually impaired people to address their everyday visual questions. This was done by collecting the images directly taken by the visually impaired individuals and recording a voice query about those images paired with ten captions for each question. This dataset includes 23,431 training images, 117,155 training captions, 7,750 validation images, 38,750 validation captions, 8,000 test images, and 40,000 test captions. The *GoodNews* dataset [67] is the largest news IC dataset that retrieves news articles, images, and captions from the New York Times API ranging from 2012 to 2018. This dataset consists of 466,000 images, 424,000 training images, 18,000 validation images, and 23,000 testing images. The captions were annotated by the journalists and had only a single caption for each image. *Open Images V6-Localized Narratives* was launched in 2020 with the form of multimodal annotations namely, localized narratives [68] for various IC tasks. The human annotator used a voice recording to describe each image in the dataset while hovering their mouse over the described regions of the image. This section with the local narratives now contains 1,671k images of the open images dataset [69].

2.3 Data augmentation

In this section, we summarize some of the basic and advanced DA techniques for the image and the text transformations respectively.

The two main problems in training a deep learning model are overfitting and underfitting. Overfitting occurs when a model learns the information and noise in the training data to the point where it has a negative effect on the model's performance on new data. When a model fits the training data too well, it is said to be overfit. Underfitting occurs when the model is unable to capture the underlying trend of the data, i.e., it performs well on

training data but poorly on testing data. Its presence simply indicates that the model or algorithm does not sufficiently fit the data. This problem frequently arises when there is insufficient data to train a model.

Many methods have been developed to alleviate the "*overfitting*" problem that persists in big data-driven model-based deep CNNs. One such strategy is data augmentation. Data augmentation lies at the heart of all successful applications of deep learning, ranging from image classification [70] to speech recognition [71] due to its ease of implementation and effectiveness in resolving the issue of underfitting and overfitting. It aims to deal with a lack of training data and tries to introduce data variability into real-world data via label-preserving changes. The outcomes of data augmentation are often referred to as synthetic data, created data, simulated data, or artificial data.

2.3.1 Image augmentation

Different techniques have been explored for data transformations by leveraging the substantial domain knowledge leading to improved generalizability. One among them is the basic image transformations [72] such as

- a) **Rotation**: The image can be rotated right or left on an axis at an angle between 1° and 359° .
- b) **Colour space**: The color channels of the image can be changed.
- Cropping**: Crop a central patch of each image with mixed height and width dimensions.
- c) **Flipping**: The image can be flipped horizontally, vertically, or both ways.
- d) **Scaling ratio**: Image size can be increased or decreased.
- e) **Occlusion**: Occlusion in an image occurs when an object hides a part of another object.
- f) **Salt and pepper**: Salt and Pepper noise refers to the addition of white and black dots in the image.
- g) **Blur**: Blurring is to make something less clear or distinct.
- h) **Translation**: The image can be moved horizontally, vertically, or both ways.
- i) **Contrast**: The contrast of the image can be changed or altered.

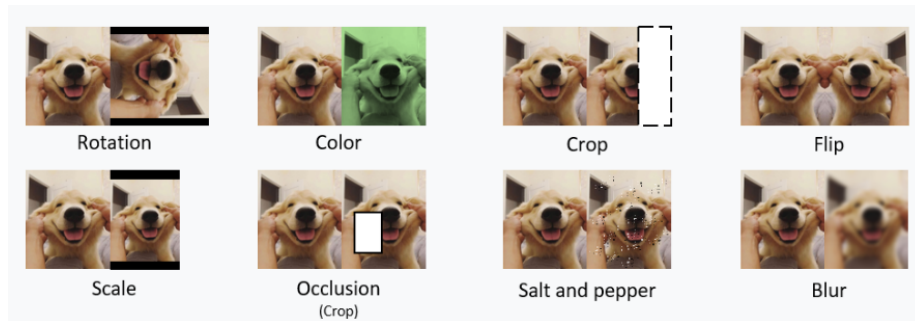


Figure 6: Different image augmentation techniques a) Rotation b) Color c) Crop d) Flip e) Scale f) Occlusion g) Salt and Pepper h) Blur [73]

Other methods include advanced techniques like *Cutout* [10], a simple augmentation technique that randomly masks out square regions of input during training to improve the robustness and the performance of the convolution neural networks. In this method, the network drops the units on the input stage rather than the intermediate layers, so that visual features, including the objects eliminated from the input image, are similarly removed from all subsequent feature maps. Cutout solves the problem of recognizing

partial or occluded images by letting the model evaluate preferably on the minor features rather than the major features of the image.

Mixup [11], a form of vicinal risk minimization [74] was proposed to reduce unenviable problems such as memorization and sensitivity to adversarial examples of large deep neural networks. This is a DA method that regularizes the neural network by training the model on the convex pairs of examples and their labels. While certainly improving classification performance, Mixup samples tend to be unnatural as it makes full use of pixels. Mixup samples suffer from the fact that they are locally ambiguous and unnatural, and therefore confuse the model, especially for localization.

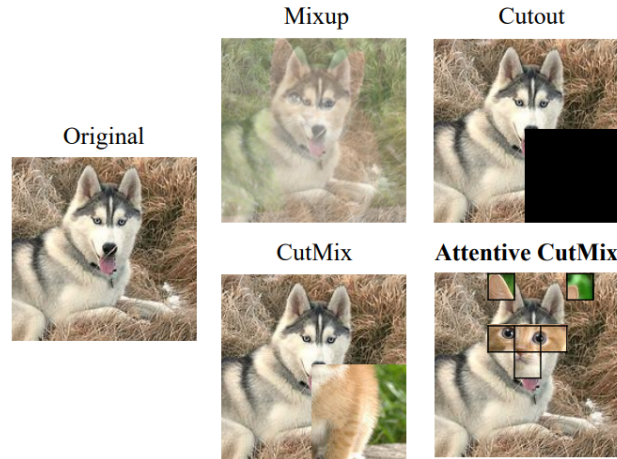


Figure 7: A visual comparison of a) Mixup b) Cutout c) CutMix d) Attentive CutMix [75]

The regional dropout methods eliminate informative pixels from the training images superimposing a patch of either black pixels [10] or random noise [76] leading to data loss and inefficiency during training. To overcome these issues, **CutMix** [13]: an augmentation strategy was proposed where the removed regions are replaced with the patches that are cut and pasted within mini-batches from the training images. The ground truth labels are combined proportionally to the area of the patches. Therefore, it has no uninformative pixels during training, making training more efficient, and improving localization by requiring the model to recognize the object from a partial view. CutMix has not only shown performance boosts in the image classification tasks but also for a wide range of localization tasks and transfer learning experiments. Furthermore, merely employing the CutMix-imageNet pre-trained model for the object detection and the IC tasks improves the overall performance.

Even though CutMix [13] has demonstrated significant effectiveness in classification and localization performance, it hinders the practical results of the method. This is due to the chances of cutting the unimportant background patch and pasting it onto the second image simultaneously since the patch is cut of random size and location from the image. As a result, it becomes more difficult for the model to overfit on a specific subject and forced to learn more significant features associated with that subject rather than the subject of interest. **Attentive CutMix** [77], an attention-based CutMix augmentation strategy was introduced to address this shortcoming. In this method, a pre-trained network decides the most meaningful or representative regions within the image then the top "N" attentive region patches are selected to cut from the first image. In the second image, these cutout patches are pasted in the exact location as in the first image.

The experimental results on CIFAR-10 and CIFAR-100 [78] exhibit that this attention method generates robust image fusing without additional training or testing costs and outperforms the baseline methods by a considerable margin.

2.3.2 Text augmentation

Text data augmentation is based on the concept of semantically invariant transformation [79], which means the data transformation should be done such that the newly generated data does not change the class label. A range of research is done on different augmentation techniques that can be applied to all types of texts, sentences, and paragraphs. Examples of these diverse text DA methods are provided in the table 1.

Augmentation method, *lexical embedding* [80] is done on the Twitter corpus [81] for analyzing annoying behaviors in social media. This approach used k-nearest neighbors (knn), where each word in a tweet is replaced by their knn words based on the cosine similarity between the word and its knn neighbors. A similar method, *synonym replacement*, was proposed as a part of the EDA [21] technique. In this method, the model chooses n words at random from the sentences that do not stop words, and each of these words is replaced with a random synonym. *Word replacement/lexical replacements* [79] using the thesaurus is the same as synonym replacement. But, in this method, hyperonyms are preferred for lexical substitution (more general word, sparrow => bird) while hyponyms are avoided (more specific word, bird => sparrow).

However, the synonym-based augmentation techniques can be applied to only a small fraction of the vocabulary as the words having identical or nearly the same meanings are too less. As a result, the synonyms are quite restricted, and the synonym-based augmentation cannot build a wide range of patterns from the original texts. Kobayashi [82] introduced a method known as *contextual augmentation*, in which the words are replaced with substitute words that are predicted by a bi-directional LM, given the context surrounding the original words to be augmented at the word positions. Here, sample words are generated at each word position in the sentences for augmentation at each update during model training. But in this case, the substitute word could be the opposite of the original word which will change the label making contextual augmentation incompatible with the annotated labels of the original sentences. As a solution, LM with label-conditioned architecture was introduced, which fed also the label into the bidirectional LM, resulting in an output calculated from the combined information from both the label and the context.

Random insertion, random swap, and random deletion are the other EDA methods [21]. *Random insertion* finds a random synonym for a non-stop word in a sentence and inserts that synonym into the sentence at random. *Random swap* chooses two words at random from the sentence and swaps their positions. *Random deletion* eliminates each word in the statement at random with a probability p . Similarly, *Textual noise injection* [79] injects weak textual sounds: changes, additions, deletions of letters in words, changes of the case, and modification of punctuation in the text. *Spelling errors injection* [79] is a way of injecting the noise to create texts based on a list of the most common misspellings in English to train the models, thus making them more resistant to this specific kind of textual noise.

Data augmentation approaches like *two-way translation* and *instance crossover* [31] were proposed for the detection of sentiment polarity of Spanish tweets. An external machine translation service is employed in two-way translation to convert tweets to other "pivot" languages and then back to Spanish. This allows to bring lexical and syntactical

variants to tweets while retaining their sense in most cases. Instance Crossover is a new augmentation concept focused on creating new data by mixing pairs of sentences with the same label. This method divides the tokenized tweets into two halves, then randomly samples and merges the first and second halves. The instance Crossover is a very rough and naive method but is beneficial to add more data variability than two-way translation.

S.No.	AUGMENTATION METHODS	ORIGINAL TEXT	AUGMENTED TEXT
1	Lexical embedding/ Synonym replacement/ Word replacement	A man laying on bench holding leash of dog sitting on ground.	A man resting on bench holding collar of dog sitting on ground.
2	Contextual augmentation	The movie is funny	The actor is funny The performance is funny The plot is funny The scene is funny
3	Random insertion	A girl is talking on the phone while walking in the park.	A girl is talking on the speaking phone while walking in the strolling park.
4	Random swap		A girl is walking on the phone while speaking in the park
5	Random deletion		A girl is on the phone while in the park.
6	Textual noise injection	Alice's dog is swimming in the lake	Alice dog is Swimming in the lke .
7	Spelling errors injection	The children are eagerly waiting to receive their gifts tomorrow .	The children are eagerly waiting to recieve their gifts tommorrow .
8	Two-way translation	Hello, how are you? Hallo, wie geht's?	Hallo, wie geht's? Hello, how are you doing?
9	Instance crossover	The perfume is fabulous. I like the perfume The perfume has the lavender aroma. Love the fragrance.	The perfume is fabulous. Love the fragrance. The perfume has the lavender aroma.I like the perfume

Table 1: Examples of text DA techniques

Chapter 3

Technical Background

The technical background chapter provides a comprehensive overview of the essential elements underpinning the thesis work. From the fundamentals of deep learning and image captioning to exploring data augmentation, object detection, and attention mechanisms, we lay the groundwork for a detailed exploration in subsequent sections.

3.1 Deep learning

In this section, we present a brief overview of fundamental deep learning architectures, focusing on CNNs and LSTM networks.

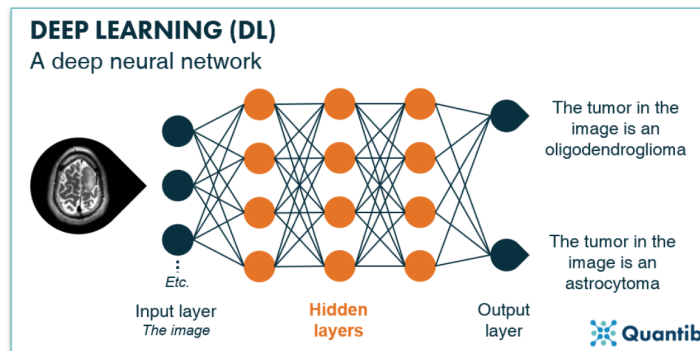


Figure 8: The image illustrates the anatomy of a deep neural network, revealing its hierarchical structure. The network comprises input, hidden, and output layers, interconnected by nodes that symbolize the neural units. This visual representation provides a detailed glimpse into the intricate design of deep learning, showcasing the flow of information through the layers, ultimately reaching the network's capacity for sophisticated computation and pattern extraction. [83]

Deep learning, a subfield of ML, has emerged as a game-changing paradigm that has

revolutionized numerous domains through its ability to automatically learn intricate patterns and representations from data. This transformative approach is built upon the foundation of neural networks, particularly deep neural networks, which are designed with multiple layers to extract complex features and hierarchies from raw information. Deep learning's ascent to prominence has been fueled by its remarkable achievements in CV, NLP, and speech recognition, among other areas. Deep learning's historical roots can be traced back to pioneering work in the late 20th century, but it gained widespread recognition and traction in the early 21st century. LeCun et al. [84] note that deep learning models, inspired by the structure of the human brain, excel in representation learning. This entails the automatic discovery of features that are essential for understanding and making predictions about data. These models, commonly referred to as artificial neural networks, have demonstrated exceptional prowess in tasks like image classification, object detection, and speech synthesis.

One of the defining characteristics of deep learning is its ability to perform end-to-end learning. This means that deep neural networks can directly map raw input data to output predictions without relying on manual feature engineering. Schmidhuber [85] highlights the transformative impact of this approach, which eliminates the need for domain-specific feature extraction and allows the model to learn relevant representations from the data itself. This end-to-end learning has significantly streamlined the development of AI systems, making them more adaptable to various applications.

Scalability is another hallmark of deep learning, enabling the handling of vast datasets and intricate tasks. Goodfellow et al. [86] emphasize that deep neural networks are designed to process high-dimensional and unstructured data, making them well-suited for tasks such as image recognition, language translation, and natural language understanding. The flexibility of deep learning architectures further contributes to their widespread adoption. CNNs [39] are tailored for CV tasks, while RNNs [42] excel in processing sequential data like natural language.

Transfer learning, a technique frequently employed in deep learning, leverages pre-trained models to bootstrap learning for new tasks. Bengio [87] discusses how this approach has the potential to save significant training time and data, as models pre-trained on one task can be fine-tuned for related tasks. Transfer learning has been particularly valuable in areas like image classification, where models pre-trained on massive image datasets can be adapted for specialized applications with relatively small datasets.

3.1.1 Convolution neural networks

The architecture of a CNN [39] is inspired by the human visual system, designed to recognize patterns and objects in images. It consists of multiple layers, each with specific functions in feature extraction, transformation, and classification. A typical CNN architecture comprises the following key components:

1. **Input layer:** The input layer represents the raw image data, usually in the form of a grid of pixel values. For color images, it consists of three channels (red, green, and blue), while grayscale images have one channel. The input size is typically fixed, but CNNs can handle various input sizes through techniques like resizing or cropping.
2. **Convolutional layers:** Convolutional layers are the core building blocks of CNNs. They consist of a set of learnable filters (or kernels) that slide over the input image. Each filter detects specific features or patterns, such as edges, corners, and textures, within a

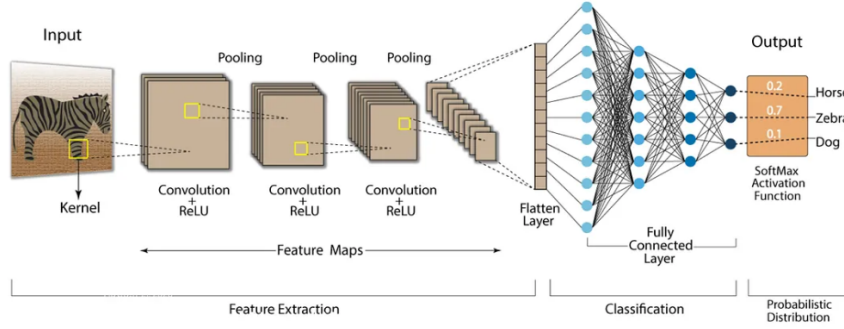


Figure 9: A convolutional neural network architecture [88]

local receptive field. Convolutional operations involve element-wise multiplications and aggregation to produce feature maps. The depth of the feature maps corresponds to the number of filters used in the layer. Multiple convolutional layers capture increasingly complex and abstract features.

$$\text{Convolution: } (I * K)(x, y) = \sum_{i,j,c} I(i, j, c) \cdot K(x - i, y - j, c)$$

3. Activation functions: Following the convolution operation, an activation function is applied element-wise to introduce non-linearity to the model. One widely used activation function is the Rectified Linear Unit (ReLU), defined as

$$\text{ReLU: } f(x) = \max(0, x)$$

In this expression, " \max " stands for the maximum function, comparing the input x with zero and outputting the greater value. ReLU, by returning zero for negative inputs and leaving positive values unchanged, facilitates the modeling of complex relationships in the data. Its computational efficiency and ability to introduce non-linearity make ReLU a preferred choice in various deep learning architectures.

4. Pooling layers: Pooling layers play a crucial role in reducing the spatial dimensions of feature maps while retaining essential information in CNNs. The commonly used technique, max-pooling, involves retaining the maximum value within a small region of the feature map. In addition to max-pooling, average pooling is another technique where the average value in a small region is calculated. These pooling operations contribute to controlling the model's size and computational complexity, while also introducing translational invariance to enhance the network's robustness.

$$\text{Max-Pooling: } P(x, y) = \max_{i,j} I(2x + i, 2y + j)$$

$$\text{Average Pooling: } P(x, y) = \frac{1}{4} \sum_{i,j} I(2x + i, 2y + j)$$

5. AdaptiveAvgPool2d: The utilization of adaptive average pooling plays a crucial role in accommodating variable input sizes and ensuring consistent feature representation. PyTorch provides the AdaptiveAvgPool2d layer, which offers a dynamic approach to spatial pooling. Unlike traditional average pooling layers that rely on fixed kernel sizes,

AdaptiveAvgPool2d allows the specification of the desired output size, adapting the pooling window dynamically based on the input dimensions. This flexibility proves especially valuable when dealing with CNN for CV tasks, where the spatial dimensions of input tensors may vary. By incorporating AdaptiveAvgPool2d in the network design, we can seamlessly transition from convolutional layers to fully connected layers, facilitating a uniform input size for subsequent operations. This layer is instrumental in achieving consistent and effective feature extraction while accommodating diverse input dimensions, contributing to the robustness and adaptability of the neural network architecture.

6. Fully connected layers: Fully connected layers come after the convolutional and pooling layers. Neurons in these layers are connected to every neuron in the previous layer. They capture high-level abstractions and spatial hierarchies in the feature maps. Fully connected layers are typically used for classification or regression tasks.

7. Output layer: The output layer's structure depends on the specific task. In classification tasks, it usually contains neurons corresponding to class labels, with a softmax activation function to output class probabilities. In regression tasks, it may have a single neuron with a linear activation function for numerical predictions.

$$\text{Softmax: } \sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}, i = 1, 2, \dots, K$$

8. Loss function: The selection of the loss function is a critical decision dependent on the specific task at hand in ML. Commonly used loss functions include: MSE is commonly employed in regression tasks, where the goal is to predict continuous numerical values. It measures the average squared difference between the predicted and true values. The formula is given by:

a. **Cross-entropy loss (or log loss) for classification:** This loss function is well-suited for classification tasks. It measures the dissimilarity between the predicted class probabilities and the true class labels. The formula for binary classification is often expressed as:

$$L(y, \hat{y}) = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})),$$

where y is the true label and \hat{y} is the predicted probability.

b. **MSE for regression:** MSE is commonly employed in regression tasks, where the goal is to predict continuous numerical values. It measures the average squared difference between the predicted and true values. The formula is given by:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

where N is the number of samples, y_i is the true value, and \hat{y}_i is the predicted value.

8. Optimization algorithm: Optimization algorithms like SGD, Adam, or RMSprop are used to update the network's parameters during training. The goal is to minimize the loss function by adjusting the weights and biases.

9. Dropout: Dropout is a regularization technique used to prevent overfitting. It randomly drops a fraction of neurons during training, forcing the network to learn more robust features.

10. Batch normalization: Batch normalization normalizes the activations of each layer to ensure stable and efficient training. It helps mitigate issues like vanishing gradients and accelerates convergence.

CNNs are designed to automatically extract hierarchical features from images. Each layer, from convolutional to fully connected, plays a crucial role in transforming the input data and learning representations that enable effective visual recognition tasks, such as image classification and object detection. The architecture and components can vary based on the specific CNN model and task at hand.

3.1.2 Long short-term memory

LSTM [43] networks represent an advancement in the domain of RNNs [40], strategically crafted to overcome inherent challenges in modeling sequential data and circumvent the vanishing gradient problem encountered by conventional RNNs. The ubiquity of LSTMs in diverse applications, such as natural language processing, speech recognition, and time series prediction, underscores their efficacy in capturing intricate temporal dependencies. This elucidation delves into the intricacies of LSTM architecture and provides a detailed exploration of the mathematical equations governing its operations.

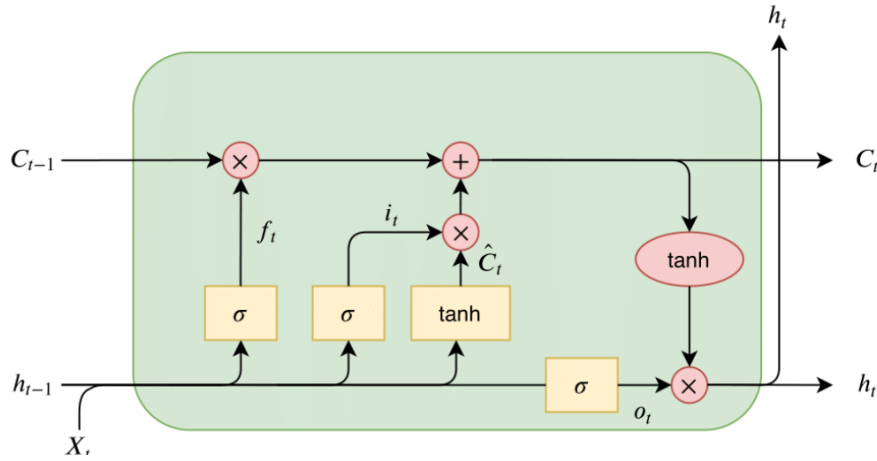


Figure 10: LSTM cell architecture, here X_t : input time step, h_t : output, C_t : cell state, f_t : forget gate, i_t : input gate, o_t : output gate, \hat{C}_t : internal cell state [89]

At its essence, an LSTM comprises a memory cell that serves as the central repository for information, maintaining and regulating the flow of data across various time steps. The hallmark of LSTMs lies in their utilization of gating mechanisms, each serving a unique purpose. The input gate i_t , forget gate f_t , and output gate o_t collectively empower the network to selectively process and store pertinent information. This adaptability positions LSTMs as powerful tools for learning complex patterns within sequential data.

Now, delving into the mathematical underpinnings of LSTMs, the input gate i_t is defined by the sigmoid function applied to a linear combination of input x_t , previous hidden state h_{t-1} , and corresponding weights and biases:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi})$$

Similarly, the forget gate f_t and output gate o_t follow analogous formulations:

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho})$$

The candidate memory cell content \tilde{c}_t is computed using the hyperbolic tangent function:

$$\tilde{c}_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg})$$

The final memory cell state c_t is determined by combining the previous memory cell content c_{t-1} , the forget gate f_t , and the input gate i_t :

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

Finally, the hidden state h_t is computed based on the output gate and the hyperbolic tangent of the memory cell content:

$$h_t = o_t \odot \tanh(c_t)$$

The intricate orchestration of architecture and mathematical formulations allows LSTMs to effectively capture and retain information over extended sequences, making them well-suited for tasks such as natural language processing, time series prediction, and speech recognition. The flexibility of the gating mechanisms empowers LSTMs to learn and adapt to complex patterns within sequential data, making them a powerful tool in the realm of deep learning.

3.2 Image captioning

Image captioning is an interdisciplinary field that combines CV and NLP to automatically generate descriptive and contextually relevant textual captions for images. This field has gained substantial attention due to its wide-ranging applications, such as enhancing accessibility tools for the visually impaired and improving content recommendation systems.

Deep learning - The foundation of image captioning: The proliferation of deep learning techniques has been instrumental in advancing IC. Deep neural networks, including CNNs for image feature extraction and RNNs for sequence generation, form the backbone of IC models. CNNs excel at capturing intricate visual details within images [90], while RNNs are adept at processing sequential data, making them suitable for generating coherent textual descriptions [37].

Multimodal data handling: IC is inherently multimodal, requiring the seamless integration of visual and textual information. This integration involves encoding the visual content of the image into a fixed-length vector using CNNs, which serves as input to an RNN-based decoder responsible for generating captions word by word [38]. Attention mechanisms have been pivotal in enhancing this process, enabling models to focus on relevant image regions while generating captions [38].

3.3 Data augmentation

Data augmentation is a critical technique in ML and CV used to artificially increase the size of a training dataset by applying various transformations to the original data. These transformations create new, slightly modified versions of the existing data, which helps improve the robustness and generalization ability of ML models, particularly in tasks like image classification and object detection. In this explanation, we will delve into data

augmentation, its importance, and its applications, supported by relevant citations. DA serves several essential purposes in ML:

- 1. Increased diversity:** By creating variations of the original data, data augmentation introduces diversity into the training dataset. This diversity is crucial for preventing models from overfitting to the specific examples in the training set.
- 2. Improved robustness:** Augmented data exposes models to a broader range of scenarios and variations, making them more robust to different conditions, lighting, and orientations.
- 3. Reduced overfitting:** With a larger dataset, models are less likely to memorize the training data and instead focus on learning the underlying patterns.
- 4. Better generalization:** Augmentation techniques encourage models to generalize better to unseen data by training them on a more representative set of examples.

Common DA techniques are:

- 1. Image augmentation:** In computer vision, image augmentation is widespread. Techniques include rotation, flipping, zooming, cropping, changing brightness, adding noise, and applying geometric transformations like affine and perspective transformations.
- 2. Text augmentation:** For natural language processing, text data can be augmented by adding synonyms, changing word order, or introducing grammatical variations. This helps improve text classification and sentiment analysis models.
- 3. Audio augmentation:** In speech recognition and audio analysis, audio data can be augmented by adding background noise, changing pitch, speed, or volume, and applying time-stretching or time-shifting operations.

DA finds applications across various domains and ML tasks:

- 1. Image classification** [90]: Augmenting image data is widely used in image classification tasks. For example, in the ImageNet Large Scale Visual Recognition Challenge, DA played a crucial role in improving model performance by artificially expanding the training dataset.
- 2. Object detection** [91, 92, 48, 93] : In object detection tasks, augmenting both images and object bounding boxes helps train models to detect objects accurately under different scales and orientations.
- 3. Semantic segmentation** [94] : Augmentation is applied to pixel-level annotation in semantic segmentation tasks to create diverse training samples, enabling models to segment objects effectively in different contexts.
- 4. Speech recognition** [95, 96]: In automatic speech recognition, augmenting audio data with variations in speed, pitch, and background noise enhances the model's ability to recognize speech in real-world environments.
- 5. Text classification** [97, 98]: For text classification tasks, augmenting textual data with synonyms, paraphrases, or slight modifications of sentences increases the model's understanding of different phrasings and expressions.
- 6. Medical imaging** [99]: In medical image analysis, DA helps train models to recognize pathologies and abnormalities across various patient populations and imaging conditions.

Data augmentation is a fundamental technique in ML and CV that enhances model performance by increasing dataset diversity, improving robustness, and reducing overfitting. It is widely employed in image classification, object detection, natural language

processing, speech recognition, and medical imaging, among other domains. By creating augmented versions of the data, ML models can generalize better and perform more effectively in real-world scenarios.

3.4 Object detection

This section provides a concise overview of object detection and its different methods like R-CNN, Fast R-CNN, and Faster R-CNN, emphasizing their key components and workflow.

Object detection is a fundamental task in CV with numerous applications, from autonomous driving and surveillance to image retrieval and augmented reality. This technique involves identifying and locating objects of interest within images or video frames, often bounding them with rectangles or polygons. In this detailed explanation of object detection, we will explore the key concepts, methodologies, and recent advancements in the field. Object detection addresses the challenge of simultaneously classifying objects and determining their precise locations within an image or video frame. Unlike image classification, which identifies the main object in an image, object detection identifies multiple objects and their positions. The ability to detect and locate objects in images is crucial for a wide range of applications, including self-driving cars, medical imaging, and content-based image retrieval. Object detection is a crucial CV task with numerous applications. Over the years, various methodologies, including R-CNN [91], YOLO [93], and SSD [100], have been developed to address the challenges in object detection. Recent advancements, such as EfficientDet [101] and DETR [102], continue to push the boundaries of accuracy and efficiency. As object detection remains a vibrant field of research, it holds the promise of further improvements in real-time, accurate, and efficient object localization and classification.

Object detection is a complex task due to several challenges:

1. **Object scale and size variation:** Objects can appear at different scales and sizes within an image, making it essential for detection models to handle scale variations.
2. **Object occlusion:** Objects may be partially or fully occluded by other objects or elements in the scene, requiring models to handle occlusions robustly.
3. **Cluttered backgrounds:** Cluttered or complex backgrounds can make it challenging to distinguish objects from their surroundings.
4. **Object pose and orientation:** Objects can appear at various poses and orientations, requiring models to be rotation-invariant.
5. **Real-time processing:** Many applications, such as autonomous driving, demand real-time object detection, imposing strict computational constraints.

Several methodologies have been developed for object detection over the years. Some of the prominent ones include:

3.4.1 Region-based convolutional neural networks

R-CNN [91] represents a landmark in the evolution of object detection methodologies within CV. Conceived to address the limitations of earlier approaches, R-CNN introduced a paradigm shift by decoupling the tasks of region proposal and object classification into a two-stage process.

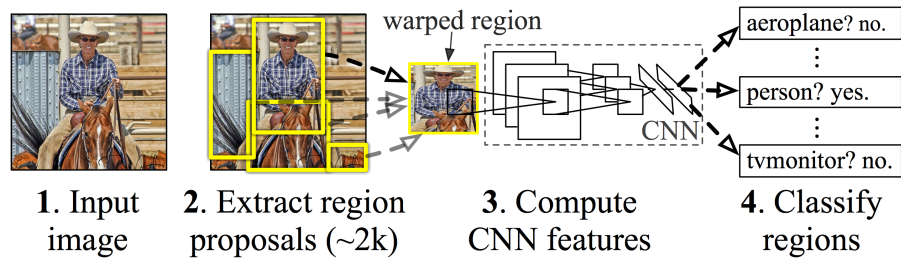


Figure 11: The architecture of R-CNN [91]

The key components of the R-CNN family include:

1. **Region proposal:** In the initial stage, a selective search algorithm is employed to propose a set of candidate regions within an image that is likely to contain objects. These proposed regions serve as the input for subsequent processing.
2. **Feature extraction:** Each proposed region is then individually processed through a CNN to extract relevant features. This stage involves resizing the region to a fixed size and passing it through pre-trained CNN layers to obtain a feature representation.
3. **Object classification:** The extracted features from each region are used for object classification through a set of fully connected layers. In the original R-CNN, a support vector machine is employed for classification.
4. **Bounding box regression:** Additionally, R-CNN includes a bounding box regression step to refine the proposed regions, improving the localization accuracy of the detected objects.

While groundbreaking, the original R-CNN suffered from computational inefficiencies due to the independent processing of each region. This limitation led to subsequent improvements in the R-CNN family, including:

3.4.2 Fast region-based convolutional neural networks

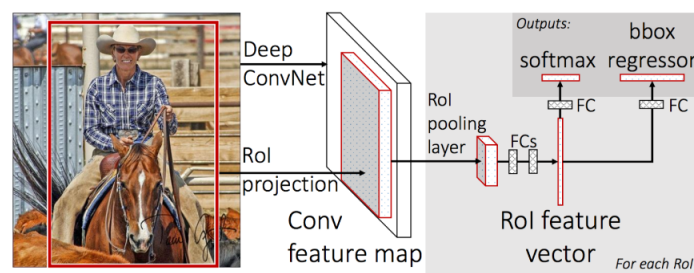


Figure 12: The architecture of Fast R-CNN [92]

Fast R-CNN [92] is a significant advancement within the R-CNN family, aiming to address computational inefficiencies present in the original R-CNN. Fast R-CNN introduced key improvements, streamlining the two-stage process of region proposal and object classification, and significantly enhancing the speed and efficiency of object detection. This method improved the speed of object detection by introducing RoI pooling and

using a single network for both region proposal and object classification. Faster R-CNN introduced the RPN for generating region proposals.

Introducing the Key Components and Workflow:

- 1. Region proposal:** Fast R-CNN replaces the selective search algorithm used in the original R-CNN with a RPN. The RPN generates region proposals directly as a part of the end-to-end training process. The RPN operates on the convolutional feature maps obtained from the input image, generating region proposals based on predefined anchor boxes and their associated scores.
- 2. Feature extraction:** The entire image is passed through a CNN to obtain convolutional feature maps. The proposed regions are then aligned with the feature maps, enabling accurate extraction of features for each region.
- 3. RoI pooling:** Fast R-CNN introduces RoI pooling, which efficiently extracts a fixed-size feature map for each region proposal, regardless of its size or aspect ratio. RoI pooling is crucial for maintaining spatial information and ensuring that the extracted features align properly with the region of interest.
- 4. Object classification and bounding box regression:** The RoI-pooled features are then passed through fully connected layers for object classification and bounding box regression. A softmax layer provides object class probabilities, and bounding box regression refines the predicted bounding box coordinates.
- 5. Training and backpropagation:** Fast R-CNN is trained end-to-end, allowing for joint optimization of both region proposal generation and object classification tasks. The model is trained using a multi-task loss function, encompassing classification loss and regression loss for bounding box refinement.

Some of the advantages are as follows:

Improved efficiency: By sharing convolutional features across region proposals, Fast R-CNN significantly reduces redundant computations, making it more computationally efficient compared to the original R-CNN.

End-to-end training: The end-to-end training process simplifies the model and enables seamless optimization, leading to improved performance.

Fast R-CNN's innovations have had a lasting impact on the field of object detection, paving the way for subsequent advancements such as Faster R-CNN and contributing to the development of faster and more accurate models for real-world applications in CV.

3.4.3 Faster region-based convolutional neural networks

Faster R-CNN [48] represents a groundbreaking evolution within the R-CNN family. Addressing the limitations of its predecessors, Faster R-CNN introduces an end-to-end trainable architecture that integrates the region proposal generation process directly into the network, further enhancing the efficiency and accuracy of object detection.

Let's explore the key components and workflow of Faster R-CNN, where a transformative addition is the integration of a RPN directly into the architecture, allowing for convolutional generation of region proposals and scores, thereby optimizing the entire object detection process

- 1. Region proposal network:** A crucial innovation in Faster R-CNN is the integration of an RPN into the network architecture. The RPN operates convolutionally on the feature maps, efficiently proposing regions with associated scores for potential objects. RPN

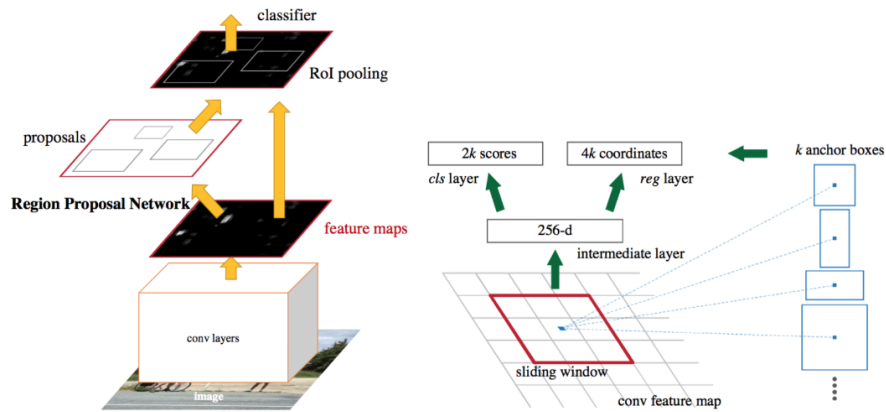


Figure 13: An illustration of Faster R-CNN model [48]

generates region proposals based on predefined anchor boxes and their corresponding scores, creating a more unified and streamlined detection pipeline.

2. Anchor boxes: Faster R-CNN employs anchor boxes of different scales and aspect ratios to efficiently capture objects of varying shapes and sizes. These anchor boxes serve as reference templates for the RPN during proposal generation.

3. RoI pooling and feature extraction: RoI pooling is utilized to align and extract fixed-size feature maps for each region proposal. This ensures that spatial information is preserved and accurately aligned with the regions of interest. The features extracted from RoI pooling are then used for subsequent object classification and bounding box regression.

4. Object classification and bounding box regression: Fully connected layers process the RoI-pooled features for object classification, providing class probabilities through a softmax layer. Simultaneously, bounding box regression refines the predicted bounding box coordinates for increased localization accuracy.

Examining the advantages of Faster R-CNN reveals its prowess in delivering an end-to-end trainable model, efficient proposal generation through the integration of an RPN, and adaptability to diverse object sizes, collectively enhancing the efficiency and performance of object detection.

End-to-end training: Faster R-CNN introduces an end-to-end trainable model, allowing for joint optimization of the RPN and object detection network. This unified training process enhances model efficiency and performance.

Improved speed and accuracy: The integration of the RPN eliminates the need for a separate proposal generation step, resulting in a more efficient and accurate detection system compared to previous R-CNN variants.

Flexibility and adaptability: The use of anchor boxes provides flexibility in handling objects of different sizes and aspect ratios, contributing to the model's adaptability across diverse datasets and scenarios.

Faster R-CNN has established itself as a cornerstone in object detection methodologies, setting the stage for subsequent advancements in the field and demonstrating the potential for unified, end-to-end trainable architectures in CV.

3.5 Attention mechanism

This section delves into attention mechanisms, exploring both soft and hard attention mechanisms. It elucidates their key components, workflow, significance, and diverse applications.

Neural networks have revolutionized the field of ML, demonstrating remarkable capabilities in various tasks, from image recognition to NLP. However, traditional neural networks have inherent limitations when it comes to processing sequences of data, where different parts of the input may carry varying levels of importance. This limitation becomes evident in scenarios where a model needs to focus selectively on specific elements within the input data. Herein lies the motivation for the introduction of attention mechanisms in neural networks. Attention mechanisms are essential in neural networks due to their ability to selectively weigh and prioritize different parts of the input data, effectively mimicking the human cognitive process of selectively focusing on relevant information. This feature becomes particularly crucial in scenarios where traditional neural networks may struggle, including:

1. **Variable-length sequences:** In many real-world applications, data comes in the form of variable-length sequences. For example, in NLP, sentences vary in length, and in image analysis, the number of objects in an image may differ. Traditional neural networks with fixed-sized inputs, struggle to handle such variability efficiently. Attention mechanisms allow the model to adapt to varying input lengths by focusing more on relevant elements while ignoring or de-emphasizing irrelevant ones.

2. **Long-range dependencies:** Certain tasks require capturing dependencies between distant elements in a sequence. For instance, in language translation, understanding the relationship between words at the beginning and end of a sentence is vital. Traditional neural networks, which typically employ fixed-size windows or filters, may fail to capture long-range dependencies effectively. Attention mechanisms enable the model to establish connections between distant elements by assigning higher attention weights to those elements, facilitating better information flow.

3. **Multiple sources of information:** In complex tasks, multiple sources of information may be available simultaneously. For example, in machine translation, a model may benefit from considering both the source sentence and its translation contextually. Traditional neural networks often struggle to incorporate and integrate information from multiple sources efficiently. Attention mechanisms provide a solution by allowing the model to attend to relevant parts of each source selectively.

4. **Ambiguity and noise:** Real-world data is often noisy and ambiguous. Traditional neural networks are vulnerable to being misled by noisy or ambiguous information, as they treat all input elements equally. Attention mechanisms enable the model to reduce the impact of noise and ambiguity by emphasizing more reliable and contextually relevant elements.

Incorporating attention mechanisms into neural networks mitigates these limitations and enhances their capacity to handle complex, variable, and information-rich data. The concept of attention itself draws inspiration from human cognition, where our brains naturally allocate attention to the most relevant aspects of a situation. This alignment with cognitive processes has contributed to the success of attention mechanisms in ML tasks. Several notable architectures have leveraged attention mechanisms to achieve state-of-the-art results. For instance, the transformer model, introduced by [50], relies heavily on self-attention mechanisms to process sequences and has become the cornerstone of many

natural language processing tasks. This demonstrates the profound impact that attention mechanisms can have on model performance. Attention mechanisms are indispensable in neural networks due to their ability to address the limitations of traditional models in handling variable-length sequences, capturing long-range dependencies, integrating information from multiple sources, and mitigating the impact of noise and ambiguity. These mechanisms align with human cognitive processes, making them a powerful tool in various ML tasks and serving as a foundation for many recent breakthroughs in the field.

Attention mechanisms in neural networks draw inspiration from the concept of "*attention*" in human cognition, a fundamental aspect of perception and information processing. In the context of ML, attention mechanisms have been adapted to replicate and enhance this cognitive process. Understanding the theoretical foundations of attention mechanisms requires an exploration of both their cognitive origins and their ML applications. Attention is a cognitive process that allows humans to focus selectively on specific elements of their sensory input, enhancing the processing of relevant information while filtering out distractions. This cognitive mechanism is closely tied to our ability to perceive and make sense of the world. It operates at various levels, from simple sensory attention, such as focusing on a specific sound in a noisy environment, to more complex cognitive attention, such as reading comprehension or problem-solving. One of the key principles of attention in human cognition is the idea that attention is not uniformly distributed but can be selectively allocated to specific regions or objects. This allocation is often guided by factors such as saliency, relevance, and context. For example, when reading a book, your attention shifts from word to word, with a greater focus on the current word and its surrounding words, while other parts of the page are processed with lower attention.

In the field of ML, attention mechanisms have been adapted to mimic and enhance this selective focus on relevant information. The introduction of attention mechanisms has significantly improved the performance of various deep learning models, particularly in tasks involving sequences or structured data. The fundamental idea behind attention mechanisms in ML is to enable models to weigh and prioritize different elements of the input data dynamically. This is achieved through the computation of attention weights, which determine how much focus or "*attention*" should be assigned to each element in the input sequence. The weights are computed based on learned parameters and the context of the current processing step.

The theoretical foundation of attention mechanisms in neural networks is rooted in the human cognitive process of selective attention. By adapting this concept to ML, attention mechanisms have transformed the field, enabling models to dynamically focus on relevant information within data sequences. This has led to significant improvements in various applications, making attention mechanisms a cornerstone of modern deep learning architectures. Understanding the cognitive origins of attention and its integration into ML is essential for appreciating the power and versatility of attention mechanisms in solving complex tasks. Attention mechanisms in deep learning can be broadly categorized into different types based on their characteristics and the way they allocate attention. These mechanisms are instrumental in enhancing the capabilities of neural networks by enabling them to selectively weigh and focus on specific parts of input data. Two primary types of attention mechanisms that have gained prominence are hard attention and soft attention.

3.5.1 Soft attention

Soft attention [103] is a mechanism in deep learning, widely used in various applications to selectively focus on specific elements of input data. It operates by assigning attention scores to elements and then calculating a context vector as a weighted sum of these elements.

Key components and workflow

Attention scores (α): For a given input sequence, soft attention computes attention scores (α) for each element. These scores represent the significance or relevance of each element to the task at hand.

$$\alpha_i = f(\text{query}, \text{key}_i).$$

Here, '*query*' represents the context or target information that guides attention, and '*key_i*' denotes the i -th element in the input sequence. The function ' f ' computes the attention scores.

Attention weights (ω): The attention scores are transformed into attention weights (ω) using a softmax function to ensure that they sum up to 1.

$$\omega_i = \frac{e^{\alpha_i}}{\sum_{j \text{ in input_sequence}} e^{\alpha_j}}.$$

The softmax function normalizes the attention scores, making them interpretable as probabilities.

Context vector (c): Finally, soft attention calculates the context vector (c) by taking a weighted sum of the input elements, where the weights are determined by the attention weights (ω).

$$c = \sum_{i \text{ in input_sequence}} (\omega_i \cdot \text{value}_i).$$

Here, '*value_i*' represents the value or content associated with the i -th element in the input sequence.

Significance

Before delving into practical applications, let's explore the significance of soft attention, understanding its pivotal role in enhancing the performance of various tasks by dynamically focusing on relevant information within input sequences. Soft attention's weighted information retrieval and differentiability make it a versatile and valuable tool in neural networks. Its adaptability and effectiveness in enhancing model performance continue to drive innovations across various domains, making it a cornerstone of modern deep learning. By intelligently selecting and processing information, soft attention contributes to the interpretability and efficiency of neural network models. Its mathematical formulation and application versatility ensure its continued prominence in the field of artificial intelligence.

Applications

Now, we turn our attention to practical applications, starting with a focus on how soft attention is instrumental in various domains.

1. **NLP:** Soft attention plays a crucial role in tasks like machine translation, where it helps models align words in the source and target languages effectively. In this context, 'query' could be the hidden state of the decoder, 'key _i' corresponds to the encoder's hidden states, and 'value _i' contains the encoded input words.
2. **CV:** Soft attention enhances image classification and object detection. In image classification, 'query' is derived from the current image features, 'key _i' represents the image regions, and 'value _i' contains the visual content. Soft attention can also be applied to text-to-image generation tasks, aligning words in descriptions with regions in images.
3. **Reinforcement learning:** Soft attention helps agents focus on relevant information when making decisions. In reinforcement learning, it aids in selecting important states or actions in a dynamic environment.

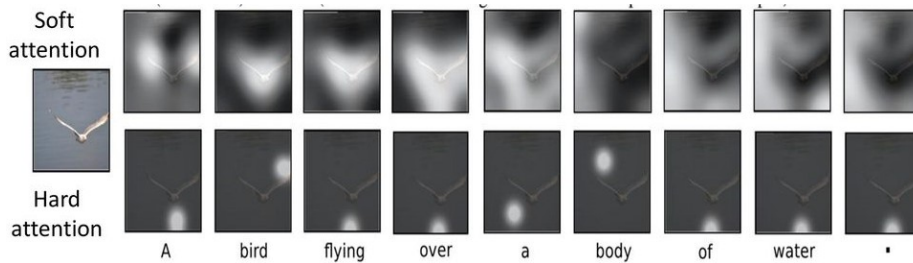


Figure 14: Examples of soft (top) and hard (bottom) attentions [38]

3.5.2 Hard attention

Hard attention [38] is a distinct approach within the spectrum of attention mechanisms, presenting a departure from the probabilistic weighting characteristic of soft attention. In contrast to soft attention, which allocates non-zero weights to all elements in the input sequence, hard attention selectively focuses on a subset of elements with non-zero weights. This deterministic selection process contributes to increased interpretability and is particularly well-suited for tasks that demand explicit and focused attention.

Key components and workflow

Stochastic selection: Within hard attention, the selection of elements follows a stochastic process, involving a sampling mechanism based on probability distributions. This introduces an element of randomness, distinguishing hard attention from its probabilistic counterpart.

Attention scores (α): Hard attention, akin to soft attention, computes attention scores (α) for each element. However, in the hard attention mechanism, these scores are transformed into a probability distribution, guiding the subsequent selection of elements based on this distribution. The calculation of attention scores (p_i) is given by the softmax

function:

$$p_i = \frac{\exp(\alpha_i)}{\sum_{j \in \text{input_sequence}} \exp(\alpha_j)}$$

Sampled index: In a departure from considering all elements, a single index is sampled from the probability distribution to determine the element that will be selectively attended to. This sampled index (sampled_index) becomes a pivotal factor in the subsequent computation.

Context vector (c): The context vector (c) is then computed based on the sampled index and the associated value (value_{sampled_index}). This contextualized representation is derived from the selectively attended element, emphasizing the deterministic focus intrinsic to hard attention:

$$c = \text{value}_{\text{sampled_index}}$$

Significance

Hard attention introduces a layer of interpretability to attention mechanisms by explicitly choosing a subset of elements for consideration. This explicitness proves advantageous in scenarios where a deterministic focus is desired, such as in image captioning tasks or when generating structured sequences.

While the sampling process in hard attention introduces a non-differentiability aspect, it offers a deliberate trade-off between interpretability and differentiability. The choice between these characteristics depends on the specific requirements of the task at hand, highlighting the nuanced application of hard attention mechanisms in diverse contexts.

Applications

Hard attention finds applications in various domains, particularly where explicit and deterministic focus is crucial. In IC, for example, hard attention can be employed to selectively attend to specific regions of an image, aiding in generating more precise and contextually relevant descriptions. Similarly, in tasks involving structured sequence generation, hard attention mechanisms can enhance the clarity and coherence of the generated sequences by focusing on key elements.

Hard attention presents a compelling alternative within the realm of attention mechanisms, introducing a deterministic selection process that aligns with interpretability needs in various applications. Its distinctive features make it a valuable tool, particularly in scenarios where explicit and focused attention is paramount.

3.5.3 Transformers

The most prominent example of attention mechanisms in ML is the transformer model introduced in [50]. The transformer architecture relies heavily on self-attention mechanisms to process sequences, making it highly effective in natural language processing tasks. In the Transformer, attention is computed between all pairs of input elements, allowing the model to capture complex dependencies and relationships within the data.

Transformers represent a groundbreaking architecture in the realm of NLP that has significantly advanced the state-of-the-art in various language-related tasks. Unlike traditional models that rely on recurrent or convolutional layers, transformers leverage

a self-attention mechanism to capture contextual information efficiently across input sequences. Introduced by Vaswani et al. [50], transformers have become the backbone of state-of-the-art models due to their inherent parallelization capabilities and ability to capture intricate relationships within sequences.

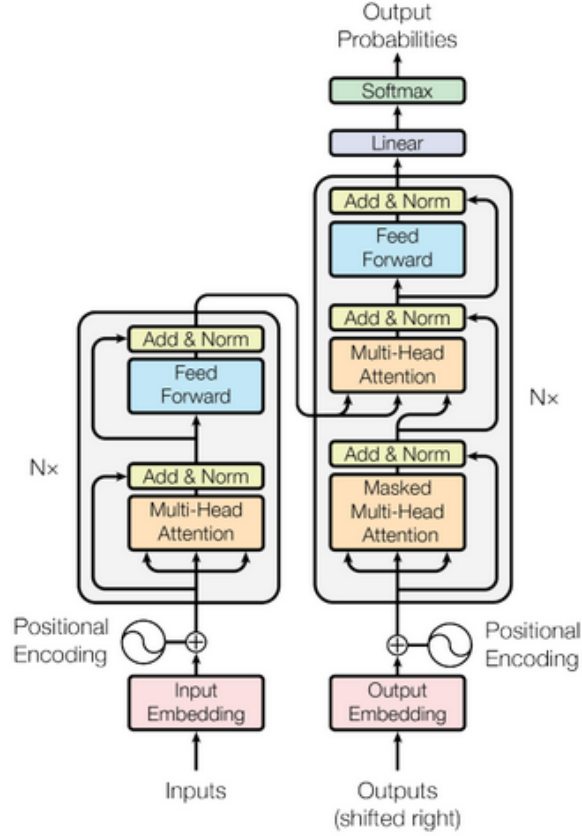


Figure 15: The transformer – model architecture [50]

Self-attention mechanism

The crux of the transformer architecture lies in the self-attention mechanism, enabling the model to weigh the importance of different words in a sequence dynamically. Unlike traditional approaches, where context modeling is sequential, self-attention allows for simultaneous consideration of all positions in the input sequence. This attention score computation is governed by the dot-product attention mechanism, fostering the model's capacity to capture contextual dependencies efficiently.

The self-attention mechanism is a pivotal component of the transformer architecture, enabling the model to weigh the importance of different elements in a sequence dynamically. This mechanism is fundamental to capturing long-range dependencies and contextual information efficiently. Let's delve into the mathematical equations that define the self-attention process.

Consider an input sequence $X = [x_1, x_2, \dots, x_n]$, where n is the length of the sequence.

1. **Calculation of attention scores (A_{ij}):** The attention score A_{ij} is computed using the scaled dot-product attention mechanism:

$$A_{ij} = \frac{e^{(X_i \cdot X_j^T)/\sqrt{d}}}{\sum_{k=1}^n e^{(X_i \cdot X_k^T)/\sqrt{d}}}$$

2. **Weighted sum (context vector S_j):** The weighted sum S_j represents the self-attention output at position j :

$$S_j = \sum_{i=1}^n A_{ij} \cdot X_i$$

3. **Vectorized form:** The attention scores and the context vector can be expressed in vectorized form. Let A be the matrix of attention scores:

$$A = \text{softmax} \left(\frac{X \cdot X^T}{\sqrt{d}} \right)$$

The context vector S can then be calculated as:

$$S = A \cdot X$$

This self-attention mechanism allows the model to dynamically adjust the importance of each element in the sequence based on its context, facilitating the capture of intricate dependencies in sequential data.

Multi-head attention

Transformers employ a multi-head attention mechanism to enhance their ability to focus on various aspects of the input sequence. By using multiple attention heads in parallel, the model can capture different patterns and relationships simultaneously. This parallelization contributes to the model's robustness and enables it to learn diverse contextual information.

Multi-head attention is a key component of the transformer architecture, allowing the model to capture different aspects of the input sequence simultaneously. It achieves this by utilizing multiple attention heads, each focusing on a distinct subspace of the input. Let's delve into the mathematical equations that define the multi-head attention mechanism.

Consider an input sequence $X = [x_1, x_2, \dots, x_n]$, where n is the length of the sequence.

1. **Single head attention:** The attention mechanism for a single head is given by the scaled dot-product attention mechanism:

$$A_{ij} = \frac{e^{(X_i \cdot X_j^T)/\sqrt{d}}}{\sum_{k=1}^n e^{(X_i \cdot X_k^T)/\sqrt{d}}}$$

The weighted sum S_j is calculated as:

$$S_j = \sum_{i=1}^n A_{ij} \cdot X_i$$

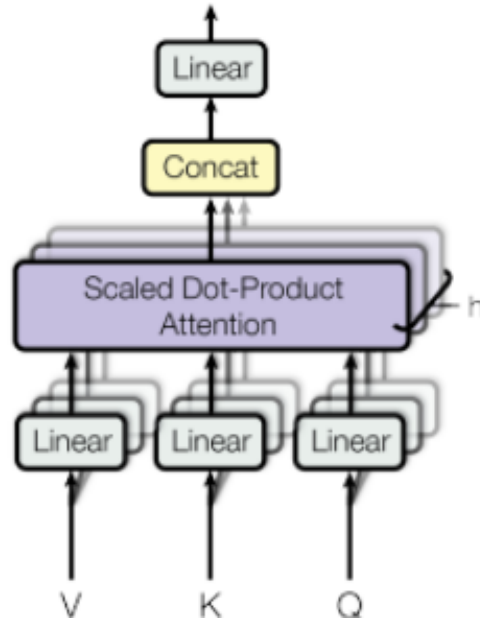


Figure 16: Multi-head attention

2. **Multiple heads:** For h attention heads, each head has its own set of learnable parameters (projection matrices W_Q^h, W_K^h, W_V^h). The outputs of all heads are concatenated and linearly transformed:

$$\text{MultiHead}(X) = \text{Concat}(\text{Head}_1, \text{Head}_2, \dots, \text{Head}_h)W_O$$

where $\text{Head}_h = \text{SingleHead}(XW_Q^h, XW_K^h, XW_V^h)$. The linear transformation W_O is applied to the concatenated outputs, allowing the model to learn how to combine information from different attention heads.

3. **Vectorized form:** The attention scores for all heads can be computed in a single matrix multiplication. Let Q , K , and V be the matrices obtained by concatenating the outputs of each head:

$$Q = \text{Concat}(Q_1, Q_2, \dots, Q_h)$$

$$K = \text{Concat}(K_1, K_2, \dots, K_h)$$

$$V = \text{Concat}(V_1, V_2, \dots, V_h)$$

The attention scores matrix A is then computed as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

The final output is obtained as the weighted sum of the values:

$$\text{MultiHead}(X) = A \cdot V$$

This mechanism enables the model to attend to different parts of the input sequence concurrently, capturing diverse relationships and enhancing the expressive power of the transformer architecture.

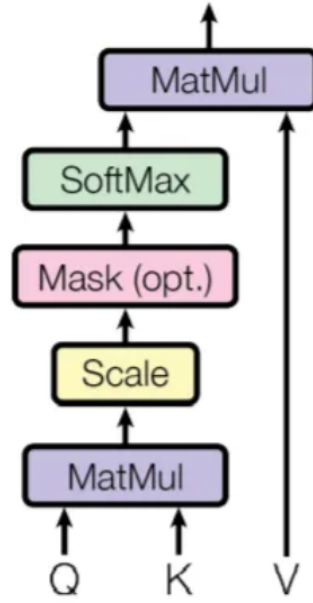


Figure 17: Scaled dot-product attention

Positional encoding

One challenge transformers overcome is the lack of inherent positional information. Unlike sequential models that inherently understand the order of elements, transformers treat input sequences as unordered sets. To address this, positional encodings are introduced, allowing the model to differentiate between elements based on their positions in the sequence. These positional encodings are typically added to the input embeddings, providing crucial information about the sequence order.

Positional encoding is crucial in transformer architectures to provide information about the position of tokens in a sequence. Since transformers do not inherently understand the order of elements, positional encoding is added to the input embeddings. This allows the model to distinguish between elements based on their positions in the sequence.

Consider an input sequence $X = [x_1, x_2, \dots, x_n]$, where n is the length of the sequence.

To incorporate positional information, positional encodings PE are added to the input embeddings X :

$$X_{\text{positional}} = X + PE$$

The positional encoding PE is calculated using the following equations:

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

Here, pos represents the position of the token, i is the dimension, and d is the dimensionality of the embeddings.

Vectorized form:

In a vectorized form, the entire positional encoding matrix PE is calculated for all positions and dimensions:

$$PE_{(pos,dim)} = \begin{cases} \sin\left(\frac{pos}{10000^{\frac{dim}{d}}}\right) & \text{if dim is even} \\ \cos\left(\frac{pos}{10000^{\frac{dim}{d}}}\right) & \text{if dim is odd} \end{cases}$$

This matrix is then added to the input embeddings X element-wise.

Positional encoding is essential for transformers to distinguish between different positions in the sequence and enable the model to capture sequential information effectively.

Feedforward neural networks

Following the self-attention mechanism, the transformer architecture incorporates FNN to introduce non-linearity and enable the learning of intricate patterns in the data. The feedforward layer allows the model to capture complex relationships and representations in a high-dimensional space, contributing to the overall expressive power of the architecture. They are a critical component of transformer architectures, introducing non-linearity and enabling the model to learn complex relationships within the data.

Consider an input sequence $X = [x_1, x_2, \dots, x_n]$, where n is the length of the sequence.

Single layer FFN:

The FFN operates independently on each position in the sequence. For a single layer FFN, the output at position i is computed as follows:

$$\text{FFN}(X_i) = \text{ReLU}(X_i W_1 + b_1) W_2 + b_2$$

Here, - X_i is the input at position i , - W_1 and W_2 are learnable weight matrices, - b_1 and b_2 are learnable bias vectors, - $\text{ReLU}(\cdot)$ is the Rectified Linear Unit activation function.

Vectorized form:

In a vectorized form, the entire sequence is processed in parallel using matrix operations. Let X be the matrix of input embeddings, W_1 and W_2 are the weight matrices, and b_1 and b_2 are the bias vectors:

$$\text{FFN}(X) = \text{ReLU}(XW_1 + B_1)W_2 + B_2$$

Here, B_1 and B_2 are matrices with repeated rows of b_1 and b_2 to match the dimensions of XW_1 and XW_2 .

The FFN introduces non-linearity to the model, allowing it to capture complex patterns within the input sequence.

Training

Transformers are trained using self-supervised learning, where a portion of the input sequence is masked, and the model is tasked with predicting the masked elements. This pre-training process allows the model to learn contextual information and general representations of the input data. This fosters the development of a robust understanding

of contextual information and semantic relationships within the data. The pre-trained transformer models can then be fine-tuned for specific downstream tasks.

Applications

Transformers have demonstrated exceptional performance across a wide range of NLP applications. Notable transformer-based models include BERT [98], GPT [104], and T5 [105]. These models have excelled in tasks such as machine translation, text summarization, and sentiment analysis. Transformers have achieved remarkable success across a spectrum of NLP applications and data augmentation. Models such as BERT, GPT, and T5 have set new benchmarks in tasks including sentiment analysis, machine translation, text summarization, and question-answering. Their versatility and performance have made them indispensable tools in the NLP practitioner's toolkit, facilitating breakthroughs in various language-related challenges.

Chapter 4

Methodology

The methodology chapter unveils our chosen methodology, highlighting our benchmark architecture, Faster R-CNN for object detection, and the significance of our dataset selection. We introduce the groundbreaking CutOver approach, delving into its pipeline and providing concise examples, laying the foundation for our experimental pursuits.

4.1 Benchmark Architecture

The *SAT* [38], plays a pivotal role in shaping the IC system. This architecture is selected due to its effectiveness in combining visual attention mechanisms with RNNs for caption generation. This architecture is renowned for its ability to dynamically focus on different regions of an image while generating textual descriptions. This is achieved through the integration of attention mechanisms. The attention mechanism allows the model to selectively attend to specific features or regions in the input image, assigning different weights to different areas based on their relevance to the current context in the caption generation process. This attention mechanism enhances the model's capability to capture intricate details and relationships within the visual input, leading to more contextually relevant and descriptive captions.

The choice of the SAT model signifies a commitment to leveraging advanced neural network architectures that excel in handling the complex interplay between visual and linguistic information. The attention mechanism within the SAT model is particularly advantageous in capturing fine-grained details in images, making it a suitable choice for the IC task. This methodology lays the foundation for a sophisticated and context-aware image captioning system that benefits from the nuanced insights offered by attention mechanisms.

4.2 Choice of Object Detector : Faster R-CNN

Facilitating the extraction of visual features from images, CutOver adopts the *Faster R-CNN* [48] object detection model. Renowned as a state-of-the-art architecture in object detection, Faster R-CNN excels in identifying objects and their spatial locations within images. This contextual information serves as a critical foundation for the subsequent caption generation process, ensuring a nuanced and informed approach to synthesizing image captions. In essence, CutOver represents a holistic and sophisticated augmentation strategy that intertwines cutting-edge techniques from both CV and NLP, aiming to significantly advance the performance of IC systems.

4.3 Datasets

The SAT model undergoes a two-stage training process to harness the power of pre-training and fine-tuning, strategically combining the strengths of the MS COCO [44] and VizWiz datasets [66]. Initially, the model is pre-trained on the MS COCO dataset, a widely recognized benchmark in CV. This pre-training phase equips the model with a foundational knowledge of fundamental visual and linguistic features present in diverse images and captions within the MS COCO dataset.

Following pre-training, the model undergoes fine-tuning on the VizWiz dataset, the target dataset specific to the application at hand. Fine-tuning is a crucial step where the model adapts its knowledge to the distinct characteristics and nuances of the VizWiz data. This process optimizes the model's performance for the challenges posed by the VizWiz dataset, ensuring it can effectively understand and generate captions that align with the intricacies of the real-world scenarios captured in VizWiz images.

CutOver employs a unique strategy by simultaneously augmenting both image and text data during fine-tuning. This simultaneous augmentation approach using the CutOver method significantly enhances the diversity of the training data, enabling the model to better handle variations, complexities, and unique features present in the VizWiz dataset. The augmentation technique introduces novel perspectives and linguistic variations, preparing the model for a more robust performance on the challenging VizWiz dataset.

The overarching goal of this approach is to adapt a pre-trained MS COCO model to excel on the more demanding VizWiz dataset. By leveraging the CutOver augmentation method during fine-tuning, the model aims to achieve superior results in caption generation. This sophisticated training strategy aligns the model's understanding of visual and linguistic elements with the intricacies of the VizWiz dataset, enhancing its capability to generate contextually relevant and diverse captions in real-world scenarios.

4.4 CutOver

The core inspiration for the CutOver methodology is rooted in a seminal paper [106] that delves into the landscape of DA techniques within NLP.

Within this survey, the authors discuss the potential and opportunities for simultaneous augmentation of both image and text data in the context of IC. This observation sparks the idea behind CutOver, aiming to incorporate and extend these insights into the realm of IC. The paper serves as a foundational motivation, suggesting that the integration of

simultaneous augmentation strategies for both modalities could lead to more robust and nuanced image captioning models. The recognition of untapped potential in concurrent augmentation becomes a pivotal motivator for CutOver, aspiring to incorporate and extend these insights into the realm of image captioning. This paper, serving as a foundational motivation, suggests that the integration of simultaneous augmentation strategies for both modalities holds promise for developing more robust and nuanced image captioning models, emphasizing the need to explore innovative approaches that exploit the synergies between image and text data in a unified manner.

4.4.1 CutOver description

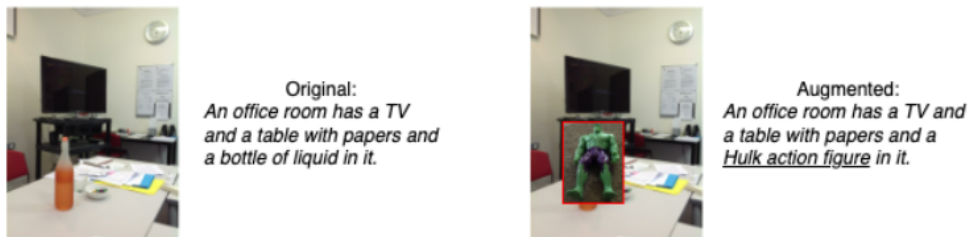


Figure 18: Joint Data Augmentation Example [65]

CutOver stands as an innovative DA method meticulously crafted for the unique requirements of IC systems. Distinguished by its joint approach, this method harmoniously blends two distinct augmentation techniques tailored for the CV and NLP modalities: CutMix and instance crossover, respectively. This synergistic combination aims to elevate the interaction between visual and linguistic elements, ultimately culminating in heightened image captioning performance.

In the realm of CV, CutOver capitalizes on the prowess of the CutMix [13] augmentation technique. Leveraging CutMix involves strategically replacing segments of one image with corresponding segments from another, thereby generating a blended, composite image. This process significantly enriches the model's robustness and generalization by exposing it to a more diverse array of visual data, fostering adaptability and resilience.

Turning to the NLP modality, CutOver introduces the Instance Crossover [31] technique. This NLP-centric augmentation method revolves around the strategic swapping or merging of textual components across captions, leading to the creation of novel and varied captions. The primary objective here is to amplify the linguistic diversity within the training data, enhancing the model's ability to understand and generate diverse and contextually relevant textual descriptions.

4.4.2 CutOver Pipeline

Figure 19 outlines the pipeline for the proposed augmentation method, CutOver. The following steps describes precisely how this augmentation method operates.

Step 1- Object detection and information collection: In this initial phase, we employ advanced object detection technology, specifically Faster R-CNN, to meticulously analyze an image. Our objective is to identify various objects within the image. Beyond mere identification, we collect crucial information about these objects, encompassing

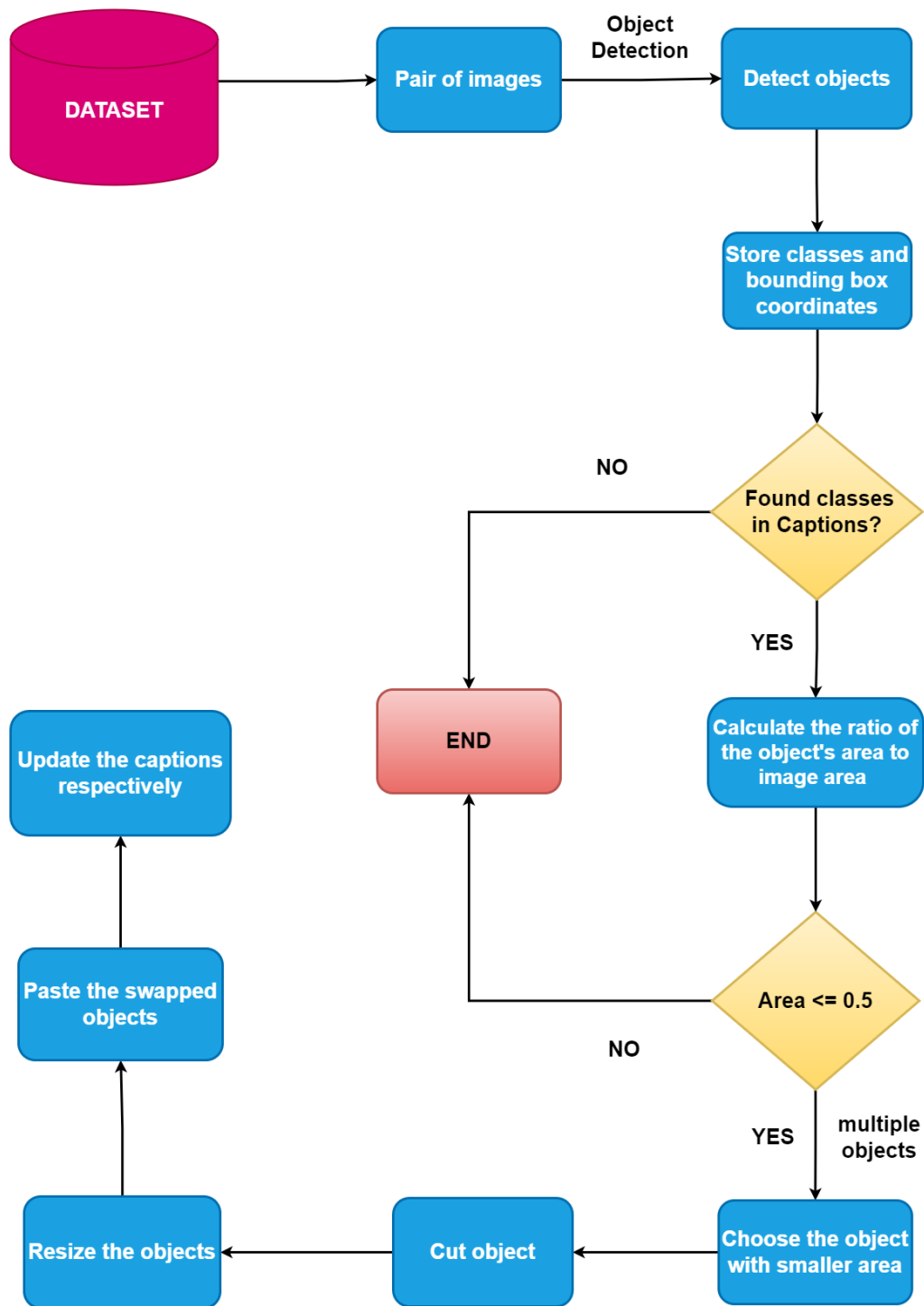


Figure 19: CutOver pipeline

their categories (classes) and the precise locations of their boundaries (bounding box coordinates). This meticulous information extraction, including object classes and spatial details, forms the foundational dataset for subsequent steps, laying the groundwork for a nuanced augmentation process.

Step 2- NLP Filtering: Moving to the second step, we integrate NLP. Here, we cross-reference the object classes identified in Step 1 with a set of five captions. By identifying matches between object classes and words in the captions, we filter out relevant information. This step streamlines our focus, ensuring that only pertinent object classes are considered for further processing, enhancing the synergy between visual and textual elements.

Step 3- Object size assessment and selection: In the third step, we introduce decision-making based on the size of the selected objects in relation to the entire image. Objects occupying less than or equal to 50% of the image area are considered for further action. In cases of multiple qualifying objects, prioritization is based on a smaller area size. This strategic assessment ensures the identification and selection of the most relevant objects for subsequent manipulation, optimizing the augmentation process.

Step 4- Replication for consistency: The fourth step replicates the entire process for a second image, ensuring a consistent and accurate approach to object detection, filtering, and selection. This step allows us to maintain uniformity in our methodology across different images, enhancing the reliability and comprehensiveness of our augmentation process.

Step 5- Object isolation (CutOut): Having identified and chosen our objects of interest in both images, we proceed to physically separate them from their respective backgrounds. This step involves "*cutting out*" the selected objects, effectively isolating them for further manipulation. This stage sets the groundwork for the subsequent exchange and rearrangement of objects between images.

Step 6- Object resizing for seamless integration: To prepare for the impending swap, we adjust the sizes of the cut objects. The goal is to ensure that these objects will seamlessly fit into each other's positions within the images. This resizing step optimizes the visual coherence of the final augmented images.

Step 7- Object position exchange: The seventh step introduces the exciting element of object position exchange. We swap the positions of the selected objects between the two images, providing a novel visual perspective and introducing dynamic changes to the overall composition.

Step 8- Caption update for consistency: To maintain consistency and coherence between the visual content and textual descriptions, the final step involves updating the captions. We modify the captions to accurately reflect the new positions and arrangements of the objects within the images, ensuring a holistic and integrated transformation of both visual and textual elements.

4.4.3 CutOver examples

In this section, we demonstrate the practical application and effectiveness of the CutOver augmentation method through examples (Figure 20). These highlight how CutOver strategically integrates image and text augmentation, improving the diversity and quality of training data for image captioning models. Each example provides a snapshot of the augmentation process, offering insights into transformations applied to both visual and textual modalities.



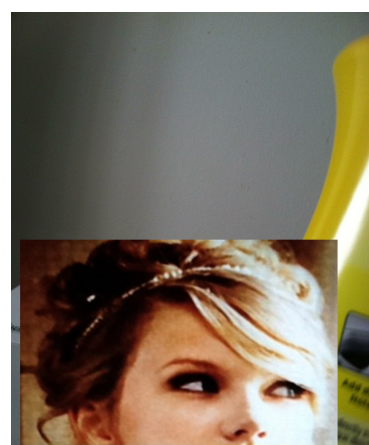
A plastic bottle of red colored spice mixture with a pop **heart** lid



A white cup of coffee with a **top** and the word love inside of it



A portrait of singer Taylor Swift with her **box** in a bun looking to the side



A yellow bottle of a chemical or cleaner next to a white **hair**



A person's arm over a **drawer** with a pattern



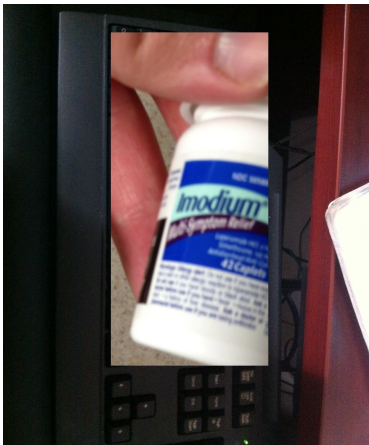
A wooden dresser with five **blankets** with two metal handles on each



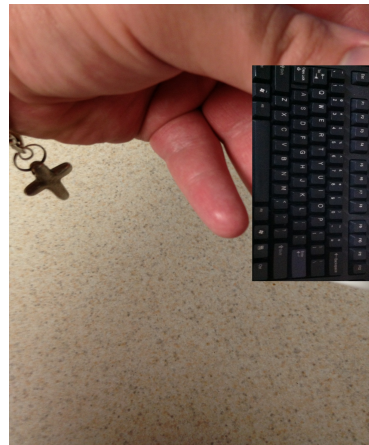
A small table **cup** on a bedroom table



A coffee **lamp** with a blue cow on it



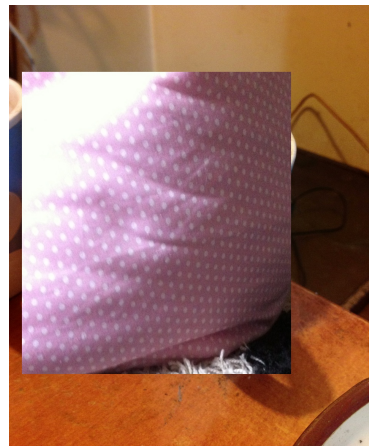
A computer **bottle** on an extended sliding bottle tray



A hand holding a **keyboard** of Imodium brand medication



A shaggy black and white carpet with a pink and white polka dotted **cup** on it



A blue and white coffee **pillow** partially filled sitting on top of a wooden surface

Figure 20: CutOver examples

4.5 Implementation details

The implementation hinges on the strategic selection of architectural frameworks that underpin the entire image captioning system. Leveraging PyTorch as the primary deep learning framework, the SAT model emerges as a linchpin for dynamic caption generation. This architecture seamlessly combines visual attention mechanisms with RNNs, enabling the model to dynamically focus on different regions of an image. PyTorch, with its flexibility and dynamic computation graph, proves instrumental in realizing the intricate workings of the SAT model.

Simultaneously, the Faster R-CNN model, powered by the torchvision library, takes center stage for object detection. Renowned as a state-of-the-art architecture, Faster R-CNN excels in identifying objects and their spatial locations within images. The contextual information extracted from this model, facilitated by the torchvision library, becomes paramount in laying the foundation for subsequent caption generation, ensuring a nuanced and informed synthesis of image captions.

4.5.1 Dataset preparation

The training process unfolds in two distinct stages, each serving a unique purpose in enhancing the model's understanding and adaptability. The initial pre-training phase occurs on the MS COCO dataset, a renowned benchmark in computer vision. This phase, facilitated by PyTorch's capabilities, imparts foundational knowledge about fundamental visual and linguistic features present in diverse images and captions within the MS COCO dataset.

Following pre-training, the model undergoes fine-tuning on the VizWiz dataset, a dataset tailored to real-world scenarios. Fine-tuning becomes a crucial step where the model adapts its knowledge to the distinct characteristics and nuances of the VizWiz data. An intriguing facet of our approach is the simultaneous augmentation of both image and text data using the CutOver method during fine-tuning. This innovative strategy, implemented with PyTorch's dynamic computation graph, significantly enhances the diversity of the training data, preparing the model to adeptly handle the intricacies of the VizWiz dataset. The dataset includes images and corresponding captions. Captions are preprocessed by adding '<start>' and '<end>' tokens, and padding to ensure uniformity in length.

4.5.2 Inputs to the model

Images: Processed using pre-trained ImageNet models available in PyTorch's torchvision module. Images are resized to 256x256, converted to a Float tensor, and normalized.

Captions: Encoded as integer tensors, with padding and additional tokens for sequence generation.

Caption lengths: Represented as Int tensors, indicating the actual length plus 2 for '<start>' and '<end>' tokens.

4.5.3 Data pipeline

In the data pipeline, we create essential elements that will be used in the subsequent stages of our process. We generate HDF5 files, which act as containers for storing image representations, providing a structured way to organize and access this visual data. Additionally, we produce JSON files that contain encoded captions along with information about their respective lengths. These files serve as a convenient and efficient means of handling textual data associated with the images.

4.5.4 Encoder

The encoder is a crucial component of our system, leveraging a pre-trained ResNet-101 model from PyTorch's torchvision library. To adapt it to our specific requirements, we discard the last two layers of the network. Furthermore, we introduce an AdaptiveAvgPool2d() layer for effective image encoding. For further customization, fine-tuning is applied to convolutional blocks 2 through 4, optimizing the model's performance to better suit our unique needs. This process ensures that the encoder efficiently extracts meaningful features from the input images, laying a solid foundation for subsequent stages in our workflow.

4.5.5 Model hyperparameters

In the initial training phase of the model, key hyperparameters are employed to guide the learning process. The model undergoes training for 10 epochs with an embedding dimension (emb_dim) of 512, attention and decoder dimensions (attention_dim and decoder_dim) set to 512, and a dropout rate of 0.5 for regularization. The choice of device for computations is determined dynamically, utilizing the GPU if available. The training is configured with a batch size of 32, one worker for data loading, and learning rates of $1e-4$ for the encoder (if fine-tuning) and $4e-4$ for the decoder. Gradients are clipped at an absolute value of 5.0, and a regularization parameter (alpha_c) of 1.0 is employed for 'doubly stochastic attention'. Here, doubly stochastic implies that the attention weights are not only conditioned on the image features but are also influenced by the generated words in the captioning process [38].

Following this initial phase, the fine-tuning process is initiated by setting the fine-tune encoder to *True*. Subsequently, the model undergoes additional training, extending either for 10 or 15 epochs, as determined by the experimental configuration. This fine-tuning step enables the model to adapt its knowledge to the specific characteristics of the dataset, enhancing its performance on the target task. The combination of these training stages, characterized by specific hyperparameter settings, aims to iteratively refine the model's understanding and improve its overall effectiveness in handling the given data.

4.5.6 Attention mechanism

The attention network is composed of linear layers and activations. It processes the encoded image and the hidden state from the decoder, generating weights through a softmax operation

4.5.7 Decoder

The decoder, named `DecoderWithAttention`, receives the encoded image and flattens it. LSTM is employed with manual iteration over timesteps to facilitate the attention mechanism. Dynamic batching is achieved using an LSTM cell, processing only valid timesteps without iterating over pads.

4.5.8 Model training

Our training process is multifaceted, combining various techniques to enhance model performance.

Cross-entropy loss and regularization: We employ cross-entropy loss during training to generate word sequences. To ensure attentive focus, a doubly stochastic regularization enforces attention weights to sum to 1 across timesteps. Notably, the loss computation strategically excludes padded regions.

Early stopping with BLEU metric: Validation is a critical phase, where we leverage the BLEU evaluation metric to assess the quality of generated captions against reference captions. The training process is designed to halt if the BLEU score deteriorates, prioritizing model generalization even if the loss metric shows improvement.

Staged training approach: For optimal results, we recommend a staged training approach. Initially, we suggest training only the Decoder without fine-tuning the Encoder. This sequential training strategy allows the model to grasp language patterns before incorporating complex image features. Subsequently, fine-tuning the Encoder enhances the model's ability to integrate visual information.

Teacher forcing for realistic inference: During the validation phase, we implement a technique known as Teacher Forcing. This mimics real-world inference conditions by using the ground truth (actual) words as input during the decoding process. This approach helps stabilize training and encourages more accurate caption generation.

In summary, the training process involves iterations over the dataset, and adjusting the model's parameters to improve performance. The SAT model learns to generate accurate and contextually relevant captions for a given input image through pre-training and fine-tuning. This comprehensive training strategy ensures that our model is not only proficient in capturing linguistic nuances but also adept at effectively incorporating visual information for accurate and contextually relevant image captions.

Chapter 5

Experiments and Results

The Experiments and Results chapter delves into the performance evaluation of IC models, offering insights into dataset characteristics, chosen evaluation metrics, and the model's effectiveness in generating contextually relevant descriptions. The analysis aims to provide a concise overview of experimental outcomes for further exploration.

5.1 Data description

This section comprehensively outlines the MS COCO and VizWiz dataset, covering its overview, sources, and size, along with details on how the dataset is split for training, validation, and testing purposes.

5.1.1 MS COCO dataset overview



Figure 21: Examples of MS COCO dataset [44]

The MS COCO [44] dataset is a widely recognized and extensively used collection of images tailored to address challenges in various CV tasks. Originating from Microsoft,

this dataset is a benchmark in the field, fostering research and development in areas such as IC, object detection, and segmentation. Unlike some datasets that are artificially curated, MS COCO images authentically capture diverse scenes with a focus on contextual relationships among objects.

The uniqueness of MS COCO lies in its depiction of real-world scenarios, offering images that encompass a wide range of objects and activities in everyday life. Unlike staged datasets, MS COCO provides a realistic representation of scenes, presenting an invaluable resource for developing and evaluating models that understand both individual objects and their contextual placement.

For this research, the MS COCO dataset offers a rich and challenging set of images. Generating captions for these images involves surpassing mere object recognition, requiring an understanding of intricate contextual details and the relationships between objects. Thus, introducing the MS COCO dataset is pivotal in establishing the foundation for exploring innovative and comprehensive image captioning methodologies.

Data sources and size

The MS COCO dataset, a cornerstone in this research, consists of images collected from a diverse array of sources, capturing real-world scenarios in a multitude of contexts. Unlike some datasets that focus on specific domains, MS COCO presents a broad and comprehensive view of everyday life, making it a robust and versatile dataset for various computer vision applications.

Encompassing a substantial collection of 200,000 images, the MS COCO dataset provides a diverse and representative sample of visual data. Each image encapsulates a unique perspective on the multifaceted nature of scenes encountered in everyday life. By reflecting the rich diversity of real-world scenarios, MS COCO challenges models to generalize effectively across a wide range of contexts, contributing to the development and evaluation of models capable of understanding and describing complex visual scenes.

Train/test split

The dataset is systematically divided into three subsets: the training set, validation set, and test set.

Training set: This foundational set comprises 120,000 images and 600,000 captions, forming the backbone of the model's learning process. The extensive training set exposes the model to diverse visual patterns, enabling it to learn intricate relationships inherent in the real-world scenarios captured by MS COCO.

Validation set: With 10,000 images and 50,000 captions, the validation set plays a crucial role in fine-tuning the model and assessing its performance during training. This set serves as an intermediate checkpoint, allowing for adjustments and optimizations before the model encounters the independent test set.

Test set: We have opted to utilize the validation set as opposed to the original test set. This decision is attributed to the unavailability of captions for the images in the original test set. Consequently, to maintain consistency and adhere to the experimental protocol, the validation set serves as a substitute for evaluating our proposed methods and ensuring rigorous testing in the absence of captioned data for the original test images.

5.1.2 VizWiz dataset overview



Figure 22: Examples of VizWiz dataset [66]

The VizWiz dataset [66] is a specialized and unique collection of images designed to address the challenges faced by visually impaired individuals in their daily lives. The dataset primarily serves the purpose of fostering research and development in the field of image captioning, particularly focusing on generating descriptive textual content for images captured by individuals with visual impairments.

Its distinctiveness lies in the real-world and uncontrolled settings of the images, reflecting the authentic experiences of visually impaired photographers. Unlike many conventional datasets, VizWiz images are not staged or manipulated; instead, they represent genuine moments captured by users facing visual challenges. This authenticity makes the VizWiz dataset an invaluable resource for studying and developing image captioning models that can effectively interpret and describe the content of images taken in diverse and dynamic environments.

For this thesis research, the VizWiz dataset presents a compelling and challenging set of images. The captions generated for these images are expected to surpass mere visual content recognition, incorporating an understanding of contextual nuances and practical scenarios encountered in everyday life by individuals with visual impairments. Therefore, introducing the VizWiz dataset is crucial in setting the stage for my exploration into innovative and inclusive image captioning methodologies.

Data sources and size

The VizWiz dataset, a pivotal component of this research, is exclusively composed of images taken by visually impaired individuals in genuine real-world settings. Its distinctiveness lies in the authenticity of moments captured by users facing visual challenges, presenting a unique and invaluable perspective. Unlike conventional datasets, VizWiz images authentically portray dynamic and uncontrolled environments, offering a rich source of real-life scenarios. The dataset encapsulates user-captured perspectives, providing a nuanced view aligned with the experiences and challenges of visually impaired individuals. By incorporating these unique characteristics, VizWiz not only challenges the traditional paradigm of image datasets but also becomes an essential resource for the development and exploration of image captioning methodologies that comprehend both visual content and the contextual intricacies of real-world scenarios.

The VizWiz dataset, a crucial element in this research endeavor, exhibits notable scale and significance. Comprising a substantial volume of images, the dataset stands as a

comprehensive collection, capturing diverse scenarios encountered by visually impaired individuals. VizWiz encompasses 39,118 images, each offering a unique perspective into the daily lives and challenges faced by individuals with visual impairments. This scale not only contributes to the richness and diversity of the dataset but also provides a substantial foundation for training and evaluating image captioning models. The considerable size of the VizWiz dataset ensures that the models developed and tested on this data can robustly generalize across a wide array of real-world situations, enhancing their adaptability and effectiveness in providing meaningful captions for images taken in diverse and dynamic environments.

Train/test split

The dataset is strategically partitioned into three subsets: the training set, validation set, and test set.

Training set: Comprising a substantial 23,431 images and 117,155 captions, the training set forms the backbone of the model’s learning process. This extensive collection exposes the model to diverse visual and linguistic elements, enabling it to grasp intricate patterns and relationships inherent in the real-world scenarios captured by VizWiz.

Validation set: With 7,750 images and 138,750 captions, the validation set plays a pivotal role in fine-tuning the model and validating its performance during training. The carefully curated validation set serves as an intermediate checkpoint, allowing for adjustments and optimizations before the model encounters the independent test set.

Test set: The test set, consisting of 8,000 images and 40,000 captions, serves as the ultimate benchmark for evaluating the model’s generalization capabilities. Kept separate during the entire training process, this set gauges the model’s ability to generate meaningful captions on previously unseen data, simulating real-world scenarios beyond its training environment.

5.2 Evaluation metrics

This section provides a concise overview of key evaluation metrics used to assess the performance of image captioning models. It covers Bleu Scores, METEOR, ROUGE_L, CIDEr, and SPICE, offering insights into different aspects of caption quality.

5.2.1 BLEU metric

In the field of natural language processing and machine translation, the BLEU score is a widely used metric for automatically assessing the quality of machine-generated translations. Introduced by Papineni et al. [26] in their seminal paper on machine translation evaluation, BLEU has become a standard benchmark for comparing the output of translation systems.

The BLEU score measures the similarity between a machine-generated translation and one or more reference translations. It ranges from 0 to 1, with 1 indicating a perfect match between the machine-generated and reference translations. The score is computed based on the precision of the n-grams (contiguous sequences of n items, typically words) in the machine-generated translation compared to the reference translations.

The BLEU score is calculated using the following formula:

$$\text{BLEU} = \text{BP} \times \exp \left(\sum_{n=1}^N \frac{1}{N} \log(P_n) \right)$$

Here, N represents the maximum n-gram order considered, P_n denotes the precision of n-grams, and BP is the brevity penalty.

Precision (P_n) calculation

The precision of n-grams (P_n) is computed as:

$$P_n = \frac{\text{Number of matching n-grams in the machine-generated translation}}{\text{Total number of n-grams in the machine-generated translation}}$$

Brevity penalty calculation

BP is determined based on the length of the machine-generated translation compared to the length of the closest reference translation. It is defined as follows

$$\text{BP} = \begin{cases} 1 & \text{if length of machine-generated translation} \geq \text{length of closest reference} \\ \exp \left(1 - \frac{\text{length of closest reference}}{\text{length of machine-generated translation}} \right) & \text{otherwise} \end{cases}$$

5.2.2 ROUGE metric

The ROUGE [29] metric is a set of evaluation measures widely used for assessing the quality of machine-generated text, particularly in tasks such as text summarization and machine translation. ROUGE comprises several metrics, each designed to capture different aspects of the quality of machine-generated text. Key ROUGE metrics include:

1. **ROUGE-N (N-gram overlap):** Measures the overlap of n-grams between the generated and reference text. ROUGE-1 considers unigrams, ROUGE-2 considers bigrams, and so on.
2. **ROUGE-L (Longest common subsequence):** Measures the longest common subsequence between the generated and reference text. This metric is sensitive to word order and useful for evaluating sentence-level coherence.
3. **ROUGE-W (Weighted N-gram overlap):** Similar to ROUGE-N, but assigns different weights to different n-grams based on their lengths. This gives more importance to longer shared sequences.
4. **ROUGE-S (Skip-bigram co-occurrence):** Measures the co-occurrence of skip-bigrams, n-grams with gaps between words, capturing partial semantic similarity.
5. **ROUGE-SU (Skip-bigram co-occurrence with unigram):** Extends ROUGE-S by including unigrams in the skip-bigram co-occurrence calculation.

ROUGE score calculation

The ROUGE score is computed by comparing the n-grams or subsequences in the generated text to those in the reference text. Precision, recall, and F1 score are commonly used to quantify the overlap:

$$\text{Precision} = \frac{\text{Number of overlapping n-grams in generated text}}{\text{Total number of n-grams in generated text}}$$

$$\text{Recall} = \frac{\text{Number of overlapping n-grams in generated text}}{\text{Total number of n-grams in reference text}}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.2.3 METEOR metric

The METEOR [27] metric is an evaluation measure commonly used in the field of machine translation. It was introduced aiming to provide a comprehensive evaluation that considers both precision and recall, along with the explicit matching of synonyms.

METEOR components

The METEOR metric incorporates several components to measure the quality of machine-generated translations:

1. **Unigram precision and recall:** METEOR calculates precision and recall based on the matching unigrams between the generated and reference text.
2. **Synonymy:** METEOR uses WordNet to identify synonyms, giving partial credit for synonymous words in the generated and reference text.
3. **Stemming:** The metric accounts for stemming, and recognizing morphological variations of words.
4. **Chunks:** METEOR considers matching chunks of words, giving credit for correctly aligned sequences.
5. **Exact matching:** METEOR also measures the percentage of exact word matches.

METEOR score calculation

The METEOR score is computed by combining precision and recall components with penalty terms for fragmentation and stemming. The formula for METEOR is as follows:

$$\text{METEOR} = \text{Precision} \times \left(1 - \beta \times \frac{\text{fragmentation penalty}}{\text{fragmentation penalty} + \gamma \times \text{stemming penalty}} \right) \times (1 - \delta \times \text{exact penalty})$$

Here, β , γ , and δ are tunable parameters to adjust the impact of fragmentation, stemming, and exact matching penalties.

5.2.4 CIDEr metric

The CIDEr [28] metric is a widely used evaluation measure for assessing the quality of image captions generated by automatic image description systems. It was introduced to capture consensus among human annotators when evaluating the similarity between generated and reference captions.

CIDEr components

CIDEr comprises several components that contribute to its evaluation of image captions:

1. **TF-IDF weighted similarity:** CIDEr utilizes the TF-IDF weighting scheme to measure the similarity between n-grams in generated and reference captions.
2. **Consensus:** CIDEr emphasizes consensus by rewarding diverse but relevant captions. It considers the consensus in human annotations, acknowledging variations in language.
3. **n-gram matching:** Similar to BLEU, CIDEr evaluates the precision of n-grams (typically up to four-grams) between the generated and reference captions.

CIDEr score calculation

The CIDEr score is calculated based on the TF-IDF weighted similarity of n-grams and the consensus measure. The formula for CIDEr is as follows:

$$\text{CIDEr} = \text{mean} \left(\frac{\text{TF-IDF weighted precision for each n-gram order}}{\text{TF-IDF weighted precision for each n-gram order} + \text{consensus penalty}} \right)$$

The consensus penalty discourages generating captions that align with only a subset of reference captions.

5.2.5 SPICE metric

The SPICE [30] metric is an evaluation measure designed to assess the quality of image captions generated by automatic image description systems. It was introduced with a focus on capturing the specificity of generated captions by comparing them to human references.

SPICE components

SPICE comprises several components that contribute to its evaluation of image captions:

1. **Scene graph representation:** SPICE leverages scene graphs to represent the relationships between objects, attributes, and their connections in an image.
2. **Semantic role labeling:** SPICE employs semantic role labeling to identify the roles of objects and their attributes within a sentence.

3. **Specificity measure:** SPICE evaluates the specificity of a caption by comparing the generated and reference captions in terms of their scene graph structures and semantic roles.

SPICE score calculation

The SPICE score is calculated based on the precision and recall of specific scene graph structures and semantic roles in the generated and reference captions. The formula for SPICE is as follows:

$$\text{SPICE} = \text{H-mean}(\text{precision, recall for scene graphs}) \times \text{H-mean}(\text{precision, recall for semantic roles})$$

The SPICE metric considers both the structural aspects of scene graphs and the semantic roles played by objects and attributes in the captions.

5.3 Evaluation

In this section, we present the outcomes of our comprehensive experiments, each conducted over 20 iterations to ensure robustness and capture potential variations.

Experiment 1: Evaluation involved benchmarking against SOTA VizWiz, VizWiz without data augmentation, and VizWiz with CutOver data augmentation.

Experiment 2: Rigorous comparison of CutOver augmentation against traditional image augmentation methods, including blur, randomBrightnessContrast, and coarseDropout.

Experiment 3: Methodical evaluation contrasting CutOver augmentation with text augmentation techniques, such as bert substitute, random swap, random delete, and synonym replacement.

Experiment 4: Comparative analysis between CutOver augmentation and joint augmentation methods, specifically blur + random swap.

Experiment 5: Assessment of performance scores on varying percentages (100%, 50%, 20%, 10%) of the VizWiz dataset without any augmentation.

Experiment 6: Examination of performance scores on varying percentages (100%, 50%, 20%, 10%) of the VizWiz dataset with the application of CutOver augmentation.

For each experiment, we provide a detailed analysis of results, including evaluation metrics such as BLEU, METEOR, ROUGE_L, CIDEr, and SPICE. The subsequent chapter (Chapter 6) delves into a comprehensive discussion of the obtained results, interpreting their implications and insights.

5.3.1 Experiments

Table 2 provides a comprehensive evaluation of the performance of three different models: SOTA VizWiz, VizWiz without DA, and VizWiz with CutOver DA. The scores across various metrics, including Bleu_1, Bleu_2, Bleu_3, Bleu_4, METEOR, ROUGE_L, CIDEr, and SPICE, offer nuanced insights into the impact of augmentation strategies on the generated captions quality and diversity.

Scores	SOTA VizWiz	VizWiz (Without DA)	VizWiz (With CutOver DA)
Bleu_1	0.725	0.593	0.344
Bleu_2	0.539	0.404	0.235
Bleu_3	0.388	0.271	0.157
Bleu_4	0.274	0.179	0.105
METEOR	0.222	0.163	0.122
ROUGE_L	0.501	0.402	0.314
CIDEr	0.807	0.355	0.282
SPICE	0.170	0.099	0.082

Table 2: Performance comparison of SOTA VizWiz [107], VizWiz (without DA), and VizWiz (With CutOver DA)

Observations and analysis

SOTA VizWiz Model: The SOTA VizWiz model is considered for reference, showcasing superior performance in various metrics.

BLEU Scores (Bleu_1, Bleu_2, Bleu_3, Bleu_4): VizWiz with DA demonstrates comparable or superior performance to VizWiz with CutOver DA across all Bleu scores. The CutOver augmentation strategy may not contribute significantly to improved n-gram matching compared to traditional augmentation.

METEOR: VizWiz with DA shows similar or better METEOR performance compared to VizWiz with CutOver DA. The augmentation strategy, particularly CutOver, may not necessarily positively influence METEOR metrics.

ROUGE_L: VizWiz with DA exhibits comparable or better ROUGE_L scores than VizWiz with CutOver DA. The CutOver augmentation strategy may not significantly contribute to better recall in longer sequences.

CIDEr: VizWiz with DA outperforms or shows similar performance to VizWiz with CutOver DA in terms of CIDEr scores. The CutOver augmentation strategy may not positively impact the diversity and informativeness of captions compared to traditional augmentation.

SPICE: VizWiz with DA achieves comparable or higher SPICE scores compared to VizWiz with CutOver DA. The CutOver augmentation strategy may not positively influence SPICE metrics related to descriptive and specific caption generation.

Scores	CutOver	Image Augmentation Methods		
		Blur	RandomBrightnessContrast	CoarseDropout
Bleu_1	0.344	0.592	0.593	0.592
Bleu_2	0.235	0.402	0.403	0.401
Bleu_3	0.157	0.270	0.269	0.267
Bleu_4	0.105	0.180	0.177	0.177
METEOR	0.122	0.161	0.162	0.160
ROUGE_L	0.314	0.403	0.404	0.402
CIDEr	0.282	0.353	0.357	0.350
SPICE	0.082	0.096	0.098	0.097

Table 3: Performance comparison of CutOver augmentation vs image augmentation methods (Blur, RandomBrightnessContrast, CoarseDropout)

Table 3 offers a detailed performance evaluation of various augmentation methods: CutOver, Blur, RandomBrightnessContrast, and CoarseDropout. The comparison encompasses key metrics such as Bleu_1, Bleu_2, Bleu_3, Bleu_4, METEOR, ROUGE_L, CIDEr, and SPICE. Notably, CutOver is assessed against different image augmentation techniques. This comprehensive analysis provides valuable insights into the effectiveness of each augmentation strategy, enabling a nuanced understanding of their impact on the generated caption’s quality and diversity.

Observations and analysis

BLEU Scores (Bleu_1, Bleu_2, Bleu_3, Bleu_4): The CutOver DA approach exhibits lower Bleu scores (Bleu_1, Bleu_2, Bleu_3, Bleu_4) compared to the specific Image Augmentation Methods (Blur, RandomBrightnessContrast, CoarseDropout). This indicates that, on average, the CutOver method generates captions that have less overlap with the ground truth across different n-grams.

METEOR: The METEOR score for the CutOver DA method is lower than that of the individual image augmentation methods. METEOR takes into account precision, recall, and alignment, suggesting that CutOver may not perform as well in terms of these metrics.

ROUGE_L: The ROUGE_L score, which measures the longest common subsequence of words, is lower for CutOver DA compared to the specific image augmentation methods. This indicates that the CutOver approach may result in captions with fewer common words with the ground truth.

CIDEr: The CIDEr score for CutOver is lower than that of the individual image augmentation methods. CIDEr considers consensus-based metrics, implying that CutOver may not capture the consensus as effectively as the specific image augmentation methods.

SPICE: The SPICE score for CutOver is lower compared to the image augmentation methods. SPICE evaluates the semantic content of the generated captions, suggesting that Joint DA may not capture semantic information as effectively as the specific image DA methods.

Among the specific image augmentation methods, **Blur** consistently demonstrates the highest scores across various evaluation metrics. This suggests that, in this particular experiment, blur may be a more effective image augmentation technique for improving the performance of image captioning models compared to RandomBrightnessContrast and CoarseDropout.

Scores	CutOver	Text Augmentation Methods			
		Bert Substitute	Random Swap	Random Delete	Synonym Replacement
Bleu_1	0.344	0.584	0.589	0.584	0.579
Bleu_2	0.235	0.395	0.401	0.398	0.388
Bleu_3	0.157	0.264	0.269	0.267	0.257
Bleu_4	0.105	0.176	0.178	0.179	0.170
METEOR	0.122	0.160	0.162	0.160	0.157
ROUGE_L	0.314	0.399	0.404	0.404	0.393
CIDEr	0.282	0.346	0.354	0.349	0.334
SPICE	0.082	0.097	0.096	0.096	0.095

Table 4: Performance comparison of CutOver augmentation vs text augmentation methods (bert substitute, random swap, random delete, and synonym replacement)

Table 4 provides a detailed evaluation of different text augmentation methods, including CutOver, bert substitute, random swap, random delete, and synonym replacement, alongside CutOver. The metrics considered, such as Bleu_1, Bleu_2, Bleu_3, Bleu_4, METEOR, ROUGE_L, CIDEr, and SPICE, offer a comprehensive perspective on the impact of these methods on caption quality and diversity.

Observations and analysis

BLEU Scores (Bleu_1, Bleu_2, Bleu_3, Bleu_4): The CutOver approach records lower Bleu scores (Bleu_1, Bleu_2, Bleu_3, Bleu_4) when compared to specific text augmentation methods (bert substitute, random swap, random delete, synonym replacement). This implies that, on average, the CutOver method generates captions with less overlap with the ground truth across various n-grams.

METEOR: The METEOR score for the CutOver method is observed to be lower than that of the individual text augmentation methods. METEOR, considering precision, recall, and alignment, suggests that CutOver may not perform as effectively in terms of these metrics.

ROUGE_L: The ROUGE_L score, measuring the longest common subsequence of words, is lower for CutOver when compared to specific text augmentation methods. This indicates that the CutOver approach may result in captions with fewer common words with the ground truth.

CIDEr: The CIDEr score for CutOver DA is found to be lower than that of the individual text augmentation methods. CIDEr, considering consensus-based metrics, implies that CutOver may not capture consensus as effectively as the specific text augmentation methods.

SPICE: The SPICE score for CutOver DA is lower compared to text augmentation methods. SPICE, evaluating the semantic content of generated captions, suggests that CutOver may not capture semantic information as effectively as the specific text augmentation methods.

Among the text augmentation methods, *random swap* consistently demonstrates higher scores across various metrics. This suggests that random swap has a notable impact on improving the quality and diversity of generated captions compared to other text augmentation methods.

Scores	CutOver	Joint Augmentation Methods
		Blur+Random Swap
Bleu_1	0.344	0.592
Bleu_2	0.235	0.402
Bleu_3	0.157	0.270
Bleu_4	0.105	0.180
METEOR	0.122	0.161
ROUGE_L	0.314	0.403
CIDEr	0.282	0.353
SPICE	0.082	0.096

Table 5: Performance comparison of CutOver augmentation vs joint augmentation methods (blur + random swap)

Table 5 presents a detailed examination of model performance across various evaluation metrics, comparing the CutOver augmentation method with a joint augmentation ap-

proach that combines blur and random swap techniques.

Observations and analysis

BLEU Scores (*Bleu_1, Bleu_2, Bleu_3, Bleu_4*): The Bleu scores (*Bleu_1, Bleu_2, Bleu_3, Bleu_4*) exhibit a consistent pattern. The model with the joint augmentation (blur + random swap) method outperforms the CutOver augmentation across all four BLEU metrics. This suggests that the combined blur and random swap augmentation strategy enhances the model’s ability to generate captions with higher overlap with ground truth across different n-grams.

METEOR: METEOR scores follow a similar trend, with the joint augmentation (blur + random swap) method demonstrating higher scores compared to the CutOver augmentation. This implies that the combined augmentation approach performs better in terms of precision, recall, and alignment metrics.

ROUGE_L: ROUGE_L scores also show consistent improvement with the joint augmentation method, indicating that the longest common subsequence of words is more effectively captured when blur and random swap augmentations are combined.

CIDEr: CIDEr scores demonstrate superior performance for the joint augmentation (blur + random swap) method compared to CutOver augmentation. This suggests that the consensus-based metric is more effectively addressed by the combined augmentation strategy.

SPICE: SPICE scores follow a similar pattern, with the joint augmentation (blur + random swap) method outperforming CutOver augmentation. This indicates that the semantic content of the generated captions benefits from the combination of blur and random swap augmentations.

The joint augmentation method (*blur + random swap*) consistently outperforms the CutOver approach across various metrics, indicating its effectiveness in enhancing the model’s caption generation capabilities. This highlights the potential benefits of combining blur and random swap augmentations for improved image captioning model performance.

Scores	Scores on n% of VizWiz Dataset without augmentation			
	VizWiz (100%)	VizWiz (50%)	VizWiz (20%)	VizWiz (10%)
Bleu_1	0.593	0.577	0.561	0.533
Bleu_2	0.404	0.391	0.366	0.346
Bleu_3	0.271	0.260	0.236	0.218
Bleu_4	0.179	0.173	0.151	0.134
METEOR	0.163	0.155	0.146	0.140
ROUGE_L	0.402	0.391	0.377	0.366
CIDEr	0.355	0.321	0.267	0.225
SPICE	0.099	0.093	0.079	0.076

Table 6: Performance scores on n% of VizWiz dataset without augmentation

Table 6 presents a comprehensive analysis of the model’s performance across different evaluation metrics on varying percentages of the VizWiz dataset without any augmentation. The scores are reported for four subsets of the dataset, representing 100%, 50%, 20%, and 10% of the original data.

Observations and analysis

BLEU Scores (*Bleu_1*, *Bleu_2*, *Bleu_3*, *Bleu_4*): As the dataset size decreases, there is a noticeable decline in BLEU scores (*Bleu_1*, *Bleu_2*, *Bleu_3*, *Bleu_4*), indicating that the model’s ability to generate captions that align with ground truth diminishes with reduced training data.

METEOR: Similar trends are observed in the METEOR scores, which consider precision, recall, and alignment. The model’s performance decreases as the dataset size shrinks, suggesting a correlation between training data volume and METEOR scores.

ROUGE_L: ROUGE_L scores, measuring the longest common subsequence of words, also exhibit a downward trend with decreasing dataset size. This implies that the model’s proficiency in generating captions with common words diminishes when trained on smaller subsets.

CIDEr: CIDEr scores, evaluating consensus-based metrics, decrease as the dataset size reduces. This indicates that the model’s ability to capture consensus in generated captions is impacted by the training data volume.

SPICE: SPICE scores, assessing semantic content, follow a similar pattern, showing a decline as the dataset size decreases. This suggests that semantic information in the generated captions is influenced by the amount of training data available.

Scores	Scores on n% of VizWiz Dataset with CutOver Augmentation			
	CutOver (100%)	CutOver (50%)	CutOver (20%)	CutOver (10%)
Bleu_1	0.344	0.116	0.170	0.143
Bleu_2	0.235	0.054	0.090	0.071
Bleu_3	0.157	0.029	0.052	0.036
Bleu_4	0.105	0.017	0.031	0.020
METEOR	0.122	0.040	0.056	0.048
ROUGE_L	0.314	0.138	0.183	0.170
CIDEr	0.282	0.028	0.059	0.034
SPICE	0.082	0.017	0.025	0.019

Table 7: Performance Scores on n% of VizWiz Dataset with CutOver Augmentation

Table 7 provides a comprehensive overview of model performance across various evaluation metrics, specifically focusing on the impact of the CutOver augmentation method on different subsets (n%) of the VizWiz Dataset.

Observations and analysis

BLEU Scores (*Bleu_1*, *Bleu_2*, *Bleu_3*, *Bleu_4*): The scores for BLEU metrics exhibit a non-linear trend with CutOver augmentation across different percentages of the VizWiz dataset. Notably, the scores at 50% of the dataset are observed to be lower than those at 20%, indicating a non-monotonic relationship between dataset size and the performance of the CutOver augmentation method. Additionally, the scores at 10% are lower, suggesting a potential saturation or diminishing returns at lower dataset sizes.

METEOR: Similarly, METEOR scores demonstrate a non-linear pattern, with a decrease at 50% compared to 20% of the dataset. This non-monotonic relationship suggests that the CutOver augmentation’s influence on precision, recall, and alignment may be influenced by factors beyond dataset size. The scores at 10% further underscore the potential challenges of using smaller datasets.

ROUGE_L: The ROUGE_L scores follow a non-linear trend, with the scores at 50% being lower than those at 20% of the dataset. This indicates that the longest common subsequence of words in generated captions may not strictly improve with an increase in dataset size when applying the CutOver augmentation method. The scores at 10% also show a decrease, highlighting potential limitations at very low dataset sizes.

CIDEr: The CIDEr scores exhibit a non-monotonic relationship, with a decrease at 50% compared to 20% of the dataset. This suggests that the consensus-based metrics considered by CIDEr may not consistently benefit from a larger training dataset when utilizing the CutOver augmentation approach. The scores at 10% provide insights into the challenges of achieving high consensus with very limited training data.

SPICE: SPICE scores also demonstrate a non-linear pattern, with a decrease at 50% compared to 20% of the dataset. This implies that the ability of the CutOver augmentation to capture semantic information may not strictly follow a linear improvement with the size of the training data. The scores at 10% highlight the potential limitations in capturing semantic diversity with extremely small datasets.

One striking observation is that, contrary to conventional expectations, the scores are higher for the 20% dataset compared to the 50% dataset. This unexpected trend suggests that the specific composition and diversity of the 20% dataset may contribute to the augmentation strategy's effectiveness in enhancing model performance. Several factors could contribute to this phenomenon:

Dataset composition: The 20% dataset may contain instances that align more favorably with the CutOver augmentation method, leading to improved performance. The subset of examples in the 20% dataset might better showcase the benefits of the augmentation strategy.

Data diversity: Despite its smaller size, the 20% dataset might retain a sufficient level of diversity, enabling the model to capture a broader range of patterns and variations during training. In contrast, the 50% dataset may be larger but lack the same diversity, limiting the model's ability to generalize effectively.

Overfitting vs. generalization: The larger dataset (50%) might introduce challenges related to overfitting, where the model may start memorizing examples rather than learning generalized patterns. The smaller dataset (20%) may strike a balance, avoiding overfitting while still having enough examples for meaningful learning.

Augmentation impact: The CutOver augmentation method might be particularly effective or complementary to the examples in the 20% dataset, introducing beneficial variations in the data that help the model learn more robust representations.

Randomness in training: Inherent randomness in the training process, especially with techniques like dropout or stochastic gradient descent, can contribute to performance differences between different dataset subsets.

Chapter 6

Discussion and Future Works

The Discussion and Future Works chapter delves into the analysis and insights derived from the study, providing a comprehensive discussion, and then outlining potential avenues for future research.

6.1 Discussion

In this section, we conduct a detailed comparative analysis of generated captions, examining the influence of the CutOver augmentation strategy on image captioning. Table 8 displays the examples of the images along with their groundtruth captions, generated captions (without augmentation), and generated captions (with CutOver)

In the first example, both the model without augmentation and the one with CutOver augmentation exhibit a deviation from the groundtruth caption, transitioning from *"tomato soup"* to *"cream of mushroom soup"*. This shift introduces inaccuracies in the generated captions, emphasizing the need for precision in capturing subtle details.

The second example exposes discrepancies in color details, with the model without augmentation erroneously mentioning a *"brown t-shirt"*, not evident in the image or the groundtruth captions. The model with CutOver augmentation worsens this by introducing an *"orange eagle"*, a color detail absent in both the visual content and the provided groundtruth captions.

The third example highlights the strengths of the model without augmentation, aligning well with the groundtruth and accurately describing the scene. However, the model with CutOver augmentation introduces *"unknown tokens"* (UNK), presenting challenges for certain images and indicating the potential need for further refinement in handling diverse scenarios.

In the fourth example, the model without augmentation offers a high-level description, correctly identifying the image as a close-up of a product label. Conversely, the model with CutOver augmentation introduces an inaccurate element, referencing a *"woman"* in the caption, which is notably absent in both the visual content and the provided groundtruth captions.




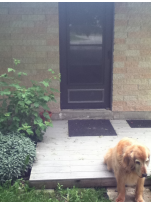
Image	GroundTruth Captions	Generated Captions (Without Augmentation)	Generated Captions (With CutOver)
	A person holding a can of Campbell's tomato soup	A person is holding a can of cream of mushroom soup.	A person holding a can of Campbell's cream of mushroom soup
	A blue shirt with the word Lions written in big letters.	A close up of a brown t-shirt with a dark background.	A close up of a t-shirt with an orange eagle on the front.
	A TV mounted in the corner of a room with a program on it.	A room with a window and a TV on the wall.	[UNK] room with a [UNK] [UNK] and a [UNK].
	A green, yellow, and red card detailing enrollment information.	A close up of a label for a product for the product.	A screenshot of a green and white background with a picture of a woman on it.
	A Golden Retriever sitting on a deck in a garden.	A picture of a house with a window behind it.	[UNK] dog is standing on the ground in a [UNK].

Table 8: Comparison of groundtruth and generated captions (without augmentation and with CutOver augmentation)

The fifth example presents notable distinctions between the generated captions and the ground truth. The ground truth accurately describes a *"Golden Retriever sitting on a deck in a garden."* In contrast, the non-augmented model deviates significantly, describing a *"house with a window behind it."* However, it's noteworthy that the CutOver-augmented model, despite introducing "[UNK]" tokens, successfully captures the main element of the scene – a *"dog standing on the ground."* This highlights a scenario where CutOver has effectively bridged the gap between the model's output and the ground truth, showcasing instances where the augmentation strategy has contributed positively to the model's performance.

These comparisons provide insights into the impacts of augmentation strategies on the precision and fidelity of generated captions. These analyses offer valuable insights into the influence of augmentation strategies on the precision and fidelity of generated captions. Each example is accompanied by its corresponding image, providing a comprehensive view of the impact of augmentation techniques on the model's captioning performance.

6.1.1 Analysis

With a comprehensive overview of image examples complemented by their respective ground truth captions, generated captions (without augmentation), and generated captions (with CutOver), it is imperative to delve into the details influencing the caption generation process. This analysis aims to uncover and elucidate the potential factors contributing to the observed disparities in caption quality and content.

1. Model is learning specific phrases

Phrases	Number of occurrences	
	Training dataset	Generated captions
a person is holding	867	445
a white piece of paper	129	222
can of food	637	190
top of a table	604	256
a computer screen	704	587

Table 9: Occurrences of phrases in training dataset and generated captions

In the context of our IC task, a crucial aspect of our model's learning process involves assimilating specific phrases from the training dataset. As an illustrative example, let's examine Table 9. The phrase *"a person is holding"* appeared 867 times within our meticulously annotated training dataset, representing a considerable frequency of occurrence.

Upon subjecting our model to rigorous evaluation, a noteworthy observation emerged: the model tended to redundantly generate the same phrase. More precisely, *"a person is holding"* surfaced 445 times in the captions produced by our model. This substantial variance in frequency between the training dataset and the generated captions signifies a distinct disparity in the model's behavior.

The recurrent generation of a specific phrase implies a limitation in the model's capacity to diversify its output effectively. Rather than encapsulating the rich contextual nuances present in the images, the model tends to rely on a restricted set of phrases.

This phenomenon underscores a crucial area for improvement in our image captioning model, as it indicates a potential shortcoming in its ability to adapt and generate varied, contextually nuanced captions.

To further illustrate this pattern, additional examples were investigated. For instance, the phrase "*a white piece of paper*" appeared 129 times in the training dataset but was generated 222 times by the model. Similarly, "*can of food*" occurred 637 times during training, yet the model produced it 190 times in captions. "*Top of a table*" was present 604 times in training but surfaced 256 times in the generated captions.

The consistent repetition of specific phrases across various examples underscores the urgency to address this tendency in our model. Exploring avenues to ameliorate this pattern is imperative for elevating the overall quality and diversity of the captions generated by our model. By addressing this challenge, we aspire to enhance the model's proficiency in understanding and encapsulating diverse visual contexts, thereby contributing to the broader goal of advancing the effectiveness of image captioning systems.

2. Semantic relationship is not captured

Groundtruth captions	CutOver captions
The front of two <i>houses</i> with four antenna towers in the background.	The front of two <i>bottles</i> with four antenna towers in the background.
The front of a partially full <i>bottle</i> of wine is shown sitting on a wood kitchen table.	The front of a partially full <i>clock</i> of wine is shown sitting on a wood kitchen table.
An image of a person standing on carpet showing their bare <i>foot</i> .	An image of a person standing on carpet showing their bare <i>spoon</i> .
A person is holding a <i>bottles</i> of pills in their lap.	A person is holding a <i>house</i> of pills in their lap.
The barcode from a bottle of <i>water</i> with a white label.	The barcode from a bottle of <i>leg</i> with a white label.

Table 10: Comparison of groundtruth captions and CutOver captions

One notable challenge observed in the analysis of the generated captions produced by the CutOver model pertains to the failure to capture semantic relationships evident in the groundtruth captions. The groundtruth captions exhibit a detailed understanding of the depicted scenes, encapsulating specific objects and their inherent semantic connections. However, the CutOver captions fall short of preserving these semantic relationships, leading to instances where the generated descriptions deviate significantly from the original context. This observation is further elucidated in the analysis presented in Table 10.

For instance, in the groundtruth caption, "*The front of two houses with four antenna towers in the background*", the model accurately identifies and describes the presence of houses in the scene. However, in the corresponding CutOver caption, the model erroneously replaces "*houses*" with "*bottles*", indicating a failure in maintaining the semantic relationship between the visual content and the generated description.

Similarly, in another example, the groundtruth caption highlights "*a person standing on carpet showing their bare foot*", precisely capturing the semantic connection between the person and their foot. In contrast, the CutOver model generates a caption where the person is portrayed as standing on carpet showing their "*bare spoon*", indicating a clear departure from the semantic context established in the groundtruth.

This discrepancy underscores a significant limitation in the CutOver model’s ability to grasp and retain semantic relationships present in the images. Addressing this issue is crucial for improving the model’s capability to generate contextually relevant and semantically coherent captions, ultimately enhancing the overall quality and fidelity of the generated descriptions. Strategies such as refining the training process or exploring alternative model architectures may be explored to mitigate this specific shortcoming and enhance the model’s semantic understanding.

6.1.2 Insights

1. Data compatibility: CutOver may encounter compatibility issues with the VizWiz dataset, known for containing images captured by blind individuals in real-world settings. VizWiz’s unique characteristics, stemming from its unconventional image sources and diverse real-world scenarios, introduce challenges for CutOver. The nature of images taken by blind individuals may include aspects such as less clarity, varying compositions, and unexpected visual perspectives. These specific characteristics, inherent to the VizWiz dataset, make it distinct from traditional datasets and may pose challenges for CutOver in effectively generating captions that align with the diverse content of the images. It might need adaptations or adjustments to accommodate the unique challenges posed by VizWiz, and this may involve refining the augmentation strategy to handle less clear images and unconventional compositions. The consideration of dataset-specific challenges is crucial for ensuring that CutOver is not only compatible with VizWiz but also capable of enhancing the dataset’s data efficiency by generating meaningful and relevant captions in real-world contexts.

2. Overly aggressive augmentation: This concern arises from the possibility that the CutOver method might be introducing excessive noise during the data augmentation process. This noise, represented by extreme variations or distortions in the augmented data, has the potential to impede the coherence of captions generated by the Show, Attend, and Tell model. Specifically applied to the diverse and real-world images in the VizWiz dataset, captured by blind individuals, CutOver must strike a delicate balance. The challenge lies in enhancing the model’s ability to generalize across various visual scenarios without introducing distortions that hinder the model’s capacity to interpret augmented data effectively. The need for caution in the augmentation process is crucial, ensuring that CutOver contributes to the model’s robustness rather than introducing excessive noise that could lead to a decrease in overall performance.

3. Data size: The consideration of data size in the application of CutOver to the VizWiz dataset is crucial, given the unique requirements of this dataset, which includes images taken by blind individuals in real-world settings. To optimize the benefits of CutOver, it may be advantageous to utilize a larger and more diverse dataset. A larger dataset provides a more comprehensive representation of the diverse scenarios within VizWiz images, allowing CutOver to learn and adapt to a broader range of visual patterns and contextual variations. Given the distinctive capture conditions and content of VizWiz, a larger dataset enhances the generalization capabilities of CutOver, enabling it to effectively handle the diverse and real-world nature of the images. The inclusion of more extensive data not only addresses the unique challenges of VizWiz but also serves as a robust foundation for training, potentially leading to the generation of higher-quality captions. Therefore, the consideration of data size becomes pivotal in ensuring that CutOver is adequately exposed to the complexities inherent in the VizWiz dataset, facilitating improved performance and caption quality.

4. Model compatibility: The evaluation of model compatibility between CutOver and the Show, Attend, and Tell architecture is pivotal, considering that the distinctive characteristics of the model might influence its receptiveness to the augmentation method. Show, Attend, and Tell rely on attention mechanisms to generate captions, emphasizing the importance of scrutinizing how well CutOver aligns with this attention-driven approach. Given that CutOver introduces variations and crossovers in both image and text modalities, the interplay with the attention mechanisms and overall model structure warrants careful examination. The model's capability to attend to relevant features during caption generation may be affected by the nature and extent of variations introduced by CutOver. Consequently, a comprehensive assessment is necessary to determine whether CutOver complements or conflicts with the underlying mechanisms of the Show, Attend, and Tell model, and whether adjustments to either the augmentation method or the model architecture are needed for optimal synergy. This compatibility study aims to uncover insights into how CutOver can enhance the model's attention mechanisms, thereby contributing to the generation of more accurate and contextually relevant captions. Understanding this interaction is crucial for making informed decisions on potential modifications to improve overall model compatibility and performance.

6.2 Future Works

1. Explore complex IC Model: A pivotal avenue for improvement lies in the exploration of a more advanced IC model to enhance the quality of text data generated by CutOver. Specifically, the investigation involves exploring the utilization of a more advanced or intricate IC model. Given its dual influence on both image and text modalities, a focused effort is directed toward refining the textual aspect. This exploration involves delving into SOTA NLP models, renowned for capturing intricate language patterns and nuances with precision. Additionally, the consideration extends to the potential development of custom-designed IC models tailored to the specific demands of the image captioning task facilitated by CutOver. The overarching goal is to elevate the quality of text data, ensuring that the augmented captions not only exhibit improved coherence, semantic richness, and linguistic diversity but also align with the cutting-edge advancements in NLP. Incorporating a more advanced IC model holds the promise of enhancing CutOver's language generation capabilities, producing captions that are not only contextually relevant but also exhibit refined linguistic nuances. This exploration signifies a commitment to continuous innovation within the Instance Crossover component of CutOver, aiming to leverage the latest NLP advancements or bespoke models for more effective and nuanced language generation during the augmentation process.

2. Enhance image quality: A critical focus in advancing the effectiveness of CutOver involves a deliberate effort to enhance the quality of images within the dataset. This improvement strategy encompasses various approaches, including preprocessing techniques, denoising procedures, and the incorporation of higher-resolution images. By implementing these measures, the objective is to ensure that CutOver operates on a foundation of high-quality visual data. Preprocessing techniques may involve methods such as contrast adjustment or normalization to standardize image features. Denoising procedures aim to reduce unwanted artifacts or distortions that could hinder the model's ability to discern key visual elements. The integration of higher-resolution images addresses the finer details within the visual content, providing a more comprehensive and detailed input for CutOver. The overarching goal is to optimize the quality of the input images, enabling CutOver to generate captions that not only benefit from improved

visual clarity but also capture more nuanced and contextually relevant information. This enhancement in image quality aligns with the broader objective of refining the input data to facilitate more accurate and coherent caption generation by CutOver.

3. Add constraints to CutOver: In the pursuit of refining the augmentation process, a strategic enhancement for CutOver involves the introduction of additional conditions or constraints to the method. This entails implementing specific rules or guidelines designed to govern the augmentation process, thereby imposing structure and ensuring that the augmented captions exhibit logical coherence and contextual relevance. These constraints could encompass a range of considerations, such as linguistic consistency, adherence to grammatical structures, or alignment with contextual themes present in the dataset. By introducing these constraints, the aim is to steer CutOver towards generating augmented captions that not only align with the semantic context of the images but also adhere to predefined criteria for linguistic quality. The design of constraints acts as a safeguard, preventing the generation of captions that may lack meaningful connections or veer away from the intended context. This strategic integration of constraints contributes to a more controlled and purposeful augmentation, fostering the production of augmented data that is not only diverse but also maintains logical and contextual fidelity. The consideration of constraints within the CutOver methodology reflects a nuanced approach to data augmentation, emphasizing the importance of incorporating semantic and contextual guidelines to enhance the overall quality and meaningfulness of the generated captions.

4. Semantic relationship preservation: A critical avenue for advancing the capabilities of CutOver involves a dedicated exploration of techniques aimed at preserving and enhancing the semantic relationships between objects in images during the augmentation process. This intricate task may encompass strategies to identify and selectively swap objects within an image, ensuring that the resulting augmented captions maintain, or ideally, strengthen the semantic context portrayed in the original image. Techniques for semantic relationship preservation go beyond traditional augmentation methods, delving into the nuanced understanding of object interactions and contextual relevance within the visual content. By identifying objects with inherent semantic connections, such as a person holding an object, or the relationship between different elements in a scene, CutOver can be fine-tuned to ensure that swaps maintain or enhance these relationships. This exploration aligns with the broader goal of not just diversifying the dataset but also enriching it with augmented instances that capture the inherent semantics of the scenes. The augmentation process, guided by semantic relationship preservation, thus contributes to the generation of captions that reflect a deeper understanding of the interconnectedness of objects within the visual context. Through this strategic integration, CutOver evolves beyond conventional augmentation approaches, aiming to produce augmented data that not only varies in content but also retains and amplifies the underlying semantic richness within the images.

5. Change object detection model: A pivotal strategy to enhance the efficacy of CutOver involves a systematic exploration of various object detection models. This entails experimenting with a diverse array of models, including more advanced or domain-specific variants, to identify the most suitable option for providing accurate and contextually relevant object annotations. The choice of the object detection model plays a foundational role in the augmentation process, influencing the precision with which objects are identified and annotated within the images. By considering advanced or domain-specific models, the aim is to elevate the accuracy of object annotations, ensuring that CutOver operates on a foundation of highly reliable and detailed object information. The integration of a superior object detection model aligns with the overarching goal of refining

the input data, facilitating more precise and meaningful augmentations. This strategic experimentation acknowledges the dynamic nature of object detection advancements and seeks to leverage the latest models that offer enhanced capabilities in capturing object features and relationships within complex scenes. The outcome of this exploration is anticipated to be a more effective and contextually aware CutOver, capable of generating augmented data that not only reflects a diverse range of scenes but also benefits from the improved accuracy of object annotations provided by the chosen detection model.

Chapter 7

Conclusion

The concluding chapter serves as the culmination of this comprehensive thesis endeavor, bringing together key insights, contributions, and reflections. This chapter not only revisits the overarching objectives set forth at the onset of the study but also delves into a synthesis of findings, shedding light on the implications of the research within the specific domain. As we navigate through the conclusion, we will not only summarize the main discoveries but also explore avenues for future research, thereby solidifying the broader impact of this work on the field of the IC system.

1. IC objective: the primary objective is to undertake the intricate task of generating human-like descriptions for images. This involves moving beyond the conventional understanding of images as visual entities and delving into the realm of linguistic interpretation. The fundamental goal of IC is to craft meaningful and contextually relevant textual descriptions that encapsulate and communicate the essence of the visual content within a given image. The process entails more than merely identifying and labeling objects within an image; it aims to capture the nuanced relationships, intricate details, and the overall narrative presented by the visual elements. By generating human-like descriptions, IC bridges the gap between the raw visual data and a linguistic understanding of that visual information. In essence, the objective is to create a seamless fusion of the visual and linguistic modalities, enabling machines to not only recognize objects but also articulate them in a manner akin to how humans describe and comprehend visual scenes. This ambitious objective finds application in various domains, from aiding accessibility for the visually impaired to enhancing human-computer interaction and enriching multimedia content with informative and engaging descriptions.

2. Role of DA: DA plays a pivotal role in the IC process, serving as a crucial component in optimizing data efficiency. At its core, DA involves the deliberate introduction of variations and diversity into the training dataset by applying various transformations to the existing images and captions. This intentional manipulation of data is essential for several key reasons.

Firstly, DA contributes to robust model training by exposing the model to a more extensive and varied set of examples. This helps prevent overfitting, a common challenge in machine learning where a model becomes too tailored to the specifics of the training data and struggles to generalize well to new, unseen data. By introducing augmented data

during training, the model learns to adapt to a broader range of scenarios, improving its ability to handle diverse real-world situations.

Secondly, the artificial expansion of the training dataset through DA techniques ensures that the model encounters a more comprehensive representation of the inherent variability in visual content and linguistic expressions. This diversity is particularly crucial in IC, where images may exhibit variations in lighting, composition, and object arrangements. Augmenting the dataset allows the model to become more resilient to these variations, ultimately enhancing its robustness and performance when faced with novel or challenging images.

Furthermore, DA aids in improving the model's generalization capabilities across different images. As it learns from an augmented and diversified dataset, the model becomes adept at extracting meaningful features and patterns that are more transferable to a wide range of image scenarios.

3. Exploration of IC architectures and DA techniques: The research involves a thorough exploration of various IC architectures and DA techniques. This exploration encompasses understanding the strengths and limitations of different IC models and augmentation strategies, laying the groundwork for informed decision-making.

4. Introduction of CutOver - A novel joint DA method: In a pivotal moment within the conclusion, the research introduces CutOver, an innovative and novel joint DA method meticulously crafted for IC systems. CutOver stands out for its unique approach, strategically combining techniques from both CV and NLP domains, thereby fostering a symbiotic relationship between visual and textual modalities.

4.1 Core principles of CutOver

CV Integration (CutMix): CutOver incorporates a CV augmentation technique known as CutMix. This involves replacing portions of one image with corresponding portions of another, creating a blended or mixed image. By leveraging CutMix, CutOver introduces diversity and complexity in the visual domain, exposing the model to a rich tapestry of visual variations. This integration is crucial for enhancing the model's ability to adapt to diverse image compositions and scenarios.

NLP (Instance Crossover): In tandem with the visual augmentation, CutOver embraces an NLP technique termed Instance Crossover. This involves swapping or merging textual elements between captions, generating new and diverse textual descriptions. By intertwining linguistic variations with visual transformations, CutOver aims to enrich the linguistic diversity of the training dataset, ensuring the model is well-equipped to handle the intricacies of varied linguistic expressions associated with different visual scenes.

4.2 Synergy between visual and textual modalities

CutOver's distinctive contribution lies in its ability to forge a synergy between the visual and textual modalities. Unlike traditional DA methods that often focus solely on images or captions independently, CutOver recognizes the interdependence of these modalities in IC. The simultaneous augmentation of both visual and textual elements creates a more holistic and nuanced learning experience for the model, aligning with the inherent complexity of real-world scenarios captured in images. **Unique Advantages of CutOver:** The introduction of CutOver represents a departure from conventional augmentation methods, offering a novel perspective on addressing the challenges of IC. The joint nature of CutOver allows for a more seamless integration of visual and textual data, potentially leading to more coherent and contextually relevant caption generation.

4.3 Strategic positioning in conclusion

Positioned as a central point in the conclusion, the introduction of CutOver signifies a deliberate and innovative attempt to enhance the augmentation strategies employed in IC. Its incorporation embodies a strategic move towards refining the model's adaptability, diversity, and performance through a unique blend of visual and textual augmentations. This introduction sets the stage for a nuanced evaluation and analysis of CutOver's impact on the overall effectiveness of IC models.

5. Comprehensive experimental examination: The conclusion incorporates a detailed examination of experimental outcomes. This includes the implementation of baseline strategies and the application of robust evaluation metrics to measure the effectiveness of the CutOver method in comparison to other approaches. Results and findings are thoroughly analyzed for insights.

6. Analysis of CutOver performance: A critical aspect of the conclusion involves a transparent analysis of why CutOver might have fallen short of expected improvements in IC models. This evaluation includes considerations of challenges, limitations, and potential areas for refinement in future iterations.

7. Walk-through of potential future enhancements: The conclusion provides a forward-looking perspective, outlining a systematic walk-through of potential future enhancements for refining the CutOver method in IC systems. This involves identifying areas of improvement, addressing limitations, and considering innovative approaches to enhance the overall performance of the model.

Bibliography

- [1] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [2] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VI-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- [3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2019.
- [4] Chengge Li, István Fehérvári, Xiaonan Zhao, Ives Macedo, and Srikar Appalaraju. Seetek: Very large-scale open-set logo recognition with text-aware metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2544–2553, 2022.
- [5] Di Qi, Lin Su, Jianwei Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *ArXiv*, abs/2001.07966, 2020.
- [6] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.
- [7] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.
- [8] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021.
- [9] Mareike Hartmann, Aliki Anagnostopoulou, and Daniel Sonntag. Interactive machine learning for image captioning. *The AAAI-22 Workshop on Interactive Machine Learning*, 2022.
- [10] Terrance DeVries and Graham Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, 2017.
- [11] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

- [12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.
- [13] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.
- [14] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.
- [15] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.
- [16] Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N Metaxas. Selfnorm and crossnorm for out-of-distribution robustness. 2020.
- [17] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.
- [18] William Yang Wang and Diyi Yang. That’s so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563, 2015.
- [19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.
- [20] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.
- [21] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [22] Claude Coulombe. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*, 2018.
- [23] Mehdi Regina, Maxime Meyer, and Sébastien Goutal. Text data augmentation: Towards better detection of spear-phishing emails. *ArXiv*, abs/2007.02033, 2020.
- [24] Ge Yan, Yu Li, Shu Zhang, and Zhenyu Chen. Data augmentation for deep learning of judgment documents. In *Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9*, pages 232–242. Springer, 2019.
- [25] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

- [26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [27] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [28] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [30] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.
- [31] Franco Martín Luque. Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. In *IberLEF@SEPLN*, 2019.
- [32] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.
- [33] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.
- [34] Himanshu Sharma and Devanand Padha. A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artificial Intelligence Review*, 56:1–43, 04 2023.
- [35] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR 2011*, pages 1601–1608, 2011.
- [36] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.
- [37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [41] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015.
- [42] Kyunghyun Cho, Bart Merriënboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.
- [43] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, pages 1735–1780, 1997.
- [44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, page 740–755, 2014.
- [45] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, pages 67–78, 2014.
- [46] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)*, page 853–899, 2013.
- [47] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.
- [48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1137–1149, 2017.
- [49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *31st International Conference on Neural Information Processing Systems*, page 6000–6010, 2017.
- [51] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10575–10584, 2020.
- [52] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022.

- [53] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [54] Rajarshi Biswas, Aditya Mogadala, Michael Barz, Daniel Sonntag, and Dietrich Klakow. Automatic judgement of neural network-generated image captions. In Carlos Martín-Vide, Matthew Purver, and Senja Pollak, editors, *Statistical Language and Speech Processing*, pages 261–272, Cham, 2019. Springer International Publishing.
- [55] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006.
- [56] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. *KI - Künstliche Intelligenz*, 34(4):571–584, 2020.
- [57] Rajarshi Biswas, Michael Barz, Mareike Hartmann, and Daniel Sonntag. Improving german image captions using machine translation and transfer learning. In Luis Espinosa-Anke, Carlos Martín-Vide, and Irena Spasić, editors, *Statistical Language and Speech Processing*, pages 3–14, Cham, 2021. Springer International Publishing.
- [58] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [59] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.
- [60] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- [61] Adrian Bussone, Simone Stumpf, and Dymrna O’Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.
- [62] Huan Ling and Sanja Fidler. Teaching machines to describe images with natural language feedback. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [63] Tingke Shen, Amlan Kar, and Sanja Fidler. Learning to caption images through a lifetime by asking questions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10393–10402, 2019.

- [64] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. Building trust in interactive machine learning via user contributed interpretable rules. In *27th International Conference on Intelligent User Interfaces*, pages 537–548, 2022.
- [65] Aliko Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. Putting humans in the image captioning loop. *Bridging Human-Computer Interaction and Natural Language Processing*, 2022.
- [66] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision (ECCV)*, pages 417–434, 2020.
- [67] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467, 2019.
- [68] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020.
- [69] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.
- [70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, pages 84–90, 2012.
- [71] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.
- [72] Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019.
- [73] Suho Cho. Image Augmentation — thecho7. <https://velog.io/@thecho7/Image-Augmentation>. [Accessed 08-01-2024].
- [74] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *13th International Conference on Neural Information Processing Systems*, page 395–401, 2000.
- [75] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv preprint arXiv:2003.13048*, 2020.
- [76] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020.

- [77] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. pages 3642–3646, 05 2020.
- [78] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto, Ontario, 2009.
- [79] Claude Coulombe. Text data augmentation made simple by leveraging nlp cloud apis. *ArXiv*, 2018.
- [80] William Yang Wang and Diyi Yang. That’s so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Conference on Empirical Methods in Natural Language Processing*, page 2557–256, 2015.
- [81] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *19th ACM International Conference on Information and Knowledge Management*, pages 759–768, 2010.
- [82] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 452–457, 2018.
- [83] Valerio Fortunati. How does deep learning in radiology work? — quantib.com. <https://www.quantib.com/blog/how-does-deep-learning-work-in-radiology>. [Accessed 08-01-2024].
- [84] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [85] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
- [86] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1, 2016.
- [87] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
- [88] Kh. Nafizul Haque. What is Convolutional Neural Network — CNN (Deep Learning) — linkedin.com. <https://www.linkedin.com/pulse/what-convolutional-neural-network-cnn-deep-learning-nafiz-shahriar/>. [Accessed 08-01-2024].
- [89] Ruilin Hu and Tianyang Luo. Xgboost-lstm for feature selection and predictions for the sp 500 financial sector. *Advances in Economics, Management and Political Sciences*, 59:249–257, 01 2024.
- [90] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [91] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

- [92] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [93] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [94] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [95] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [96] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Gregory Frederick Diamos, Erich Elsen, Ryan J. Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and A. Ng. Deep speech: Scaling up end-to-end speech recognition. *ArXiv*, abs/1412.5567, 2014.
- [97] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- [98] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.
- [99] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pages 234–241. Springer, 2015.
- [100] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14, pages 21–37. Springer, 2016.
- [101] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.
- [102] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [103] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [104] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [105] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

- [106] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.
- [107] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A Young, and Brian Belgodere. Image captioning as an assistive technology: lessons learned from vizwiz 2020 challenge. *Journal of Artificial Intelligence Research*, 73:437–459, 2022.