Integrating Classification in Image Captioning Tasks: A Study

Master's degree in Data Science

Gustavo Rocha Luz

Leiria, March of 2025

Integrating Classification in Image Captioning Tasks: A Study

Masters degree in Data Science

Gustavo Rocha Luz

Dissertation developed under the supervision of Professor Carlos Fernando de Almeida Grilo, Professor Rolando Lúcio Germano Miragaia, Professor José Carlos Bregieiro Ribeiro and Professor Luís Miguel de Oliveira Pegado de Noronha e Távora

Leiria, March of 2025

Originality and Copyright

This dissertation report is original and was created solely for this purpose. All authors whose studies and publications were used to complete it are duly acknowledged.

Acknowledgments

I would like to express my sincere gratitude to my academic supervisors, who supported me with dedication throughout the entire dissertation journey. From the very beginning, they were always available to guide me, provide valuable insights, and offer constructive feedback. Several months of meetings, discussions, and alignment efforts made the completion of this work possible.

I also extend my heartfelt thanks to my family, who have been my foundation during this entire process. Their unwavering support, constant encouragement, and motivating words were essential in helping me move forward, take the next step, and continually strive to exceed my own expectations.

Abstract

Image captioning combines computer vision and natural language processing to generate descriptive text for images. This dissertation evaluates whether integrating image classification into captioning models improves the quality of generated descriptions. Experiments were conducted with LSTM and Bidirectional LSTM architecture, using CNN-based feature extractors on the FLOWERS dataset. Each configuration was trained 35 times with controlled random seeds to ensure consistency and reproducibility .

Although all standard evaluation metrics were computed, the focus was on METEOR and SPICE for their balanced view of linguistic and semantic quality. ResNet50 yielded the best overall results among CNNs. The inclusion of classification labels showed mixed outcomes: in the Base Case, it increased variability; in BiLSTM models, it led to better METEOR scores and more consistent results.

Further tests with varied classification accuracy showed limited impact on caption quality. The model remained robust, with no significant drop in performance observed down to 80% accuracy, and top performance recorded at 95% and 90% classification accuracy. These findings suggest classification can enhance performance under favorable conditions, especially when paired with BiLSTM architectures, which is valuable for real-world settings where classification errors are expected.

In summary, the results underscore the subtle but meaningful role of classification in image captioning and offer guidance for building more robust multimodal systems.

Keywords: Image Captioning, Multimodal Learning, BiLSTM, Classification, METEOR, SPICE.

Resumo

A geração de legendas para imagens combina visão computacional e processamento de linguagem natural para produzir descrições textuais descritivas a partir de imagens. Esta dissertação avalia se a integração da classificação de imagens em modelos de legendagem melhora a qualidade das descrições geradas. Os experimentos foram realizados com arquiteturas LSTM e Bidirecional LSTM, utilizando extratores de características baseados em CNNs sobre o conjunto de dados FLOWERS. Cada configuração foi treinada 35 vezes com sementes aleatórias controladas, a fim de garantir consistência e reprodutibilidade dos resultados.

Embora todas as métricas padrões de avaliação tenham sido calculadas, a análise concentrou-se nas métricas METEOR e SPICE, por fornecerem uma visão equilibrada da qualidade linguística e semântica. A ResNet50 apresentou os melhores resultados gerais entre as CNNs avaliadas. A inclusão de rótulos de classificação gerou resultados variados: no modelo Base Case, aumentou a variabilidade; nos modelos BiLSTM, levou a melhores pontuações de METEOR e maior consistência nos resultados.

Testes adicionais com diferentes níveis de acurácia de classificação indicaram impacto limitado na qualidade das legendas. O modelo manteve-se robusto, sem queda significativa de desempenho até o limite de 80% de acurácia, com os melhores resultados sendo registrados nos níveis de 95% e 90%. Esses achados sugerem que a classificação pode melhorar o desempenho sob condições favoráveis, especialmente quando associada a arquiteturas BiLSTM, o que é relevante para contextos reais onde erros de classificação são esperados.

Em síntese, os resultados ressaltam o papel sutil, porém relevante, da classificação na tarefa de geração de legendas para imagens e oferecem orientações práticas para o desenvolvimento de sistemas multimodais mais robustos.

Palavras-chave: Geração de Legendas, Aprendizado Multimodal, BiLSTM, Classificação, METEOR, SPICE.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
AoANet	Attention on Attention Network
BiLSTM	Bidirectional Long Short-Term Memory
BLEU	Bilingual Evaluation Understudy
BLIP-2	Bootstrapped Language-Image Pretraining
CIDEr	Consensus-based Image Description Evaluation
CLIP	Contrastive Language–Image Pretraining
CNN	Convolutional Neural Network
ESTG	Escola Superior de Tecnologia e Gestão
GAN	Generative Adversarial Network
GIT	General Image Transformer
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
L1/L2	L1 and L2 Regularization Techniques
LDA	Latent Dirichlet Allocation
LRCN	Long-Term Recurrent Convolutional Network
LSTM	Long Short-Term Memory
MBConv	Mobile Inverted Bottleneck Convolution
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ML	Machine Learning
MLE	Maximum Likelihood Estimation
mPLUG	Multimodal Pretraining with Latent Unified Guidance
MS COCO	Microsoft Common Objects in Context
MTL	Multi-Task Learning
NLP	Natural Language Processing
OSCAR	Object-Semantics Aligned Pretraining
ReLU	Rectified Linear Unit
RMSProp	Root Mean Square Propagation
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SIFT	Scale-Invariant Feature Transform
SimVLM	Simple Visual Language Model
SPICE	Semantic Propositional Image Caption Evaluation
STL	Single-Task Learning
VGG	Visual Geometry Group (CNN architecture)

1. Introduction

Image captioning is a rapidly evolving field that bridges the gap between computer vision and natural language processing, enabling machines to generate descriptive text for visual content. The ability to automatically describe visual content has profound implications across various domains, including assistive technologies (Ordonez et al., 2011), content retrieval (Feng & Lapata, 2010), and human-computer interaction (Cornia et al., 2020). Early image captioning methods relied on predefined templates and rule-based approaches, which limited their flexibility and adaptability (Farhadi et al., 2010). With the advent of deep learning, models have become increasingly sophisticated, leveraging neural networks to learn contextual relationships between visual and textual data, improving the accuracy and fluency of generated captions (Vinyals et al., 2015).

Recent advancements in image captioning have incorporated techniques such as convolutional neural networks (CNNs) for feature extraction (He et al., 2016) and recurrent neural networks (RNNs) for sequential text generation (Hochreiter & Schmidhuber, 1997). These improvements have enhanced the ability of models to generate more accurate and contextually relevant captions across various datasets.

Building upon these foundations, attention mechanisms and transformer-based architecture have been introduced to overcome RNNs' sequential limitations (Vaswani et al., 2017). These models dynamically align visual and textual components, improving coherence and specificity in caption generation.

Despite these advancements, challenges persist in capturing intricate semantic details, abstract concepts, and complex image structures. Anderson et al. (2018) highlights the difficulty of aligning textual descriptions with visual elements, underscoring the need for continued innovation in multimodal learning. This dissertation explores one approach: integrating classification as an auxiliary task to enhance the accuracy and relevance of captioning.

1.1.Motivation

Image captioning has become a cornerstone of assistive technologies. For example, assistive systems designed for visually impaired individuals can describe objects in their surroundings, fostering greater autonomy. Devices equipped with captioning capabilities can identify and define visual elements, such as "a car" or "a tree", enabling users to interact with their environment more effectively. Research by Ordonez et al. (2011) highlighted the potential of large-scale datasets in enhancing captioning models for accessibility applications, underscoring the importance of ongoing advancements in this domain.

Similarly, in the context of e-commerce, visual search systems such as those used by Amazon help users find related products by analyzing uploaded images. These applications demonstrate the transformative potential of image captioning in improving accessibility and efficiency in various domains. Studies, such as that by Feng and Lapata (2010), have explored how multimodal learning techniques enhance product retrieval, further validating the impact of image captioning beyond accessibility.

Despite the progress achieved with models like "Show and Tell" (Vinyals et al., 2015), significant gaps remain, particularly in generating contextually accurate captions resembling those produced by humans. Earlier studies, such as those by Farhadi et al. (2010), revealed that combining traditional semantic methods with modern classifiers can address these gaps. More recently, Cornia et al. (2020) demonstrated how integrating auxiliary tasks, such as object classification, improves caption coherence and contextual alignment. These approaches highlight the importance of refining multimodal learning techniques to establish robust semantic relationships and improve caption fluency and accuracy.

1.2.Objectives

This research examines the role of classification as an auxiliary task in image captioning models, assessing its impact on the coherence and contextual relevance of generated captions. By incorporating classification, models can gain a structured understanding of objects within an image, resulting in more semantically rich and precise descriptions.

The integration of classification tasks has been widely explored in literature, particularly in improving contextual understanding of image-related challenges. While traditional image captioning models have focused primarily on direct feature extraction from images to generate text, classification introduces an additional layer of semantic understanding by

categorizing objects and attributes within an image. This process allows for a more structured approach to captioning by incorporating explicit contextual cues.

Previous studies have highlighted the benefits of multimodal learning, where classification aids in refining object recognition before the captioning process. Research, such as that by Farhadi et al. (2010), has demonstrated how structured representations of objects and actions contribute to more semantically accurate captions.

More recent approaches have leveraged classification to guide neural networks in refining the descriptive power of generated captions. As illustrated in Figure 1.1, the model studied in this work follows the original structure, represented in gray, which includes visual feature extraction and caption generation. The highlighted classification input, introduced in this work, is integrated into the pipeline with the aim of potentially enhancing the descriptive accuracy of the generated captions.

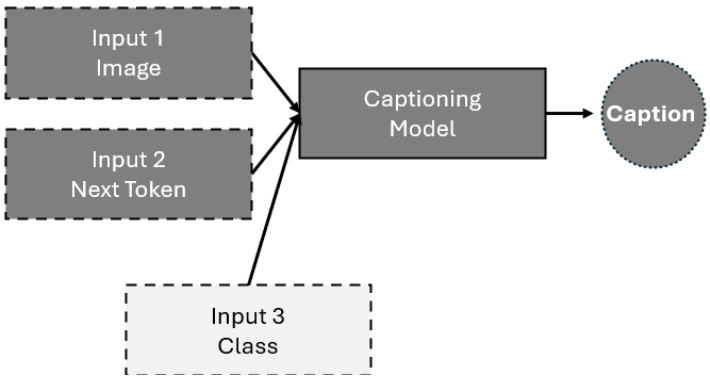

Figure 1.1 – Diagram proposed in the work

1.3.Methodology

This research employed an iterative methodology that followed a structured process consisting of dataset preparation, model development, and performance evaluation.

The dataset was pre-processed by converting tokenized text into words, as the collection initially stored captions in a tokenized format. An analysis was conducted to understand the distribution of words and phrases, revealing a significant variance in the number of captions per image. To enhance learning efficiency, captions contributing to greater vocabulary diversity were prioritized.

With the refined dataset, preprocessing was performed, including converting all text to lowercase, removing special characters, standardizing spaces, filtering out words with fewer with just one character, and appending *startseq* and *endseq* tokens to mark sentence boundaries. The dataset was then tokenized and split into training and testing sets.

Following this, image features were extracted using deep learning-based methods, and various model architectures were trained. The models were then evaluated using standard metrics, including BLEU, METEOR, ROUGE, and SPICE, followed by comparative analyses to assess the impact of different configurations on captioning performance. This entire process, as visually summarized in Figure 1.2, reflects the workflow adopted in this study, from the initial stages of data collection and preprocessing, through training and evaluation, to iterative refinements leading to the results.

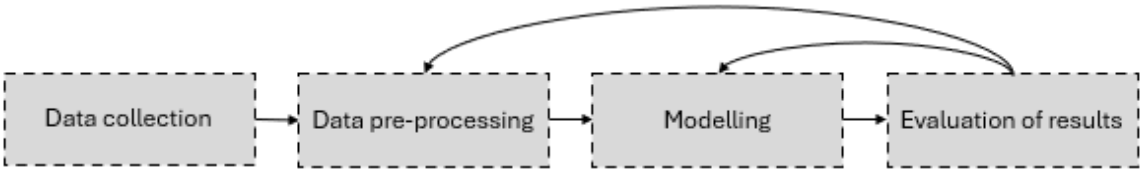

Figure 1.2 – Workflow adopted in this study

1.4.Structure of the Document

The rest of this dissertation is structured into four main chapters:

- **Background and Literature Review**: Covers the foundational concepts and recent developments related to image classification, image captioning, and the use of classification as an auxiliary task within multi-task learning frameworks.
- **Problem and Modelling**: Details the problem addressed in this study, the rationale behind the chosen methodologies, the data preparation steps, and the adaptations made to the "Show and Tell" model. This chapter also introduces and explains the different model variations evaluated, including versions with and without classification input, and with modifications to the architecture such as the removal of residual connections or the use of BiLSTM layers.

- **Results and Discussion**: Presents the results obtained from multiple experiments, providing comparative analyses of different model configurations and discussing the observed outcomes.
- **Conclusion**: Summarizes the key findings of the research, acknowledges its limitations, and outlines directions for future investigation.

2. Background and Literature Review

Early approaches relied heavily on rule-based systems and manually crafted templates, which constrained flexibility and scalability. For example, these systems might generate a caption like "A person riding a bicycle" by mapping specific image features, such as "bicycle" or "person", to predefined templates. While effective for simple and controlled scenarios, they struggled with complex or unseen data, limiting applicability in dynamic environments.

The advent of deep learning introduced a transformative shift, enabling models to learn patterns and relationships directly from large datasets. These advancements have expanded the applicability of image captioning systems, transitioning the field from rigid frameworks to highly adaptable, data-driven solutions. Furthermore, they emphasize the need to address ongoing challenges, such as enhancing the correlation between automated evaluation metrics and human judgment.

2.1. Image Classification

Image classification is a foundational task in computer vision that involves assigning labels to images based on their visual content. This field has undergone significant evolution, transitioning from manually crafted feature extraction techniques, such as edge detection and texture analysis, to data-driven deep learning models (Krizhevsky et al., 2012). These earlier approaches often struggled with complex image variations and required domain expertise to design compelling features.

The emergence of deep learning has revolutionized image classification by enabling models to learn hierarchical feature representations directly from data, thereby surpassing traditional handcrafted approaches. Convolutional Neural Networks played a pivotal role in this transformation, providing a structured way to process spatial information through convolutional layers, as shown in Figure 2.1, which traces this progression from LeCun's 1998 model to the most advanced convolutional architectures.

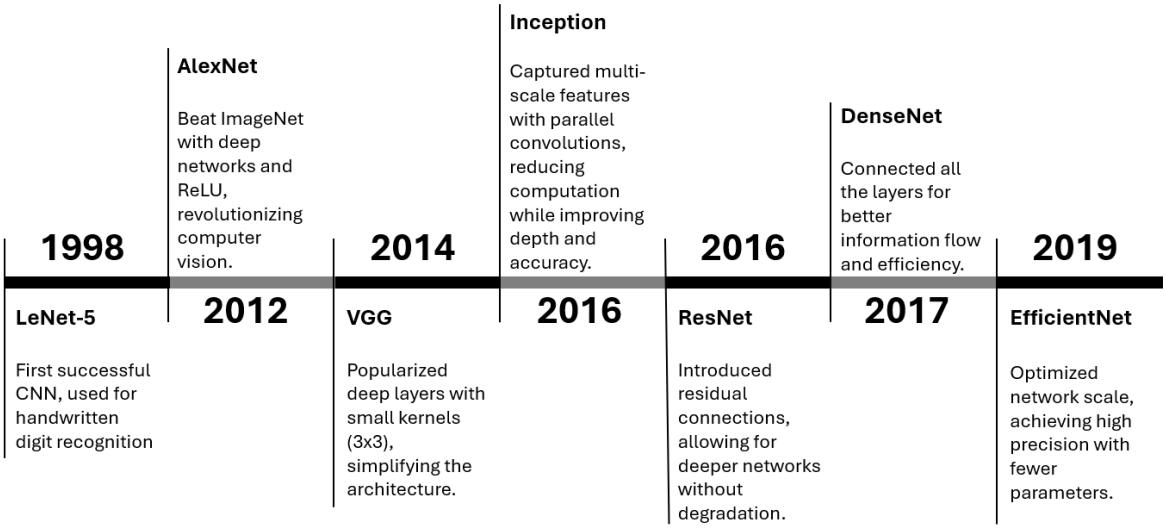

Figure 2.1 – Image Classification Evolution

CNNs were introduced by LeCun et al. (1998) with the LeNet-5 model, which was designed for handwritten digit recognition. It efficiently captured spatial hierarchies using convolutional and pooling layers, laying the groundwork for modern CNNs.

A breakthrough in large-scale image classification was achieved with AlexNet (Krizhevsky et al., 2012), which significantly improved the accuracy of the ImageNet dataset (Deng et al., 2009). AlexNet's architecture consisted of eight layers, including five convolutional layers and three fully connected layers, making it one of the deepest networks of its time. The model introduced several innovative techniques, such as ReLU activation functions to accelerate training and dropout layers to prevent overfitting. GPU acceleration was crucial in enabling AlexNet to be trained on the massive ImageNet dataset, which comprises over 1.2 million labeled images across 1,000 classes.

This performance established CNNs as the dominant paradigm in image classification, setting a new benchmark for accuracy and efficiency. AlexNet's results inspired subsequent research, driving innovation in architectural design and training methodologies. However, its computational demands, particularly the reliance on GPUs and the need for substantial labeled data, also highlight challenges that will shape future developments in model efficiency and data augmentation techniques.

The impact of AlexNet extended beyond its immediate success, shaping the direction of future convolutional network designs and inspiring a wave of increasingly deep and efficient

architectures. By demonstrating the advantages of deeper networks and the effectiveness of GPU acceleration, AlexNet established a precedent for future models that prioritize depth and computational efficiency. However, its reliance on substantial labeled data posed challenges that remain relevant for future research. These challenges include the need for lightweight architectures suitable for edge devices and techniques to mitigate the reliance on vast annotated datasets, such as semi-supervised or self-supervised learning approaches.

Building on the success of AlexNet, VGG (Simonyan & Zisserman, 2014) advanced deep learning by increasing network depth while maintaining a simplified structure. VGG employed small, fixed-size convolutional filters (3x3) throughout its layers, which helped capture fine-grained details while maintaining computational efficiency. Unlike AlexNet, VGG introduced uniform configurations, making the architecture easier to generalize and expand.

The model achieved state-of-the-art performance on the ImageNet dataset, with VGG-16 and VGG-19 being the most prominent versions. These variants contained 16 and 19 layers, respectively, emphasizing the importance of depth in improving classification accuracy. However, VGG's computational complexity was a notable drawback, as the network required significant memory and processing power, limiting its scalability for real-time or resource-constrained applications. Despite these limitations, VGG remains widely used as a feature extractor in transferring learning tasks because it produces rich and detailed visual data representations.

Inception (Szegedy et al., 2015), also known as GoogLeNet, introduced a novel architectural concept based on parallel convolutional pathways within the same layer, referred to as Inception modules. These modules allowed the network to capture multi-scale spatial information by combining filters of different sizes. This design reduced computational costs while maintaining high representational power, making it possible to build deeper and broader networks without a proportional increase in parameters.

The first version of Inception achieved strong performance on the ImageNet classification challenge while being more efficient than its contemporaries, such as VGG. Its modularity also allowed easy scalability and formed the basis for several subsequent improvements. Inception-v3 introduced enhancements such as factorized convolutions, label smoothing, and RMSProp optimization, improving both training stability and accuracy. These refinements allowed Inception-v3 to achieve state-of-the-art performance with fewer

parameters and reduced computational cost compared to its predecessors (Szegedy et al., 2016).

ResNet (He et al., 2016) introduced residual connections to address the vanishing gradient problem, enabling stable training of intense networks. These connections allow information to bypass specific layers, ensuring adequate gradient flow and making optimization more efficient. Traditional deep networks faced challenges where gradients would diminish as they propagated backward through multiple layers, leading to ineffective weight updates. ResNet overcame this issue through identity mappings, which bypassed one or more layers and carried the gradient information directly to earlier layers.

This breakthrough enabled training networks with over one hundred layers, which was previously impractical due to optimization difficulties. One of the most widely used variants is ResNet-50, which comprises 50 layers and strikes a balance between depth and computational efficiency. ResNet-50 became a standard backbone in many computer vision tasks due to its robust performance and adaptability, particularly in transferring learning scenarios and feature extraction pipelines.

ResNet maintained high accuracy by leveraging residual connections without suffering from degradation as network depth increased. This approach led to state-of-the-art performance on benchmarks such as ImageNet, surpassing human-level accuracy in image classification tasks. The introduction of deep residual networks significantly improved gradient flow and optimization, making it possible to train deep models that generalize well across multiple domains.

Building on these innovations, DenseNet (Huang et al., 2017) enhanced feature reuse by densely connecting all layers within a block. This design ensured that each layer had direct access to the feature maps of all preceding layers, promoting feature reuse and improving gradient flow during backpropagation. By mitigating the vanishing gradient problem, DenseNet enabled more efficient training and facilitated the development of deeper networks.

One of DenseNet's key advantages is its parameter efficiency. Unlike traditional deep networks, which require many parameters to learn increasingly complex representations, DenseNet optimizes parameter utilization by ensuring that learned features are effectively shared across layers. This results in a model that requires fewer parameters while achieving

competitive accuracy, making it highly suitable for applications with limited computational resources.

Furthermore, due to its improved feature propagation and reduced redundancy in learned representations, DenseNet has been widely adopted in various domains, including medical imaging, remote sensing, and autonomous driving, where precise and computationally efficient image analysis is required.

The next evolution in CNN architecture came with EfficientNet (Tan & Le, 2019), which addressed the limitations of prior architectures by introducing a systematic approach to scaling neural networks. Unlike previous models that scaled depth, width, or resolution independently, EfficientNet employed a compound scaling method that proportionally adjusted all three dimensions. This holistic scaling approach optimized computational efficiency while maintaining high accuracy, distinguishing EfficientNet from earlier architectures that relied on arbitrary or manual scaling adjustments.

EfficientNet also leveraged a mobile inverted bottleneck convolution (MBConv) and squeeze-and-excitation layers to enhance feature recalibration. This combination improved the model's ability to capture fine-grained image details while significantly reducing computational overhead. As a result, EfficientNet outperformed prior architectures on benchmarks such as ImageNet, achieving higher accuracy with fewer parameters compared to ResNet and DenseNet.

Another key advantage of EfficientNet is its ability to scale across various sizes, from EfficientNet-B0 to EfficientNet-B7, where each successive variant increases model capacity while maintaining efficiency. This made it highly adaptable for applications ranging from mobile devices to large-scale cloud-based classification tasks. EfficientNet's balance between performance and computational cost solidified its role as a state-of-the-art architecture for efficient deep learning models.

The evolution of CNNs, from LeNet-5 to modern architectures, has continuously improved efficiency, scalability, and accuracy in image classification. The transition from early, manually designed networks to deep, automated models has underscored the importance of innovations in feature extraction, parameter optimization, and computational efficiency. Each breakthrough, whether through more profound architecture, enhanced connectivity, or

novel scaling techniques, has improved the accuracy and generalization of image classification models across diverse applications.

Beyond classification, advancements in CNNs have impacted object detection (Ren et al., 2015), medical imaging (Litjens et al., 2017), and generative modeling (Radford et al., 2018), demonstrating their versatility in computer vision. With ongoing research focusing on lightweight architectures for edge computing and hybrid approaches integrating transformers with CNNs, the future of deep learning in image classification continues to expand. CNNs' adaptability ensures they will remain a cornerstone of computer vision, evolving to meet new challenges and drive further innovations in artificial intelligence.

In recent years, attention mechanisms have emerged as a powerful concept in deep learning, enabling models to dynamically focus on the most relevant parts of an input. While initially popularized in natural language processing (Vaswani et al., 2017), attention has also enhanced vision models by improving their ability to capture spatial relationships and contextual dependencies (Wang et al., 2018).

The integration of attention mechanisms with CNNs, either through self-attention layers or hybrid transformer-CNN architectures, has led to improved performance in image classification and related tasks (Chen et al., 2015). By selectively weighing feature maps or spatial regions, attention augments CNNs' capacity to model complex structures and global context, bridging the gap between local feature extraction and holistic scene understanding.

2.2.Image Captioning

Image captioning has evolved significantly, transitioning from handcrafted, rule-based systems to advanced deep-learning models that integrate computer vision and natural language processing. This section provides a historical overview of image captioning, highlighting key milestones, methodologies, and challenges faced at each stage of development.

2.2.1. Early Approaches: Rule-Based System and Templates

The earliest methods for image captioning were based on rule-based systems and manually defined templates. These approaches mapped detected objects and their attributes to structured sentence patterns. A foundational example is Winograd's "Picture Description Task" (Winograd, 1972), which aimed to link objects and their spatial relationships with

linguistic structures. The model analyzed structured scene descriptions and applied predefined grammatical rules to generate captions.

The system relied on a knowledge-based framework where objects, spatial relations, and actions were manually encoded into logical forms, which were then translated into grammatically correct sentences. Unlike modern methods that extract visual features automatically, this model required extensive human intervention to define relationships between objects and generate textual output. Due to the computational limitations of the time, it lacked machine learning-based feature extraction, relying instead on manually structured scene representations.

The model was evaluated on a dataset of manually annotated images, which was constrained. Its performance was assessed by comparing the captions generated with the expected outputs, with a primary focus on syntactic accuracy and logical consistency. However, it lacked adaptability to novel images and variations in descriptions since it strictly adhered to predefined rule sets. This rigidity limited its ability to scale and manage more complex or ambiguous scenarios.

One of the significant shortcomings of this model was its reliance on static templates, which rendered it inflexible when confronted with images containing objects or relationships that were unseen. Additionally, the absence of probabilistic reasoning meant it struggled to manage uncertainty in object recognition. At the time, deep learning-based models, such as CNNs, had not yet been developed, restricting the model's ability to generalize beyond its training data.

Despite its limitations, Winograd's approach laid the groundwork for structured visual-linguistic integration. His model demonstrated the potential for linking vision and language through systematic rule-based representations. However, the extensive manual effort required to expand its knowledge base made it impractical for large-scale applications. The challenges of this early approach underscored the need for more flexible and scalable models, which would later emerge with the advent of statistical and deep learning-based techniques.

2.2.2. The Rise of Statistical and Probabilistic Methods

Following the limitations of rule-based captioning models, researchers in the 1980s and 1990s began developing probabilistic frameworks to introduce flexibility and data-driven learning into image captioning. Mori et al. (1999) pioneered a Hidden Markov Model (HMM) approach for caption generation, leveraging probabilistic modeling to improve caption coherence and diversity. This approach represented sequences of visual features as latent states, with transitions between states determining the word sequences in captions.

Mori's model utilized pre-defined feature extraction techniques, such as edge and texture analysis, to identify key visual elements in an image. These extracted features were then mapped to corresponding textual descriptions using an HMM framework, which assigned probabilities to word sequences. Unlike rule-based systems, this approach enables captions to be dynamically generated based on statistical likelihoods rather than rigid templates, allowing for greater flexibility in handling unseen images.

The model was evaluated using perplexity scores, a standard metric for assessing the probability distribution of generated text. Lower perplexity indicated more confident and coherent caption predictions. While introducing probabilistic modeling was a step forward, the reliance on handcrafted visual features still posed significant limitations. The system struggled with complex scenes that contained multiple interacting objects and lacked a mechanism for capturing hierarchical relationships within images.

At the time, the absence of deep learning techniques restricted the use of more advanced feature representations, such as CNNs. Consequently, Mori's approach was constrained by the quality of manually engineered features, limiting its ability to scale effectively across diverse datasets. Despite these limitations, this work established a foundation for subsequent models that incorporated machine learning with enhanced visual processing methods, facilitating the transition from rule-based captioning to deep learning-driven approaches.

Shortly after, Barnard and Forsyth (2001) expanded on this approach by introducing co-occurrence models that mapped words to image regions based on statistical learning. Their method provided a more dynamic and automated mapping of textual descriptions to visual features, marking one of the first significant departures from deterministic rules.

Evaluation methodologies for these statistical models primarily relied on perplexity and likelihood scores, which measured the probability of a generated caption given an observed

image. However, these models still relied on manually crafted image descriptors, such as the Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG). SIFT extracted distinctive key points from images, allowing for robust object recognition across variations in scale and rotation. On the other hand, HOG captured the distribution of gradient orientations, making it particularly effective for detecting edges and textures. While these methods improved feature extraction, they struggled to capture deeper semantic relationships and were limited in their ability to generalize across diverse datasets.

A key limitation of these approaches was their inability to generate complex, context-aware captions. The models were often restricted to narrow datasets, limiting their generalization to diverse image distributions. Furthermore, statistical approaches were unable to fully model the long-range dependencies between words in captions, resulting in inconsistencies in sentence structure and coherence.

Several models attempted to address these issues by incorporating additional probabilistic reasoning and structured learning approaches. Barnard et al. (2003) explored generative models that linked image regions with words through a hierarchical framework, aiming to improve contextual alignment using statistical co-occurrence techniques. Their method involved analyzing large-scale image datasets to establish word-region associations, but it struggled with rare objects or non-standard compositions, leading to inaccuracies in captioning novel images.

Blei and Jordan (2003) proposed another notable approach: the introduction of Latent Dirichlet Allocation (LDA) for visual-language modeling. LDA identified underlying themes, or “hidden topics”, in images and captions, providing a structured way to represent visual elements and enabling a more structured representation of visual content. However, while LDA improved caption coherence, it failed to fully capture fine-grained spatial relationships, making it unsuitable for highly detailed descriptions.

Fei-Fei et al. (2006) applied LDA-based techniques to explore object co-occurrence patterns, improving captioning generalization across varied datasets. Despite these advances, the reliance on manually defined visual features remained a bottleneck, preventing models from fully capturing complex visual semantics.

Gupta and Davis (2008) later introduced a multimodal Bayesian network to capture higher-level dependencies between objects, actions, and scenes. This model integrated object

recognition with contextual reasoning, allowing it to infer relationships between detected elements. Despite this advancement, the reliance on predefined probabilistic dependencies limited its scalability, requiring extensive manual tuning to generalize effectively across diverse datasets.

The inability to efficiently scale probabilistic models to large datasets led to the search for more expressive learning frameworks. Computational limitations and the scarcity of large-scale, labeled datasets have restricted the potential of statistical methods. The development of deep learning, particularly with the emergence of CNNs, has changed the way we think by enabling automated feature extraction from images. This transition allowed models to move beyond hand-crafted feature descriptors and instead learn hierarchical visual representations directly from data.

2.2.3. The Shift to Deep Learning.

One of the earliest deep learning-based approaches to image captioning was introduced by Kiros et al. (2014) with their multimodal neural language model. This model integrated CNNs for visual feature extraction and a Recurrent Neural Network for sequential text generation, shifting from traditional probabilistic methods to fully end-to-end deep learning architectures.

Kiros et al. (2014) employed an encoder-decoder framework in which a CNN processed image data into a compact feature representation. This feature vector was then fed into an RNN, specifically a Long-Short-Term Memory (LSTM) network, responsible for sequentially generating words to form captions. The CNN, trained on large-scale datasets such as Flickr8k and Flickr30k, extracted high-level visual features, while the LSTM modeled the sequential dependencies in textual descriptions.

The training process utilized maximum likelihood estimation (MLE), where the model was optimized to maximize the probability of generating reference captions given the extracted image features. The evaluation used well-established language generation metrics such as BLEU and METEOR, which are further detailed in the following sections, assessing lexical similarity between generated and human-annotated captions.

A key architectural aspect of this model was its ability to learn visual and linguistic representations in an end-to-end manner jointly. However, the model lacked an explicit mechanism for attending to different regions of an image, treating all extracted features

uniformly. This limitation sometimes led to incomplete or overly generic captions, as the model could not selectively focus on salient regions within an image. Additionally, while integrating deep learning improved caption fluency, the model still struggled with maintaining contextual coherence over longer descriptions.

Despite these challenges, Kiros et al. (2014) provided a crucial steppingstone for image captioning, demonstrating the feasibility of deep learning in this task. The absence of attention mechanisms and the need for improved alignment between image and text representations were identified as key areas for future research, leading to later innovations that addressed these shortcomings.

Attention mechanisms had not yet been widely adopted in image captioning, limiting the model's ability to dynamically emphasize relevant visual areas during caption generation. Kiros et al. (2014) acknowledged these limitations and suggested future improvements, such as incorporating mechanisms to refine the alignment between image features and textual descriptions.

Shortly after, Donahue et al. (2015) introduced the Long-term Recurrent Convolutional Network (LRCN), a model that further refined the integration of CNNs and RNNs for sequential prediction tasks. Unlike early methods that handled image processing and language generation as separate tasks, LRCN allowed feature representations to be dynamically updated throughout the caption generation process. This model extended beyond image captioning, demonstrating applications in activity recognition and video description.

The LRCN framework used CNN to encode images into a feature vector, which was then fed into a deep LSTM network that recursively generated captions. Unlike Kiros et al. (2014), who trained their model primarily on static image datasets like Flickr8k and Flickr30k, Donahue et al. (2015) leveraged the MS COCO dataset, which contains a broader variety of images and multi-caption annotations. The evaluation was conducted using BLEU and METEOR scores, with results indicating improved fluency and coherence over previous methods.

Despite its advances, LRCN still struggled with efficiently capturing dependencies between distant words in more extended captions. Furthermore, the model processed all image features simultaneously, treating them equally without prioritizing the most salient visual

regions. This limitation resulted in captions that, while coherent, sometimes failed to emphasize critical image elements.

The challenges presented by LRCN highlighted the need for better alignment between visual features and textual outputs descriptions. Researchers explored different strategies to enhance contextual awareness in captioning models, leading to improvements in recurrent architectures and multimodal learning approaches. These developments refined the connection between CNN-based feature extraction and RNN-based sequence modeling, ensuring more structured and semantically meaningful captions.

2.2.4. Breakthrough in Deep Learning-Based Image Captioning

As deep learning methodologies evolved, models incorporating more sophisticated architecture were introduced. The "Show and Tell" model, introduced by Vinyals et al. (2015), represented a pivotal advancement in image captioning by refining the encoder-decoder architecture and leveraging deep learning for end-to-end training. This model incorporates into CNNs for feature extraction and RNNs, specifically LSTM networks, for sequential caption generation.

The approach employed in "Show and Tell" followed an encoder-decoder framework. Unlike earlier models such as Kiros et al. (2014) and Donahue et al. (2015), which experimented with various levels of integration between vision and language, "Show and Tell" formalized a fully end-to-end approach. The pre-trained CNN (Inception-v3) encoder extracted high-level feature representations from images, which were then passed to the LSTM decoder to generate captions. While Kiros et al. (2014) incorporated a multimodal neural language model and Donahue et al. (2015) allowed for dynamic feature updates with LRCN, "Show and Tell" optimized visual feature extraction and caption generation within a unified end-to-end learning process, rather than employing a multi-task learning approach.

The preprocessing pipeline involved resizing images, normalizing pixel values, and extracting feature vectors from the final convolutional layer of the CNN. Textual data underwent tokenization, and captions were padded to ensure uniform input sequences. Integrating these components into a single architecture marked a shift from separately trained components to a unified learning process, improving fluency and contextual relevance in generated captions.

The architecture consisted of a CNN encoder that extracted visual features from input images, converting them into a fixed-length representation. These features were then passed to an LSTM decoder, which sequentially generated captions by modeling dependencies between words. The LSTM effectively captured temporal relationships, ensuring coherence in sentence construction. Additionally, a word embedding layer transformed words into dense vector representations, enhancing the model's ability to generalize across different linguistic structures.

Figure 2.2 illustrates the architecture of an image captioning model based on a CNN-LSTM framework. The CNN extracts feature representations from the input image, which are then passed through a linear transformation before being fed into an LSTM-based decoder. The LSTM sequentially generates words, conditioned on both the image features and previously predicted words, ultimately forming a complete caption.

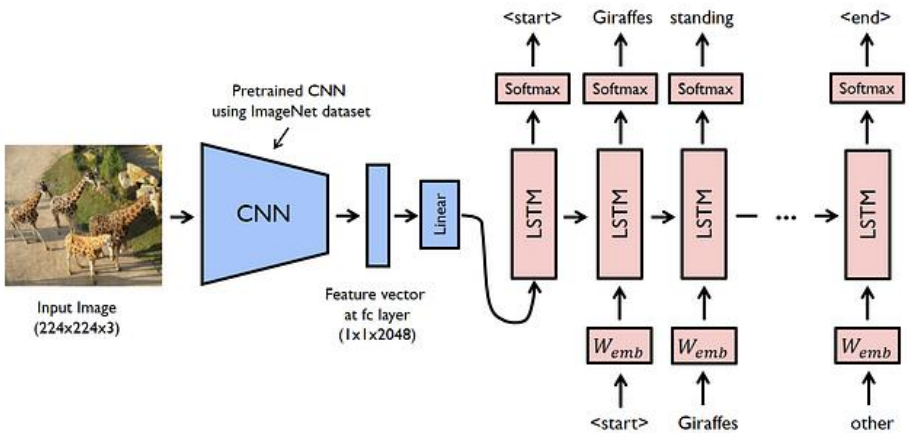

Figure 2.2 – Overview of the Image Captioning Model Architecture

The "Show and Tell" model was evaluated on benchmark datasets such as MS COCO and Flickr8k. Performance was measured using automatic evaluation metrics, including BLEU, METEOR, ROUGE, CIDEr, and SPICE. The results demonstrated a significant improvement over the previous captioning model. The model exhibited higher fluency and contextual alignment between images and text, outperforming traditional statistical approaches and earlier deep learning-based models.

A key advancement introduced by the "Show and Tell" model improved image captioning performance compared to earlier models. Vinyals et al. (2015) reported that the model achieved a BLEU-4 score of 27.7 on MS COCO, a notable increase over prior approaches.

CIDEr scores further highlighted its effectiveness, demonstrating the model's ability to generate captions that closely align with human-annotated references. Additionally, METEOR and ROUGE metrics indicated enhanced lexical diversity and improved recall-based performance, solidifying "Show and Tell" as a benchmark in the field.

This recognition is further supported by Bernardi et al. (2016), who noted that recent models such as "Show and Tell" achieved state-of-the-art results on widely used benchmarks like MS COCO.

However, while "Show and Tell" introduced substantial advancements, it still faced challenges, particularly in handling fine-grained object distinctions and novel image compositions. This led to the exploration of attention-based mechanisms in later models.

The lack of explicit attention mechanisms meant that the model sometimes generated overly generic captions or failed to capture key details. These limitations paved the way for subsequent improvements in attention-based architectures, which aimed to focus on relevant image regions dynamically during caption generation.

2.2.5. Modern Developments in Image Captioning

As image captioning progressed beyond the "Show and Tell" model, researchers introduced several innovations to address its limitations. While a comprehensive exploration of these advancements could fall outside the primary scope of this dissertation, it is worth to acknowledge some key developments that have significantly influenced modern approaches.

One such advancement was the integration of attention mechanisms, as introduced in the 'Show, Attend and Tell' model (Xu et al., 2015). This approach allowed captioning models to dynamically focus, on relevant image regions, improving contextual accuracy and reducing generic captioning tendencies. Later, transformer-based architecture further expanded on this by leveraging self-attention mechanisms to capture both fine-grained local details and long-range dependencies within the image, providing a more holistic understanding of visual content.

Additionally, transformer-based architecture has reshaped the field, with models like "Image Transformer" (Parmar et al., 2018) and CLIP (Radford et al., 2021), demonstrating improved generalization across diverse datasets. These models leverage self-attention mechanisms to

capture long-range dependencies in visual and textual information, overcoming the sequential processing limitations of RNN-based models.

Recent work has also expanded into multi-modal learning, incorporating image captioning with complementary tasks such as object detection and scene understanding. The emergence of vision-language pretraining, seen in models like Oscar (Li et al., 2020) and SimVLM (Wang et al., 2021), has pushed the boundaries of image captioning by leveraging extensive datasets to improve contextual representation and caption diversity.

Despite these advances, challenges persist, including dataset biases (Zhao et al., 2017), difficulties in capturing abstract concepts (Lu et al., 2017), and the heavy reliance on large-scale annotated data (Schuhmann et al., 2021). Future research continues to explore solutions to these challenges, ensuring that image captioning systems evolve toward more robust, context-aware, and explainable models, as illustrated by the timeline in Figure 2.3, which presents the evolution of these models from their origins in 1972 to the most modern architecture.

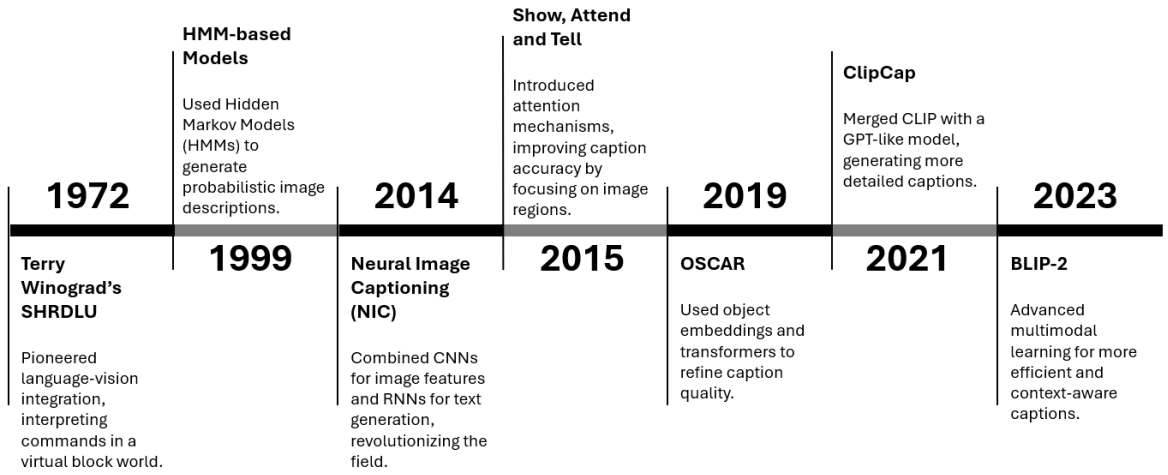

Figure 2.3 – Image Captioning Evolution

2.3.Single-Task Learning vs. Multi-Task Learning

Machine learning models often adopt different training paradigms depending on the complexity of the task and the necessity for generalization. Among these paradigms, sequential-task learning and multi-task learning (MTL) represent two distinct approaches for handling learning objectives (Caruana, 1997; Ruder, 2017). Each approach has its strengths

and weaknesses, influencing how models acquire and transfer knowledge, which is particularly relevant in deep learning applications.

2.3.1. Single-Task Learning

Single-task learning (STL) is a paradigm in which a model is trained to pursue a single objective without leveraging knowledge from other tasks. Each task is treated independently, and no shared representation is used across multiple objectives. This approach is common when data from different tasks arrive sequentially, or models require incremental learning to adapt to new domains.

A key advantage of single-task learning is that it enables models to optimize performance for a specific task without interference from other tasks, ensuring task-specific optimization. However, a significant challenge is catastrophic forgetting, where knowledge from previously learned tasks deteriorates as new tasks are introduced. Techniques such as knowledge distillation and replay mechanisms (Lopez-Paz & Ranzato, 2017) have been proposed to mitigate this issue.

Notable applications of single-task learning include domain-specific optimizations in natural language processing (Howard & Ruder, 2018), task specialization based on reinforcement learning (Mendez et al., 2021), and highly specialized computer vision models (Rebuffi et al., 2017). In image captioning, Single-Task learning could refine descriptions by first learning object detection before generating textual descriptions.

2.3.2. Multi-Task Learning

Multi-task learning (MTL) is a strategy in which a model is trained simultaneously on multiple tasks, leveraging shared representations to enhance generalization. MTL is particularly beneficial when tasks are related, allowing the model to extract standard features and reduce overfitting by providing additional training signals.

MTL has been widely studied in deep learning (Caruana, 1997) and has been applied successfully in various domains, including natural language processing (Liu et al., 2019), autonomous driving (Kendall et al., 2018), and medical image analysis (Xu et al., 2021). One of the primary challenges of MTL is task interference, where competing objectives lead to partial learning. Strategies such as adaptive weighting of loss functions (Sener & Koltun, 2018) and complex parameter sharing have been proposed to address these issues.

In image captioning, MTL has been explored to integrate auxiliary tasks such as object classification, scene recognition, and sentiment analysis to enhance caption quality. Studies have demonstrated that incorporating classification as an auxiliary task improves the semantic richness of generated captions by guiding the model to focus on relevant objects and attributes (Cornia et al., 2020).

2.3.3. Comparison and Relevance to Image Captioning

While STL is well-suited for scenarios requiring incremental adaptation, MTL is often preferred when multiple tasks share underlying representations. In image captioning, MTL offers a compelling framework by enabling models to learn auxiliary tasks, such as object classification, alongside caption generation, resulting in more accurate and contextually relevant descriptions.

This dissertation investigates how classification, employed as an auxiliary task within an STL framework, can enhance image captioning performance. The study employs single task learning principles to assess whether integrating classification improves the fluency, coherence, and accuracy of generated captions.

2.4.Evaluation Metrics

Evaluating image captioning models is crucial for understanding their effectiveness and relevance in generating descriptive text. Over the years, automatic metrics have become the primary tools for this purpose, providing standardized methods to assess accuracy, linguistic quality, and contextual relevance. These metrics enable researchers to compare models quantitatively and ensure consistent benchmarks. Studies by Cui et al. (2018) in "Learning to Evaluate Image Captioning" and Bernardi et al. (2016) in " Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures " emphasize the significance of these metrics in driving advancements in the field, while also acknowledging their limitations in accurately capturing human judgment.

Before the advent of automated evaluation metrics, image captioning models were assessed primarily through human evaluation, where experts rated captions based on fluency, relevance, and accuracy. While human judgment offers a nuanced understanding of caption quality, it is subject to limitations, including subjectivity, time constraints, and limited reproducibility. As image captioning research expanded, the need for scalable and

standardized evaluation methods became evident, leading to the development of automatic metrics.

The introduction of automated metrics transformed the evaluation landscape. Early methods, such as BLEU (Papineni et al., 2002), focused on n-gram precision, measuring lexical similarity between generated and reference captions. Subsequent metrics, including ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and SPICE (Anderson et al., 2016), sought to address BLEU's limitations by incorporating recall, semantic similarity, and scene structure analysis.

Despite the efficiency of automated metrics, they face several challenges in aligning with human judgment. One major limitation is their reliance on predefined reference captions, which may not capture the full range of valid descriptions for a given image. This often penalizes captions that use synonyms or alternative phrasing despite conveying the same meaning. Additionally, these metrics struggle with assessing high-level semantic relationships, creativity, and contextual nuances, which are crucial for human-like captioning. Researchers continue to explore hybrid approaches that integrate human-in-the-loop assessments and learned models to improve alignment with human perception.

Several factors must be considered when comparing results across different studies to ensure fair and meaningful comparisons. Variations in dataset composition, preprocessing techniques, model architectures, and hyperparameter tuning can significantly impact reported performance. Direct comparisons should ideally be made between models evaluated under the same conditions using identical datasets and evaluation metrics. However, due to differences in experimental setups, it is often necessary to consider relative improvements within each study rather than absolute metric scores. Additionally, variations in human-labeled reference captions can influence automated evaluation metrics, adding another layer of complexity to cross-study comparisons. Researchers should, therefore, contextualize their findings and highlight any discrepancies in evaluation conditions when presenting comparative analyses.

Furthermore, evaluating results across studies should be conducted cautiously, as different models may be optimized for specific datasets or tasks. Automated metrics provide a standardized evaluation framework, but their scores are influenced by dataset diversity, model architecture, and training methodology. Consequently, absolute metric scores may not always be directly comparable across research efforts. Instead, relative improvements

within the same experimental conditions should be prioritized. Additionally, integrating human evaluations alongside automated metrics remains an essential practice to validate the real-world effectiveness of generated captions, as automated methods alone may fail to capture finer nuances in linguistic quality and contextual appropriateness.

2.4.1. BLEU

The BLEU (Bilingual Evaluation Understudy) metric, introduced by Papineni et al. (2002), measures the overlap between generated captions and reference texts by analyzing n-gram precision. An n-gram is a sequence of n words that appear consecutively in a text. For example, a unigram consists of a single word, while a bigram is a two-word sequence. BLEU evaluates the number of n-grams from the generated caption that appears in the reference captions, assigning higher scores to outputs with substantial lexical similarity. It is divided into levels, from BLEU-1 (which considers unigrams) to BLEU-4 (which evaluates up to four-gram sequences).

Despite its widespread use, BLEU has limitations in capturing semantic equivalence or creativity. It heavily relies on exact matches and does not account for synonyms or paraphrased expressions. For example, a caption with synonymous wording may receive a lower BLEU score even though it conveys the same meaning as the reference, illustrating its lexical bias. These limitations are particularly evident in creative or diverse datasets, where semantic richness is more critical than exact lexical matching.

Recent benchmarks in image captioning indicate state-of-the-art models, such as mPLUG (Li et al., 2023), which utilizes a vision-language pretraining framework combining transformer-based architectures for both image and text encoding, achieve BLEU-4 scores of 41.0 on the MS COCO dataset (Li et al., 2023). Compared to previous models, such as AoANet (Huang et al., 2019), which scored 38.9 BLEU-4 on the same dataset, mPLUG demonstrates improved contextual understanding and caption fluency, highlighting advancements in multimodal learning.

2.4.2. ROUGE

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, outlined by Lin (2004), evaluates textual recall by analyzing overlapping n-grams, word sequences, and sentence-level matches between generated and reference captions. ROUGE is robust in assessing the completeness of information in generated texts, making it valuable for tasks

that prioritize recall over precision, such as summarization. In the context of image captioning, ROUGE captures how well-generated captions include all key elements described in the reference. However, its focus on recall may lead to overemphasizing verbosity and penalizing shorter, concise captions. Additionally, like BLEU, ROUGE struggles with semantic diversity, as it relies on exact or near-exact lexical matches, making it less effective in capturing context or nuanced meaning.

Recent benchmarks indicate that state-of-the-art models, such as GIT (Wang et al., 2022), achieve ROUGE scores of 58.2 on the MS COCO dataset, demonstrating significant advancements in capturing textual richness and diversity in image captioning.

2.4.3. CIDEr

The CIDEr (Consensus-based Image Description Evaluation) metric, developed by Vedantam et al. (2015), assesses image captions based on their alignment with human consensus. CIDEr calculates a weighted similarity score for n-grams, giving higher importance to terms that are frequent across reference captions but rare in generic text corpora. Unlike BLEU and ROUGE, CIDEr does not operate on a fixed 0-100% scale; its scores vary depending on dataset-specific distributions and linguistic diversity. A higher CIDEr score indicates more substantial alignment with human-written captions, with values typically ranging from 50 to 150 in benchmark evaluations.

Recent benchmarks indicate that state-of-the-art models, such as BLIP-2 (Li et al., 2023), achieve CIDEr scores of 149.6 on the MS COCO dataset, outperforming previous models like mPLUG (Huang et al., 2022), which scored 141.7. These improvements highlight advancements in multimodal pretraining and the effectiveness of vision-language architecture in generating human-like image captions.

2.4.4. METEOR

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric, proposed by Banerjee and Lavie (2005), improves upon BLEU by addressing its inability to capture linguistic variations and semantic equivalences. Unlike BLEU, METEOR evaluates captions by considering synonyms, stemming, and word order, providing a more flexible and comprehensive similarity measure. METEOR assigns higher scores to captions that are semantically like reference texts, even if they do not share exact word matches. This

characteristic makes it particularly useful for datasets with common paraphrasing and linguistic diversity.

Recent benchmarks indicate that state-of-the-art models, such as BLIP-2 (Li et al., 2023), achieve METEOR scores of 31.2 on the MS COCO dataset, surpassing previous models like mPLUG (Huang et al., 2022), which scored 30.1. These improvements highlight advancements in vision-language pretraining and better alignment between generated and human-written captions.

2.4.5. SPICE

The SPICE (Semantic Propositional Image Caption Evaluation) metric, introduced by Anderson et al. (2016), evaluates image captions by analyzing their semantic content through scene graphs. Unlike BLEU or ROUGE, which focus on surface-level n-gram matches, SPICE decomposes captions into a set of semantic propositions (e.g., objects, attributes, and relationships) and compares these with reference captions. This structured approach to evaluation highlights both the strengths and limitations of automated metrics in image captioning research, reinforcing the need for continuous refinement and human validation in assessing model performance.

Recent evaluations indicate that top-performing models, such as BLIP-2 (Li et al., 2023), achieve SPICE scores of 26.8 on the MS COCO dataset, surpassing previous models like mPLUG (Huang et al., 2022), which scored 25.4. These improvements reflect advancements in multimodal understanding and the ability of modern vision-language models to capture more prosperous semantic relationships in image captions.

2.4.6. Strengths & Limitation

This structured approach to evaluation underscores the strengths and limitations of automated metrics in image captioning research. While these metrics provide a standardized and scalable evaluation framework, they often fail to capture the nuanced aspects of human language and creativity. This limitation underscores the ongoing need for refinement in automated assessments and highlights the irreplaceable role of human validation in ensuring that model performance aligns with real-world expectations. Table 2.1 summarizes these metrics, outlining their key characteristics, including strengths, limitations, and interpretation guidance.

Table 2.1 – Metrics

Metric	Strengths	Limitation	Interpretation
BLEU	Measures n-gram precision, effective for lexical similarity.	Fails to capture semantics and is sensitive to word order.	Models like mPLUG achieve **a BLEU-4 score of 41.0** on MS COCO (Li et al., 2023). Higher scores indicate more substantial lexical similarity.
ROUGE	Focusing on recall evaluates content completeness.	Overemphasizes verbosity and struggles with semantic diversity.	Models like GIT achieve **58.2** ROUGE-L on MS COCO (Wang et al., 2022). Higher scores indicate better recall quality.
CIDEr	Aligns well with human consensus and weights important n-grams.	Limited applicability to diverse linguistic settings. Scores do not follow a fixed 0-100% scale.	Models like BLIP-2 achieve a score of **149.6** on MS COCO (Li et al., 2023). Scores above 120 are considered excellent.
METEOR	Accounts for synonyms, stemming, and word order.	Computationally expensive does not fully capture context.	Models like BLIP-2 achieve a score of **31.2** on MS COCO (Li et al., 2023). Scores above 60% indicate strong alignment.
SPICE	Evaluates semantic content using scene graphs.	Relies on structured scene representations may not fully assess fluency.	Models like BLIP-2 achieve a score of **26.8** on MS COCO (Li et al., 2023). Scores above 25% indicate excellent semantic accuracy.

3. Problem and Modeling

This study aimed to evaluate the impact of classification on the quality of captions in image captioning models. To do so, the "Show and Tell" model was modified by introducing a third input—image classification—processed through a dense layer. This input was combined with image features and text sequences to enrich caption generation with semantic context. Figure 3.1 illustrates the modified architecture, highlighting the added input. Experiments compared the performance of standard and bidirectional LSTM configurations.

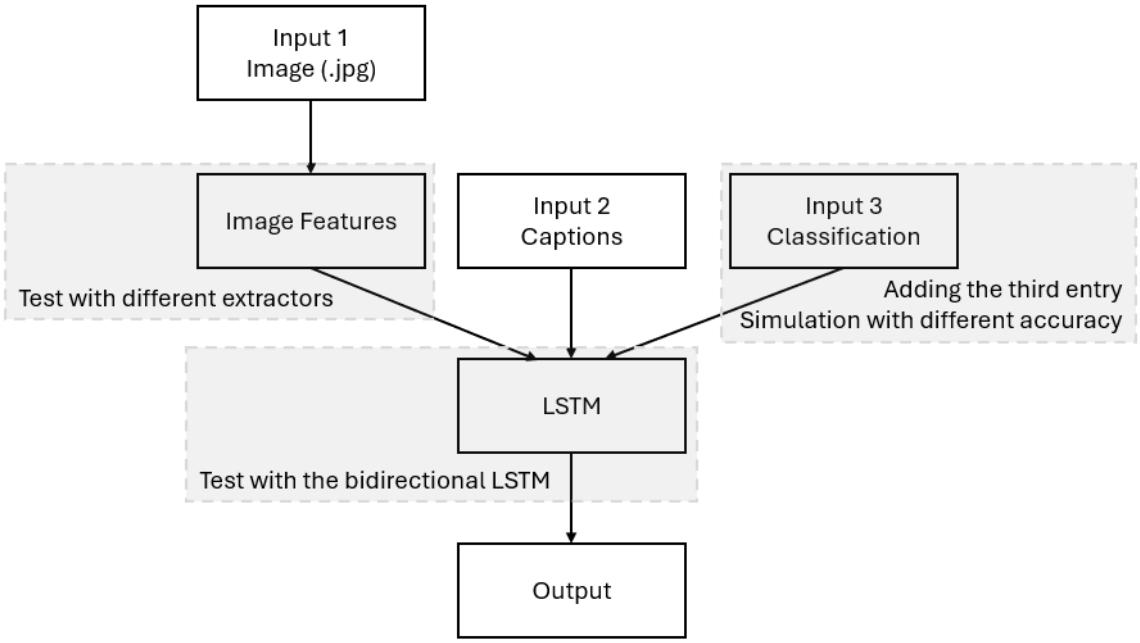

Figure 3.1 – General outline of the model and modifications made.

The following sections outline each stage of the methodology in detail, providing a structured approach to understanding the processes involved in dataset preparation, model development, and evaluation.

3.1. Dataset preparation

Image captioning research has benefited from large-scale datasets designed to improve model training and evaluation. The MS COCO and Flickr datasets are widely used benchmarks in image captioning tasks.

MS COCO comprises over 330,000 photos with multiple human-annotated captions per image, offering a diverse and comprehensive dataset for training and evaluating captioning models. The Flickr datasets, including Flickr8k and Flickr30k, contain images sourced from Flickr with corresponding captions, making them suitable for image-to-text generation research.

Instead of widely used datasets like MS COCO or Flickr30k, this study employed the Oxford Flowers 102 dataset due to its moderate size and rich annotations. This choice aligned with available computational resources and enabled focused evaluation within a specific domain. Details of the implementation using this dataset are discussed in Section 3.1.1.

3.1.1. Dataset Description

The Oxford Flowers 102 dataset, introduced by Nilsback and Zisserman (2008), is a widely recognized benchmark in image classification and recognition tasks. This dataset comprises 8,189 high-resolution images, categorized into 102 distinct flower species, with each species represented by several images ranging from 40 to 258. It was generated by capturing images from diverse environments, including gardens, flower shows, and online sources, ensuring a range of lighting, angles, and backgrounds. The dataset is accompanied by detailed annotations, including class labels and segmentation masks, making it especially valuable for fine-grained image classification tasks. Over the years, it has been extensively used in research to benchmark algorithms in image classification tasks, both supervised and unsupervised learning, as a critical resource for evaluating advancements in computer vision models.

The comprehensive structure of this dataset facilitated its integration into this study. Specifically, the availability of high-resolution images, detailed classifications, and corresponding textual descriptions from the Kaggle competition allowed for the creation of a robust framework. This combination enabled the evaluation of image captioning models with both visual and semantic precision, effectively bridging the gap between classification and captioning tasks.

For example, the study by Wang et al. (2020) used the Oxford Flowers 102 dataset to evaluate a novel attention-based model for fine-grained image captioning, demonstrating its suitability for generating detailed textual descriptions of floral images. Similarly, Zhang et al. (2019) employed the dataset to evaluate a hybrid deep learning approach that integrates classification and captioning tasks, highlighting the dataset's flexibility for multi-task learning scenarios.

The Kaggle DataLab Cup Reverse Image Caption dataset is designed for text-to-image generation, where models transform textual descriptions into corresponding images. Participants develop deep learning architectures and GAN-based models to generate visual representations from single-sentence descriptions, such as "this flower has petals that are yellow and has ruffled stamen." This dataset provides a benchmark for evaluating the effectiveness of text-to-image models in generating realistic images from textual input.

The Oxford Flowers 102 dataset and the complementary textual data from Kaggle competition data were combined into a single dataset. The high-resolution images and classifications were sourced from the Oxford site, while the Kaggle competition contributed textual descriptions for the same pictures. This integration created a comprehensive dataset that supports classification and captioning tasks, thereby enhancing its applicability for domain-specific tasks, such as floral image captioning.

3.1.2. Data Preprocessing

Although the original Oxford Flowers 102 dataset contains 8,189 images, only 7,370 were included in this study. This reduction is because not all images had corresponding textual descriptions in the Kaggle dataset used for captioning. Since the objective of this study requires a direct mapping between image and caption, images without at least one valid caption were excluded during preprocessing. This filtering step ensured that all data points used in training and evaluation contained both visual and semantic information, preserving the consistency and relevance of the dataset for the image captioning task.

The dataset files were organized into directories containing images, caption annotations, tokenized dictionary mappings, and classification labels, ensuring a structured framework for efficient data processing. This structure facilitated seamless access to the necessary components for analysis and pre-processing, particularly in the subsequent steps involving

caption refinement and classification mapping. To enable logging and model tracking, a dedicated directory was created to store training logs.

The classification labels, stored in a .mat file, were loaded and converted into a Pandas DataFrame, mapping each image to its respective category. A separate label mapping file was read to associate numerical class identifiers with their respective flower names. The caption dataset, stored in a .xlsx file, contained tokenized text representations that required a detokenization process to restore readable sentences. A dictionary file mapping token IDs to words was loaded and used to reconstruct captions. A custom function iterated over the dataset, converting token sequences into complete text captions.

Initially, an analysis of the dataset revealed that each image had between 5 and 10 associated captions, with an average of 9.57 captions per image. The variation in the number of captions per image is significant because it can lead to an imbalance in model training. Images with more captions may substantially influence the learning process, while those with fewer captions contribute less. Ensuring a standardized number of captions per image prevents this issue, allowing the model to learn more evenly across the dataset and improving generalization.

In detail, the distribution analysis showed that most images (4,844) had exactly ten captions, while 1,954 images had nine, and only a tiny fraction had fewer captions. Table 3.1 below illustrates the entire distribution of captions per image, ranging from 5 to 10 captions. This summary highlights the degree of imbalance and provides quantitative evidence to support the standardization process.

Table 3.1 summarizes the number of captions per image and their frequency across the dataset, reinforcing the observation of caption imbalance and motivating the need for standardization.

Table 3.1 – Distribution of Captions per Image

Captions per Images	Number of Images
5	1
6	10
7	75
8	486
9	1,954
10	4,844

Further examination of the captions revealed both redundancy in textual descriptions and a limited lexical diversity across the dataset. The total word count was 887,832, comprising only 4,715 unique words, resulting in a lexical diversity score of 0.0053. This low score indicates a high level of repetition in word usage, which can reduce the richness of the dataset and affect the model's ability to learn diverse patterns effectively.

Many captions contained repetitive phrases, which could introduce noise and bias into the training process. This was confirmed by two complementary analyses: word repetition and semantic similarity. The former revealed that 45.77% of all captions had at least one repeated word. One extreme example contained 15 repeated words in a single caption: "light purple and white petals, white and dark purple middle petals, green and yellow middle, dark green leaves."

In parallel, a cosine similarity analysis between captions associated with the same image revealed an average similarity score of 0.25, with values ranging from 0.13 to 0.44. These results indicate that while some semantic overlap exists among captions for the same image, the overall level of similarity is relatively low. This suggests that redundancy in the dataset is primarily lexical rather than semantic in nature. Figure 3.2 illustrates the distribution of similarities across images. If left unprocessed, such lexical redundancy can reduce descriptive diversity and lead the model to generate overly generic outputs, ultimately compromising its ability to generalize effectively.

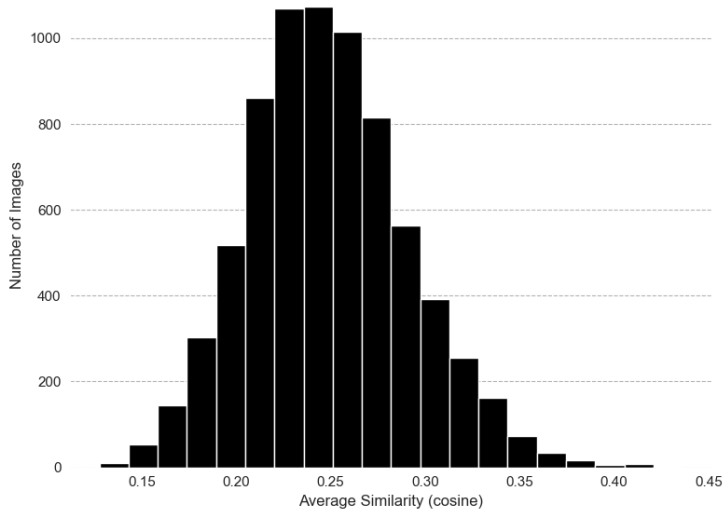

Figure 3.2 – Distribution of Average Caption Similarity per Image

Given these findings, which highlighted variability in the number of captions per image and redundancy in textual descriptions, a refinement step was necessary to ensure consistency in the dataset. To achieve this, a function was implemented to count the number of repeated words within each caption, identifying descriptions with excessive redundancy. The dataset was then sorted by classification labels and word repetition, ensuring that captions with fewer redundant words were prioritized. A selection process was applied to retain the five most representative captions per image, enhancing dataset consistency and caption diversity.

As a result, the refined dataset provided a more balanced and structured input for model training, improving generalization and ensuring that each image contributed equally to the learning process.

A text preprocessing function was applied to enhance further the quality of captions used in model training. This function converts all text to lowercase, removes special characters, eliminates multiple spaces, and filters single-character words. To structure the text data for model training, the function also added the *startseq* and *endseq* tokens at the beginning and end of each caption, providing clear sequence boundaries for the neural network.

Following the text preprocessing step, a tokenizer was initialized and fitted on the captions, transforming words into numerical indices. This process generated a vocabulary where each unique word was assigned to a corresponding index, thereby increasing efficiency in text representation. The total vocabulary size and the maximum caption length were determined, which defined the sequence length for the model.

The images were divided into two subsets to prepare the dataset for training and evaluation: 85% was allocated for training, while 15% was reserved for testing. The split was performed based on unique image identifiers to maintain consistency. Separate data frames were created for both sets, and indices were reset to ensure structured data access. Additionally, the number of unique classification labels was computed to account for categorical variations in the dataset.

Examples of tokenized captions were printed to validate the text preprocessing and tokenization step, illustrating how words were transformed into numerical sequences. Table 3.2 presents sample outputs, demonstrating the relationship between tokenized captions and their corresponding textual representations.

Table 3.2 – Tokenization

Tokenized Caption Sequence	Corresponding Text Caption
[1, 6, 3, 7, 185, 174, 22, 311, 15, 4, 8, 20, 82, 5, 147, 134, 63, 13, 140, 186, 21, 2]	startseq this flower has six leaf shaped off white petals with red spots and four prominent pale-yellow pistils or stamen endseq
[1, 6, 3, 7, 185, 27, 198, 4, 12, 98, 2146, 404, 1095, 18, 302, 17, 5, 29, 25, 2]	startseq this flower has six long tapered petals in an interestingly mottled mix of hot pink and bright orange endseq
[1, 6, 3, 7, 64, 13, 35, 12, 10, 24, 18, 238, 53, 18, 15, 4, 2]	startseq this flower has wide yellow pistil in the center of different layers of white petals endseq
[1, 6, 3, 14, 15, 17, 5, 19, 12, 16, 5, 7, 4, 11, 9, 30, 22, 2]	startseq this flower is white pink and purple in color and has petals that are oval shaped endseq

This table illustrates how tokenization converts each word into a numerical index, ensuring a structured representation of the textual data for model training. This verification step ensured that the vocabulary encoding was correctly applied, maintaining the integrity of the textual data before feeding it into the image captioning model.

3.2.Feature Extraction from Images

Four different pre-trained CNNs were utilized to extract meaningful visual features from images: DenseNet-201, ResNet-50, EfficientNet B4, and VGG-16. The same image captioning model was trained with varying feature extractors to analyze differences in performance.

Each CNN produced feature vectors of distinct dimensionality: DenseNet-201 generated a 1920-dimensional vector, ResNet-50 produced 2048 dimensions, EfficientNet B4 yielded 1792 dimensions, and VGG-16 resulted in a 4096-dimensional feature vector.

The feature extraction process involved using a CNN with its final classification layer removed, allowing the extraction of deep feature representations rather than classification outputs.

3.3.Model Architecture Exploration

Multiple architectures were developed and evaluated to assess the impact of incorporating classification into the image captioning models. These models varied in structure, integrating classification at various stages to determine its effect on caption generation performance.

3.3.1. Base Case

This model builds upon the "Show and Tell" architecture, integrating image features and tokenized captions to generate descriptive text. It takes two inputs: a vector of image features and a sequence of tokenized words. The image features are processed through a dense layer, which reduces their dimensionality and simultaneously learns a meaningful projection aligned with the textual embedding space. These representations are concatenated and passed through an LSTM layer, where the sequential dependencies in the caption are modeled. The image features are then reintroduced to enhance context awareness before the final dense layer generates the output caption.

The model processes image features as a vector of shape and captions as sequences of fixed length (max_length). The image features pass through a Dense(64, activation='ReLU') layer, reducing dimensionality before being reshaped into (1, 64) for compatibility with the textual representation. Captions are embedded into a (max_length, 64) space using an Embedding(vocab_size, 64) layer. The concatenation of these inputs enables the model to establish connections between the visual and textual modalities, which are further refined by an LSTM (64) layer to capture sequential patterns.

The LSTM output undergoes element-wise summation (using the add() function) with the transformed image representation to integrate image features into the caption generation process. This operation ensures that visual and textual components contribute meaningfully to the final prediction. The resulting features are refined through a dense (32, activation='ReLU') layer, followed by dropout regularization to mitigate overfitting. Finally, a dense (vocab_size, activation='SoftMax') layer produces a probability distribution over the vocabulary, predicting the following word in the sequence.

Before feature extraction, images are resized to 224×224×3 and normalized to the [0, 1] range. A pre-trained CNN extracts a feature vector of dimension (1920,) for each image, which is stored for retrieval during training. Captions are tokenized and padded to ensure consistency in input size. The model is trained using a teacher-forcing strategy, where captions are fed word by word to optimize predictive accuracy, following the overall structure illustrated in Figure 3.3.

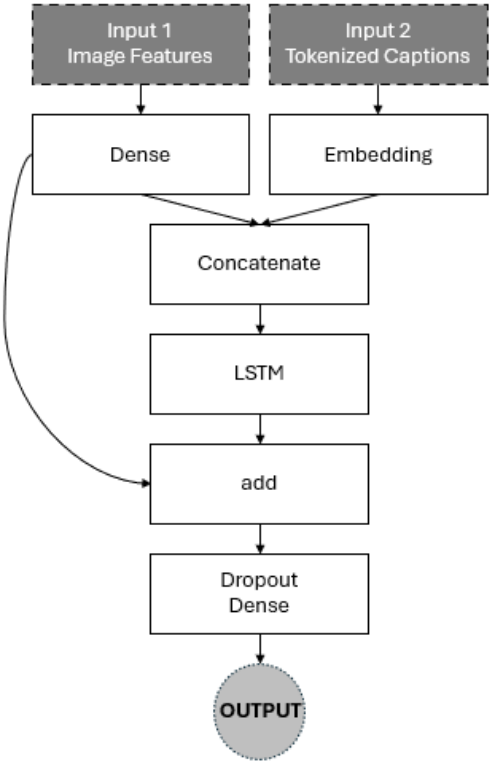

Figure 3.3 – Base Case Model

3.3.2. Base Case with Class

This version extends the previous model by incorporating classification as an additional input to assess its influence on caption generation. The classification input, structured as a one-hot encoded vector representing class labels, is processed through a dense layer with L1 and L2 regularization, batch normalization, and reshaping before concatenating with image features and tokenized captions.

This inclusion allows the model to leverage categorical information, potentially improving semantic alignment in generated captions by grounding them in explicit class-based context. By integrating classification directly into the captioning pipeline, the model gains an additional source of structured information, which may help disambiguate visual elements and refine textual descriptions. This combined representation is then passed through the LSTM layer, where sequential dependencies in the caption are learned.

An accurate label is initially used to isolate the impact of classification, ensuring that any observed improvements stem solely from this additional information rather than classification errors. This enables a controlled assessment of the actual effect of

classification by isolating it from the influence of misclassification noise, thus revealing whether flawless class labels enhance contextual alignment in captions. Future experiments will explore the effects of varying classification accuracy, examining how distinct levels of misclassification affect fluency, coherence, and overall model performance in caption generation.

Apart from this modification, the model architecture remains unchanged. Image features are extracted from a pre-trained CNN, processed through dense layers, and merged with embedded captions within a LSTM framework. The final output, generated by a dense SoftMax layer, predicts the following word based on the learned representations in the sequence, as illustrated in the overall model architecture shown in Figure 3.4.

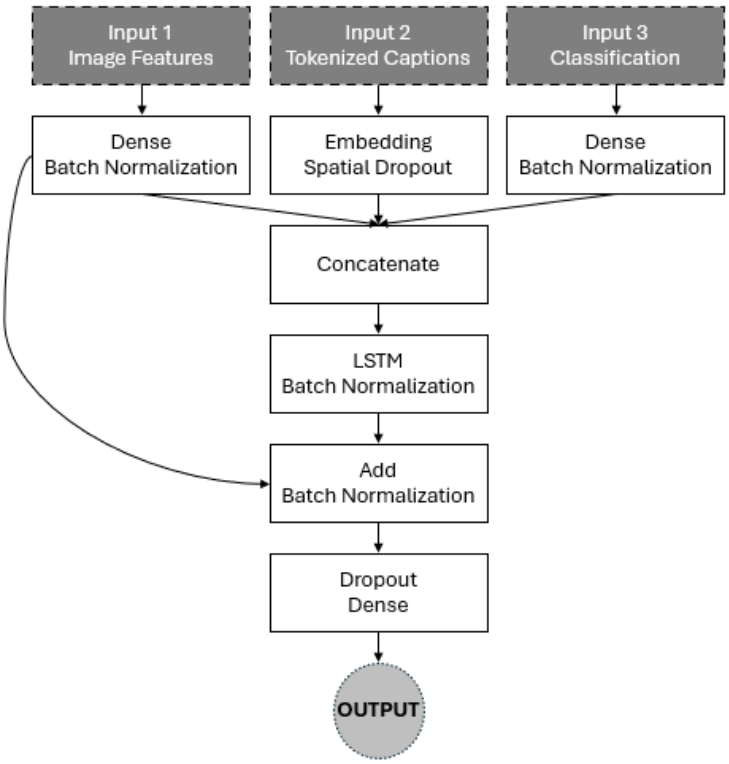

Figure 3.4 – Base Case with Class Model

3.3.3. Base Case with Class and without Residual

This model retains the core structure of the previous version but eliminates the residual connection that previously reintroduced image features after the LSTM layer. By removing this connection, the model relies solely on the sequential processing of textual inputs and the image class to generate captions. This adjustment enables a focused analysis of how much

contextual understanding can be achieved from text alone, without the direct influence of visual layers.

By isolating textual dependencies, it is possible to assess the model's ability to infer descriptive elements from text structure and word relationships, as represented in the architecture depicted in Figure 3.5. Due to the superior performance observed after removing the residual connection, the results for this specific model are not presented separately. Since all subsequent models also exclude this connection, only the results from the other configurations are reported and analyzed.

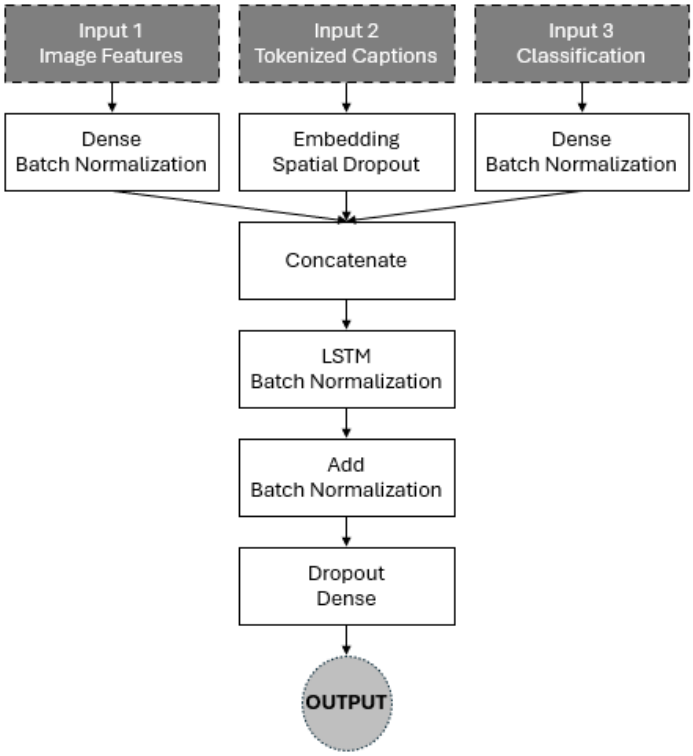

Figure 3.5 – Base Case with Class and without Residual Model

3.3.4. BiLSTM without Class

This variant maintains the core architecture of the BiLSTM model but removes the classification input to assess the specific effects of transitioning from LSTM to BiLSTM. The primary objective of this modification is to isolate the influence of bidirectional processing on caption quality, eliminating any potential confounding factors introduced by categorical information. By doing so, the model clarifies how capturing past and future contextual dependencies affects the fluency, coherence, and accuracy of generated captions.

The comparison between this model and the following versions will highlight whether the inclusion of bidirectional sequence learning enhances the captioning process independently of classification-based features, as represented in the architecture shown in Figure 3.6.

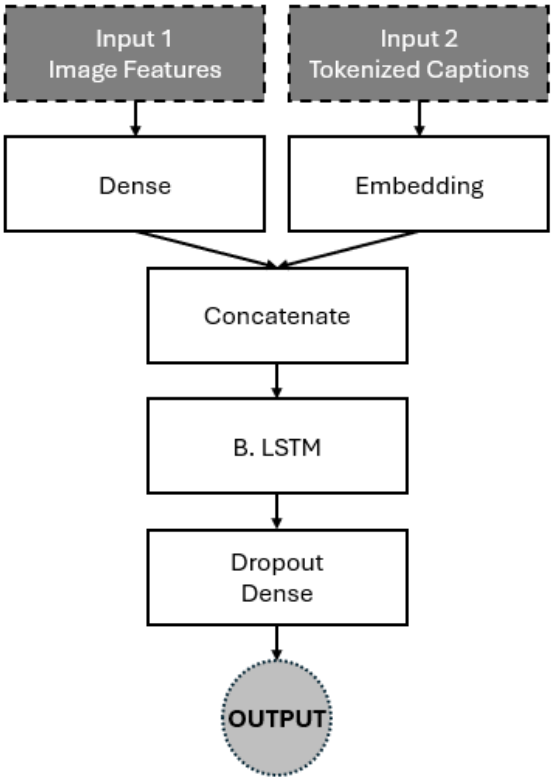

Figure 3.6 – Bidirectional LSTM without Class

3.3.5. BiLSTM with Class

This model maintains the architecture of the previous case but replaces the standard LSTM with a Bidirectional LSTM (BiLSTM). Unlike a conventional LSTM, which processes sequences in a single direction, the BiLSTM captures contextual dependencies from both past and future words within the sequence, thereby improving the model's ability to generate more coherent and contextually rich captions. Additionally, batch normalization and dropout layers are incorporated after the BiLSTM to enhance generalization and mitigate overfitting, as reflected in the architectural layout shown in Figure 3.7.

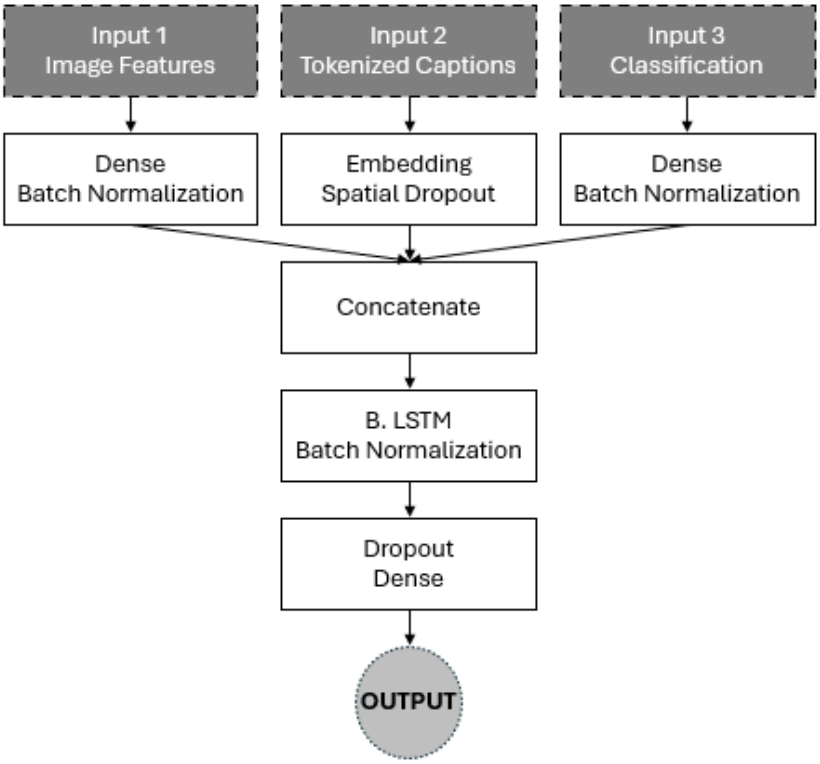

Figure 3.7 – Bidirectional LSTM with Class Model

3.3.6. BiLSTM (Accuracy Variation Simulation)

This variation deliberately alters classification accuracy to assess its impact on caption generation. Instead of using the proper classification labels, artificially modified outputs were generated by introducing controlled errors. A predefined accuracy level determined the proportion of correct labels retained, with misclassifications introduced at 95%, 90%, 85%, and 80% accuracy levels.

The correct classification labels were first duplicated as a baseline to introduce misclassification systematically. Using *keras.utils.set_random_seed()* for reproducibility, a subset of labels was randomly selected via NumPy's *np.random.choice* function. These selected labels were then replaced with incorrect values randomly sampled from the available label set, ensuring a controlled distribution of label noise. This process enabled a structured evaluation of classification errors while maintaining a realistic distribution of errors.

This approach allows a detailed analysis of how varying classification accuracy affects the generated captions. By progressively reducing classification precision, the study examines

whether increased misclassification leads to semantic inconsistencies or impacts fluency and coherence. Additionally, it evaluates whether the model adapts to classification inaccuracies by shifting reliance toward textual context in caption generation.

4. Results and Discussion

This chapter presents the results of the different model configurations and analyzes their implications. The primary objective is to evaluate the impact of integrating classification, consider LSTM variations, and modifications to model architecture on image captioning performance.

Each model configuration was trained thirty-five times to ensure stability and mitigate variability. This repeated training process helps reduce the influence of random factors, such as weight initialization and mini-batch selection, ensuring that the reported results reflect consistent model performance rather than chance variations. Additionally, for each training iteration, the keras.utils.set_random_seed function was set to match the training index (e.g., seed = 1 for the first run, seed = 35 for the last run). This approach ensured reproducibility across experiments while maintaining diversity in initialization, thereby providing a robust assessment of model performance

The first set of experiments focused on evaluating the impact of different CNN-based feature extractors within Base Case architecture. Overall, the results were relatively similar across models, with no extractor showing drastic advantages in captioning performance. This suggests that the captioning model can operate effectively with a range of feature representations.

Four pre-trained networks were compared, with the feature vectors from each served as the sole visual input to the captioning model, allowing an isolated assessment of each extractor's influence on caption quality.

Among all tested configurations, ResNet-50 achieved the highest SPICE score (37.33%) within the Base Case setup (Table 4.1), standing out as the most effective CNN backbone. Although the differences across models were relatively small, this result positioned ResNet-50 as the strongest overall performer across all experimental conditions.

To ensure that any observed improvements in later stages could be confidently attributed to the classification component—and not to variations in the CNN architecture, ResNet-50 was selected as the standard feature extractor for all subsequent experiments. This decision

preserves experimental consistency and avoids introducing architectural bias into the analysis.

Furthermore, prior studies reinforce this selection. Hossain et al. (2019) identified ResNet as a foundational component in many image captioning frameworks due to its balance between complexity and generalization. Sitaula et al. (2021) emphasized its reliable performance across datasets, and Zhou et al. (2022) noted its robustness to noise and adaptability to a variety of image inputs.

Table 4.1 – Average performance of Base Case models with and without classification (SPICE)

Model		DenseNet 201	EfficientNet B4	ResNet50	VGG16
Base Case	Mean	37.20%	37.21%	37.33%	37.02%
	s	*0.17%*	*0.21%*	*0.25%*	*0.27%*
Base Case with Class	Mean	37.18%	36.92%	36.47%	36.68%
	s	*0.79%*	*1.11%*	*3.16%*	*0.11%*
BiLSTM with Class	Mean	35.13%	36.27%	35.55%	35.12%
	s	*0.80%*	*0.76%*	*1.01%*	*1.05%*

Figure 4.1 visually summarizes the SPICE scores across model groups and CNN architectures. Each dot corresponds to the mean performance obtained from 35 independent training runs, while the vertical lines represent the interval defined by the sample mean ± sample standard deviation (± s). This notation not only conveys the central tendency of each configuration but also provides a direct measure of its stability.

The use of the sample standard deviation (s) quantifies the variability across runs and highlights the sensitivity of each model to stochastic training factors such as weight initialization and data shuffling. Configurations with lower s values exhibit more consistent performance, whereas those with higher s indicate greater variance and potentially reduced robustness. This dual representation, mean and dispersion, ensures a more reliable and interpretable comparison of model performance across experimental settings (Hossain et al., 2019; Cornia et al., 2020).

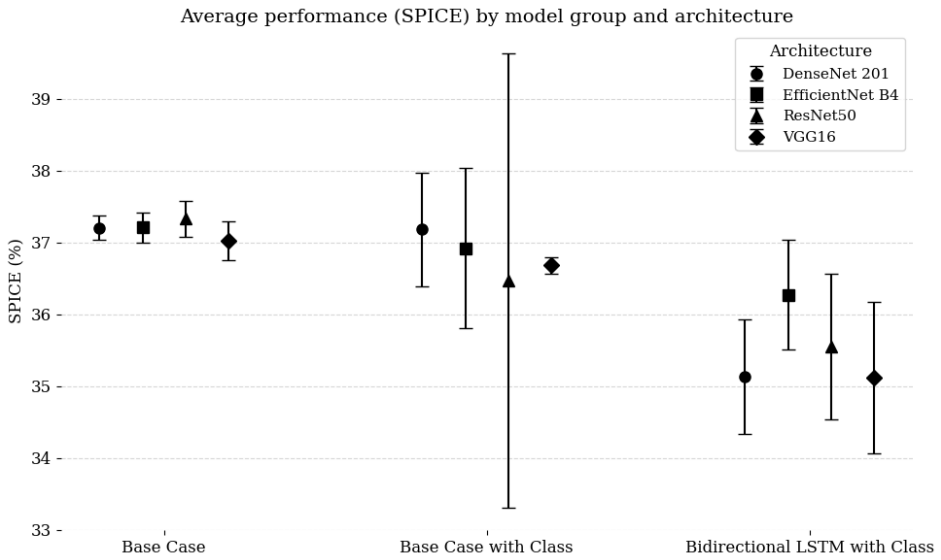

Figure 4.1 – Average performance (SPICE) by model group and architecture

The models discussed in Chapter 3 were evaluated using standard metrics, with a focus on METEOR and SPICE, each offering distinct advantages. This combination of metrics ensures a balanced evaluation, capturing both linguistic accuracy and the meaningful representation of image content.

To evaluate the effectiveness of model modifications, four comparative analyses were conducted: (1) comparing the Base Case model with and without the use of category information; (2) comparing the Bidirectional LSTM architecture with and without category; (3) comparing the Base Case with category to the Bidirectional LSTM with category; and (4) assessing the accuracy of the generated captions using different classification accuracy thresholds. These analyses considered both the average performance across all 35 runs and the best-performing models in each configuration.

The discussion that follows interprets these results in terms of both performance gain (via METEOR and SPICE) and consistency (via standard deviation), aiming to identify the most promising configuration for enhancing the quality of image captions through architectural or input-level adjustments.

4.1. Comparison between Base Case models

The comparison between Base Case with and without category input shows that, on average, the results are quite close, suggesting that the introduction of class information does not drastically alter overall performance.

Specifically, the average METEOR shifted slightly from 19.81% to 19.37%, and SPICE from 37.33% to 36.47%. However, incorporating the class input led to greater variability, as indicated by the increase in standard deviation (e.g., SPICE s = 0.25% → 3.16%). This suggests the model becomes more sensitive to training conditions. Notably, when examining the best-performing runs, SPICE shows a significant improvement with the class-enhanced model, indicating that under optimal conditions, the added classification signal has the potential to enhance performance.

These values are summarized in Table 4.2, which presents the average results across the 35 training runs. In contrast, Table 4.3 displays the best individual run for each model. This dual perspective enables a more comprehensive evaluation, highlighting not only overall stability and reliability but also the potential peak performance of each configuration.

Table 4.2 – Average performance of Base Case models with and without classification

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
Base Case	Mean	42.94%	28.99%	20.64%	14.48%	50.15%	41.52%	19.81%	37,33%
	s	*0,52%*	*0,33%*	*0,35%*	*0,40%*	*2,85%*	*0,18%*	*0,31%*	*0,25%*
Base Case with Class	Mean	41.83%	28.17%	20.03%	13.99%	45.23%	40.48%	19.37%	36.47%
	s	*3.55%*	*2.40%*	*1.75%*	*1.31%*	*5.49%*	*3.42%*	*0.36%*	*3.16%*

Table 4.3 – Best individual performance of Base Case models with and without classification

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
Base Case	Best	43.81%	29.80%	21.15%	15.01%	54.62%	41.85%	20.37%	38.04%
Base Case with Class	Best	43.36%	29.18%	20.83%	14.75%	52.10%	41.58%	19.93%	39.79%

One possible inference for the increase in variance is that the inclusion of class information further increases the complexity to the learning process. Although the classification input uses the true labels and is correctly aligned with the images, the model may become more sensitive to this added semantic signal. This may explain a higher dependency on initialization or other stochastic training factors.

Some prior works (e.g., Hossain et al., 2019; Cornia et al., 2020) suggest that auxiliary inputs, such as class embeddings, require careful integration mechanisms, like gating or attention, to stabilize learning. Without such mechanisms, even valid inputs can create fluctuations in optimization. While this is a plausible explanation, it remains an inference. Nevertheless, the fact that several runs with class input outperformed the baseline supports the potential of category-aware training when well-integrated.

4.2. Comparison between BiLSTM models

In contrast to the Base Case, BiLSTM architecture exhibits more consistent benefits from the integration of category information. The average METEOR increases from 19.62% in the model without class input to 20.67% in the class-informed model, and the best run with category input reaches 21.03%. While the average SPICE score was slightly lower in the model with class input (35.55%) compared to the version without it (37.04%), the results remain relatively close. This indicates that adding class information does not substantially hinder semantic quality. Furthermore, the best SPICE scores are nearly equivalent (37.90% vs. 37.14%), suggesting that, under favorable training conditions, the classification input can perform competitively and even approach the top-tier performance of the baseline model.

Despite the added input, the standard deviation for METEOR in the class-informed model remains low (s = 0.22%), suggesting stable learning across runs. These results are detailed in Tables 4.4 and 4.5, which present the average and best performance, respectively, for this comparison.

Table 4.4 – Average performance of BiLSTM models with and without classification

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
BiLSTM without Class	Mean	42.55%	28.80%	20.55%	14.44%	50.08%	41.41%	19.62%	37.04%
	s	*0.69%*	*0.53%*	*0.59%*	*0.65%*	*5.37%*	*0.42%*	*0.48%*	*0.35%*
BiLSTM with Class	Mean	45.50%	29.12%	19.76%	13.02%	51.70%	40.08%	20.67%	35.55%
	s	*0.32%*	*0.62%*	*0.96%*	*1.20%*	*4.42%*	*0.91%*	*0.22%*	*1.01%*

Table 4.5 – Best individual performance of BiLSTM models with and without classification

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
BiLSTM without Class	Best	43.97%	29.80%	21.44%	15.31%	60.49%	42.05%	20.63%	37.90%
BiLSTM with Class	Best	46.28%	29.94%	21.12%	14.72%	59.77%	41.43%	21.03%	37.14%

4.3.Comparison between Base Case with Class vs. BiLSTM with Class

Comparing both architectures under the condition of category input, BiLSTM outperformed the Base Case. In terms of average performance, METEOR improves from 19.37% in the Base Case to 20.67% in the BiLSTM, while SPICE decreases from 36.47% to 35.55%. Despite this minor drop, the BiLSTM model maintained significantly better stability across runs (s = 1.01% in BiLSTM vs. 3.16% in the Base Case), suggesting a more consistent integration of the class input.

Regarding peak performance, the BiLSTM achieves a maximum METEOR score of 21.03%, outperforming the Base Case's best score of 19.93%. These results are presented in Tables 4.6 and 4.7, corresponding to the average and best-performing configurations, respectively. Overall, this comparison supports the view that the BiLSTM architecture is more effective at leveraging semantic guidance when category information is available, combining improved performance with reduced variability.

Table 4.6 – Average performance of Base Case versus BiLSTM models with classification

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
Base Case	Mean	41.83%	28.17%	20.03%	13.99%	45.23%	40.48%	19.37%	36.47%
	s	*3.55%*	*2.40%*	*1.75%*	*1.31%*	*5.49%*	*3.42%*	*0.36%*	*3.16%*
BiLSTM	Mean	45.50%	29.12%	19.76%	13.02%	51.70%	40.08%	20.67%	35.55%
	s	*0.32%*	*0.62%*	*0.96%*	*1.20%*	*4.42%*	*0.91%*	*0.22%*	*1.01%*

Table 4.7 – Best individual performance of Base Case versus BiLSTM models with classification

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
Base Case	Best	43.36%	29.18%	20.83%	14.75%	52.10%	41.58%	19.93%	39.79%
BiLSTM	Best	46.28%	29.94%	21.12%	14.72%	59.77%	41.43%	21.03%	37.14%

One possible reason for the superior performance of the BiLSTM is its ability to capture context in both temporal directions. This bidirectional flow enables the model to access both preceding and succeeding words during the caption generation process, thereby enhancing coherence and semantic alignment. This architectural advantage has been noted in prior work (e.g., Huang et al., 2015; Liu et al., 2019), where bidirectional RNNs were shown to improve performance in sequence modeling tasks by leveraging complete contextual information. When combined with class information, this architectural advantage may enable better

conditioning across the sequence, contributing to more consistent and context-aware predictions.

4.4.Impact on variation in classification accuracy

To further investigate the effect of classification quality, controlled experiments were conducted using customized datasets with varying classification accuracy levels. These datasets were generated by artificially injecting random errors into the class labels. The simulated accuracies ranged from 100% (using the ground-truth labels) down to 80%, in 5% decrements. This setup allowed for an isolated analysis of how misclassification in auxiliary input influences the quality of the generated captions.

The initial analysis shows that, until reaching approximately 80% classification accuracy, the results remained relatively stable, with no significant variation observed across all evaluation metrics. This suggests a degree of robustness in the model's performance, even under moderate levels of classification uncertainty.

The results show that the best overall METEOR score (21.32%) was achieved at 95% classification accuracy, accompanied by a SPICE of 36.86%. Interestingly, while the best SPICE score observed across all runs (37.45%) occurred at 90% classification accuracy, the average SPICE values generally declined as classification accuracy decreased. This highlights the model's capacity to adapt to favorable scenarios, while also demonstrating a certain robustness under less ideal conditions.

These results are presented in Tables 4.8 and 4.9, which show the average and best scores for each classification accuracy level. In terms of METEOR, both the average and best results remained relatively stable across all accuracy levels, with minimal variation observed. For SPICE, although some fluctuations are present—particularly a slight inversion between 85% and 80% accuracy—the overall trend shows improvement as classification accuracy increases. When examining the best runs, the trend becomes even clearer, with the 90% and 95% classification levels yielding the strongest results. This suggests that higher classification accuracy generally contributes to more semantically rich and consistent captions, despite a few outlier cases at lower accuracy levels.

Table 4.8 – Average performance of the BiLSTM model with different classifications

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
ACC 100%	Mean	45.50%	29.12%	19.76%	13.02%	51.70%	40.08%	20.67%	35.55%
	s	*0.32%*	*0.62%*	*0.96%*	*1.20%*	*4.42%*	*0.91%*	*0.22%*	*1.01%*
ACC 95%	Mean	45.61%	29.13%	19.74%	13.03%	51.32%	40.02%	20.67%	35.50%
	s	*0.64%*	*0.53%*	*0.70%*	*0.84%*	*3.39%*	*0.73%*	*0.23%*	*0.75%*
ACC 90%	Mean	45.74%	29.15%	19.71%	12.95%	51.06%	39.99%	20,70%	35,40%
	s	*0,52%*	*0,51%*	*0,69%*	*0,85%*	*3,42%*	*0,63%*	*0,20%*	*0,69%*
ACC 85%	Mean	45,69%	28,90%	19,36%	12,54%	49,24%	39,70%	20,60%	34,96%
	s	*0,52%*	*0,57%*	*0,74%*	*0,86%*	*3,35%*	*0,76%*	*0,23%*	*0,78%*
ACC 80%	Mean	45,64%	28,98%	19,46%	12,64%	49,81%	39,83%	20,64%	35,19%
	s	*0,42%*	*0,45%*	*0,59%*	*0,71%*	*2,50%*	*0,51%*	*0,16%*	*0,53%*

Table 4.9 – Best individual performance of the BiLSTM model with different classifications

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
ACC 100%	Best	46.28%	29.94%	21.12%	14.72%	59.77%	41.43%	21.03%	37.14%
ACC 95%	Best	47.55%	30.53%	21.10%	14.36%	60.54%	41.77%	21.32%	36.86%
ACC 90%	Best	47.46%	30.52%	21.31%	14.93%	58.91%	41.75%	21.21%	37.45%
ACC 85%	Best	47.16%	29.78%	20.72%	14.20%	55.73%	41.06%	21.02%	36.53%
ACC 80%	Best	46.45%	29.90%	20.40%	13.55%	54.46%	40.50%	20.89%	36.00%

Research by Graves and Schmidhuber (2005) demonstrated that BiLSTM outperform LSTM in tasks that require understanding complex sequential dependencies, especially when inputs are noisy or incomplete. More recently, Zhou et al. (2016) confirmed that BiLSTM can mitigate the negative impact of inconsistent inputs by better modeling semantic relationships in both directions of a sentence.

In the context of this work, this means that even when the auxiliary class label is misclassified, the BiLSTM can still generate coherent and semantically rich captions by drawing from the full sequence context. This behavior reinforces the model's suitability for scenarios where input uncertainty is a concern. Its bidirectional design likely contributes to its ability to contextualize sequences more effectively, which buffers the impact of occasional misclassifications.

From a practical standpoint, these findings suggest that while maintaining high classification accuracy can support more semantically meaningful and consistent descriptions, its overall influence appears limited in this context. The relative stability of results across different accuracy levels indicates that the model is reasonably robust to variations in class input

quality. Therefore, although reliable class information remains beneficial, the system's resilience under moderate classification uncertainty offers promising flexibility for real-world applications.

5. Conclusion

The conclusion of this dissertation synthesizes the findings of a systematic investigation into the integration of classification in image captioning models. Motivated by the growing interest in multimodal learning, this work sets out to determine whether class information could serve as a valuable auxiliary signal to enhance the quality and consistency of generated image captions.

To address this objective, a series of experiments was conducted to evaluate the impact of incorporating image classification into image captioning models. By combining different recurrent architectures (LSTM and BiLSTM) with multiple CNN-based feature extractors, systematic experiments were conducted to evaluate the quality of the generated captions. All models were evaluated using multiple metrics, with a focus on METEOR and SPICE, and each configuration was trained 35 times to ensure result consistency and mitigate random variability.

The results indicated that, although the inclusion of classification does not always lead to consistent improvements, it can provide meaningful gains when properly integrated, particularly in bidirectional architectures. The BiLSTM architecture demonstrated greater stability and performance in scenarios using classification input, outperforming the baseline model in several metrics, especially METEOR. The SPICE metric also showed notable improvements in specific runs, reinforcing the potential of the approach.

Among all CNN backbones evaluated, ResNet-50 achieved the highest average SPICE score and was therefore adopted as the standard extractor for all subsequent experiments. This choice helped isolate the effect of the classification input and prevent architectural bias. Additionally, results from experiments with simulated classification noise showed that model performance remained relatively stable even as classification accuracy decreased to 80%. This indicates a degree of resilience, suggesting that such models may be suitable for deployment in real-world environments where classification uncertainty is expected.

Moreover, the comparison between average and best-performing models provided a broader perspective on architecture stability and peak capability. While average results often showed modest differences, best-run evaluations revealed that classification input could unlock higher semantic quality in some configurations, especially under optimal training conditions.

As a contribution, this work provides a detailed and reproducible analysis of the effects of classification on captioning performance, considering multiple architectures, metrics, and experimental scenarios. It highlights the nuanced role of classification and its potential when combined with robust sequence models such as BiLSTM. The methodology adopted, including seed-controlled training and classification accuracy simulation, contributes to more reliable and generalizable insights.

For future work, it is recommended to explore more sophisticated integration strategies between classification and captioning, such as attention mechanisms, conditional embeddings, or multi-task learning strategies. Investigating transferability to other datasets, extending the classification task to hierarchical or multilabel settings, and incorporating human evaluations are also valuable directions to enhance the relevance and applicability of this line of research.

References

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). SPICE: Semantic Propositional Image Caption Evaluation. *Proceedings of the European Conference on Computer Vision (ECCV), 382–398. https://arxiv.org/abs/1607.08822*

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6077–6086. https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.pdf*

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72. https://aclanthology.org/W05-0909/*

Barnard, K., & Forsyth, D. (2001). Learning the Semantics of Words and Pictures. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), II-408–II-415. https://www.researchgate.net/publication/2371608_Learning_the_Semantics_of_Words_and_Pictures*

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching Words and Pictures. *Journal of Machine Learning Research, 3, 1107–1135. https://www.jmlr.org/papers/v3/barnard03a.html*

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research, 55, 409–442. https://arxiv.org/abs/1601.03896*

Blei, D. M., & Jordan, M. I. (2003). Modeling Annotated Data. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 127–134. https://www.cs.columbia.edu/~blei/papers/BleiJordan2003.pdf*

Caruana, R. (1997). Multi-task Learning. Machine Learning, *28(1), 41–75. https://link.springer.com/article/10.1023/A:1007379606734*

Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325. https://arxiv.org/abs/1504.00325*

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10578–10587. https://arxiv.org/abs/1912.08226*

Cui, Y., Yang, G., Veit, A., Huang, X., & Belongie, S. (2018). Learning to Evaluate Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5804–5812. https://arxiv.org/abs/1806.06422*

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 248–255. https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf*

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2625–2634. https://arxiv.org/abs/1411.4389*

Farhadi, A., Hejrati, M., Sadeghi, M. A. R., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every Picture Tells a Story: Generating Sentences from Images. *European Conference on Computer Vision (ECCV), 15–29. https://link.springer.com/chapter/10.1007/978-3-642-15561-1_2*

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-Shot Learning of Object Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4), 594–611. https://www.sciencedirect.com/science/article/pii/S1077314206001688*

Feng, Y., & Lapata, M. (2010). How Many Words Is a Picture Worth? Automatic Caption Generation for News Images. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 1239–1249. https://aclanthology.org/P10-1126/*

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042SCIRP+1PubMed+1*

Gupta, A., & Davis, L. S. (2008). Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. *European Conference on Computer Vision (ECCV), 16–29. https://link.springer.com/chapter/10.1007/978-3-540-88682-2_3*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735*

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR), 51(6), 1–36. https://doi.org/10.1145/3295748*

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 328–339. https://doi.org/10.18653/v1/P18-1031*

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4700–4708. https://doi.org/10.1109/CVPR.2017.243*

Huang, J., Li, K., Change Loy, C., & Tang, X. (2015). Learning deep representation for imbalanced classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5375–5384. https://doi.org/10.1109/CVPR.2016.580*

Huang, Z., Zeng, Z., & Yu, Z. (2022). Multi-modal transformer fusion for continuous emotion recognition. *IEEE Transactions on Multimedia, 24, 2112–2124. https://ieeexplore.ieee.org/abstract/document/9053762*

Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), 7482–7491. https://doi.org/10.1109/CVPR.2018.00781

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539. https://arxiv.org/abs/1411.2539*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25, 1097–1105. https://doi.org/10.1145/3065386*

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791*

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *European Conference on Computer Vision (ECCV), 121–137. https://doi.org/10.1007/978-3-030-58577-8_8*

Li, Y., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2023). VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557. https://arxiv.org/abs/1908.03557*

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out, 74–81. https://aclanthology.org/W04-1013/*

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005*

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4487–4496. https://doi.org/10.18653/v1/P19-1441*

Lopez-Paz, D., & Ranzato, M. (2017). Gradient Episodic Memory for Continual Learning. *Advances in Neural Information Processing Systems, 30, 6467–6476. https://arxiv.org/abs/1706.08840*

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 375–383. https://arxiv.org/abs/1612.01887*

Mendez, J. A., Kalogeiton, V., & Ferrari, V. (2021). Continual Learning in Visual Search. *arXiv preprint arXiv:2205.13384. https://arxiv.org/abs/2205.13384*

Mori, Y., Takahashi, H., & Oka, R. (1999). Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words. *MISRM '99: Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1–8. https://www.semanticscholar.org/paper/Image-to-word-transformation-based-on-dividing-and-Mori-Takahashi/8b29ffb4207435540ddecf4b14a8a32106b33830*

Nilsback, M.-E., & Zisserman, A. (2008). Automated Flower Classification over a Large Number of Classes. *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 722–729. https://www.robots.ox.ac.uk/~vgg/publications/2008/Nilsback08/nilsback08.pdf*

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs. *Advances in Neural Information Processing Systems, 24, 1143–1151. https://papers.nips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://aclanthology.org/P02-1040/*

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., & Tran, D. (2018). Image Transformer. *arXiv preprint arXiv:1802.05751. https://arxiv.org/abs/1802.05751*

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning, 139, 8748–8763. https://arxiv.org/abs/2103.00020*

Rebuffi, S.-A., Bilen, H., & Vedaldi, A. (2017). Learning Multiple Visual Domains with Residual Adapters. *Advances in Neural Information Processing Systems, 30, 506–516. https://arxiv.org/abs/1705.08045*

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems, 28, 91–99. https://arxiv.org/abs/1506.01497*

Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv preprint arXiv:1706.05098. https://arxiv.org/abs/1706.05098*

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114. https://arxiv.org/abs/2111.02114*

Sener, O., & Koltun, V. (2018). Multi-Task Learning as Multi-Objective Optimization. *Advances in Neural Information Processing Systems, 31, 527–538. https://arxiv.org/abs/1810.04650*

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.1556*

Sitaula, C., Hossain, M. A., & Joshi, L. R. (2021). Attention-based VGG-16 model for COVID-19 chest X-ray image classification. *Applied Intelligence, 51(5), 2850–2863. https://doi.org/10.1007/s10489-020-02055*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–9. https://arxiv.org/abs/1409.4842*

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer *Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818–2826. https://doi.org/10.1109/CVPR.2016.308*

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML), 6105–6114. https://arxiv.org/abs/1905.11946*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008. https://arxiv.org/abs/1706.03762*

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4566–4575. https://arxiv.org/abs/1411.5726*

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3156–3164. https://arxiv.org/abs/1411.4555*

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7794–7803. https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Non-Local_Neural_Networks_CVPR_2018_paper.html*

Wang, W., Chen, C., Wang, W., & Wang, W. (2020). An Overview of Image Caption Generation Methods. *Computational Intelligence and Neuroscience, 2020, 1–14. https://doi.org/10.1155/2020/3062706*

Wang, A., Yu, A. W., Ettinger, D., Grefenstette, E., & Kong, L. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020. https://arxiv.org/abs/2103.00020*

Wang, Y., Xu, J., & Sun, Y. (2022). End-to-End Transformer Based Model for Image Captioning. *arXiv preprint arXiv:2203.15350. https://arxiv.org/abs/2203.15350*

Winograd, T. (1972). Understanding Natural Language. *Cognitive Psychology, 3(1), 1–191. https://www.sciencedirect.com/science/article/abs/pii/0010028572900023*

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning (ICML), 2048–2057. https://arxiv.org/abs/1502.03044*

Xu, J., Ren, X., Lin, Z., & Wang, X. (2021). Image Captioning with Transformer and Knowledge Graph. *Pattern Recognition Letters, 146, 187–193. https://doi.org/10.1016/j.patrec.2020.12.020*

Zhang, Y., Yin, Z., Jin, Z., & Zhang, Y. (2019). Image Captioning via Semantic Element Embedding. *Neurocomputing, Vol. 357, 212–221. https://doi.org/10.1016/j.neucom.2018.02.112*

Zhao, B., Fu, L., & Tao, D. (2017). Image Captioning with Semantic Attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4651–4659. https://doi.org/10.1109/CVPR.2016.503*

Zhou, L., Xu, J., Koch, P., & Corso, J. J. (2016). Watch What You Just Said: Image Captioning with Text-Conditional Attention. *arXiv preprint arXiv:1606.04621. https://arxiv.org/abs/1606.04621*

Zhou, T., Li, S., & Li, B. (2022). Aligned Visual Semantic Scene Graph for Image Captioning. *Image and Vision Computing, 120, 104395. https://doi.org/10.1016/j.displa.2022.102210*