Image Captioning using CNN and RNN

Sai Surya Vidul Chinthamaneni
(SID - 862466578)

Devaki Kalyan Chandra Yadav Podila
(SID - 862468795)

Group Number: 11

Abstract

This project develops a sophisticated image captioning model by integrating Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs), enhanced with soft and hard attention mechanisms. Using the VGG16 architecture for feature extraction and LSTM networks for dynamic text generation, the model focuses on the most salient image features to produce accurate, contextually rich captions. This advanced model was implemented using the TensorFlow-Keras framework and was rigorously evaluated on the Flickr8k dataset, where it demonstrated significant improvements over conventional captioning methods. The effectiveness of our approach is quantified through BLEU scores, indicating optimal caption precision and relevance. These results highlight the model's ability to transform visual data into meaningful textual descriptions, setting a new standard for automated image captioning. Future work will explore integrating more advanced attention mechanisms and expanding the dataset for further improvement.

1 Introduction

Image captioning represents a dynamic intersection between computer vision and natural language processing, aimed at transforming visual data into coherent textual descriptions. This technology significantly enhances the accessibility of digital content for visually impaired individuals by providing them with descriptive narratives of visual scenes they cannot see. Additionally, it plays a crucial role in the digital management of images, facilitating improved indexing and retrieval across extensive media databases, thus aiding in efficient data handling and utilization.

The potential and impact of image captioning are vast, yet the challenge remains to accurately capture and convey the complexity of visual scenes through text. Current image captioning systems often struggle to provide contextually appropriate captions, particularly in complex scenes involving multiple elements and interactions. Our project seeks to address these limitations by deploying advanced deep learning strategies that effectively combine the perceptual strengths of CNNs with the linguistic capabilities of RNNs.

Our innovative model not only uses the renowned VGG16 architecture for extracting robust visual features but also employs LSTM networks to manage the sequence generation for captions. The integration of soft and hard attention mechanisms further refines the focus of the model, allowing it to concentrate dynamically on the most relevant features of the image during the caption generation process. This approach ensures that each generated caption is not only precise but also rich in context.

This project was tested on the Flickr8k dataset, where our model has shown superior performance over traditional captioning models. This success is measured through improved BLEU scores. Through this project, we demonstrate the potential of attention-enhanced deep learning models to revolutionize image captioning, paving the way for more sophisticated and effective captioning systems in the future.

2 Related Work

Image captioning has dramatically changed over the years, transitioning from simple rule-based methods to complex systems powered by neural networks. Initially, image captioning methods were pretty straightforward they recognized objects in images and attached basic descriptions to them using simple language rules. This early approach laid the groundwork for more sophisticated techniques.

A significant advancement came with the work of Vinyals et al, who introduced a model that combines CNNs (Convolutional Neural Networks) for understanding the image and RNNs (Recurrent Neural Networks) for generating text. This model, known as the CNN-RNN framework, was a big step forward because it began to treat image captioning more like a translation problem from the "language" of images to the language of words. Their research proved crucial for future developments. Following this, Xu et al added a new layer of sophistication with attention mechanisms, which helped the model focus on specific parts of the image that are most relevant while generating captions. This method allows the model to produce more detailed and contextually relevant captions by paying closer attention to the important features in an image.

Recent advancements by Anderson et al, have further refined these techniques by introducing both bottom-up and top-down attention processes, enhancing the model's ability to notice and describe finer details and the overall context of the image. Their work shows how blending different types of attention can lead to even better results in understanding and describing images.Each of these developments has helped push the field towards creating more nuanced and context-aware image captioning systems, making them smarter at translating the visual world into descriptive language.

3 Problem Formulation

The primary challenge in image captioning is developing a system that can accurately describe the contents of an image in natural language. This involves not just recognizing objects within the image but also understanding the context in which they appear and how they relate to each other. The goal is to generate captions that are not only factually correct but also contextually rich and fluent, mirroring human-like perception and description capabilities. To address this challenge, we formulated the problem as one of translating visual data into textual data. This approach draws parallels with machine translation, where the task is to convert the "language" of images (visual features and patterns) into words. We identified two main components necessary for solving this problem effectively: feature extraction and sequential text generation.

For feature extraction, we chose the VGG16 architecture, a deep convolutional neural network known for its robustness in capturing complex image features across diverse contexts. VGG16 serves as the backbone for translating raw pixel data into a structured format that our model can interpret. Next, to convert these structured visual inputs into descriptive text, we employed Long Short-Term Memory (LSTM) networks. LSTMs are a type of RNN (Recurrent Neural Network) particularly suited for generating sequences of text due to their ability to remember information over extended periods, and hence, maintaining the context needed for coherent caption generation.

We further enhanced our model by integrating attention mechanisms both soft and hard attention. These mechanisms allow the LSTM to focus on different parts of the image at different times, making the caption generation process more dynamic and context aware. By combining these elements, our solution leverages the strengths of CNNs for visual understanding and RNNs for text generation, enhanced by attention mechanisms to bridge the gap effectively between seeing and describing images.

4 Experimental Results

Our image captioning project includes three types of models: base, base with soft attention, and hard attention. Each model processes image and text data to generate captions:

4.1 Handling the Image

The input layer receives a 512-dimensional vector from VGG16. A dropout layer (0.5) improves generalization by randomly ignoring features. A dense layer (256 units, 'relu') compresses the features to 256 dimensions for further processing.

4.2 Handling the Text

The input layer takes the caption sequence as integers. An embedding layer (256 units) transforms these integers into richer 256-dimensional vectors. A dropout layer (0.5) prevents overfitting, and an LSTM layer (256 units) processes the sequence to maintain context.

4.3 Making the Caption

The combination layer merges processed image and text data. This merged data is passed through a dense layer (256 units, 'relu') and an output layer using softmax to predict the next word in the caption based on the vocabulary.

4.4 Soft Attention Mechanism

Soft attention allows the model to focus on different parts of the image iteratively, thus considering the entire image as a whole more effectively:

- **Attention Weights Calculation**: Uses a mathematical operation (dot product) to calculate how much each part of the 512-dimensional image feature should influence the caption at each step, based on the current word being generated.

- **Context Vector Generation**: It combines these calculated attention weights with the image features to create a 'context vector'. This vector is then concatenated with the current state of the LSTM output, ensuring the model focuses on relevant parts of the image.

4.5 Hard Attention Mechanism and Challenges

Hard attention makes more definitive choices about which parts of the image to focus on:

- **Feature Selection**: Like soft attention, it starts with calculating attention weights but then picks the most significant features directly using a process that identifies the maximum values (using argmax) in the attention scores.

- **Integrating the Focus**: The selected features, based on the highest attention scores, are then directly used to influence the generation of the next word in the caption.

4.5.1 Note on Hard Attention Challenges

Implementing hard attention can be tricky because it involves making discrete choices (specific areas of focus) rather than blending them smoothly as in soft attention. This makes it computationally intensive and more complex to implement correctly, requiring advanced techniques to integrate these choices smoothly within the learning process. For our project, we mimicked this approach to demonstrate its potential, but fully realizing it in practical applications can be significantly more challenging.

4.6 Results

4.6.1 Training Loss

Here below is the training loss obtained when we applied cross entropy loss and updated the parameters using Adam optimizer during our training process of ten epochs. The below figures display how the training loss decreases as the number of epochs increase, showcasing optimal model optimization.

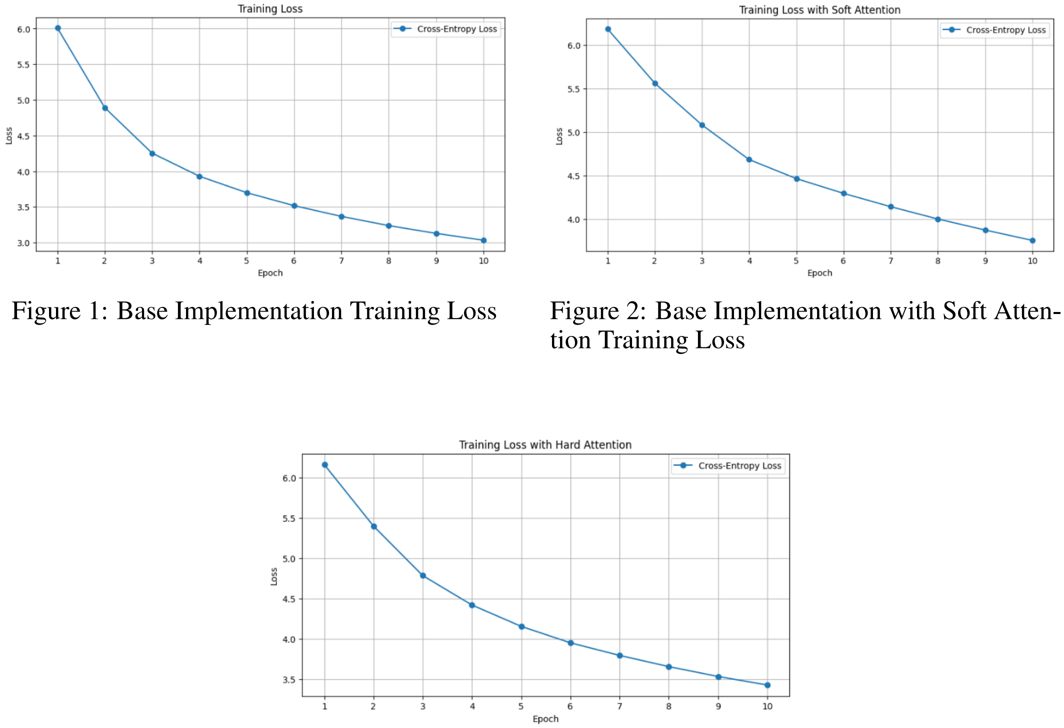

Figure 1: Base Implementation Training Loss

Figure 2: Base Implementation with Soft Attention Training Loss

Figure 3: Base Implementation with Hard Attention Training Loss

Figure 4: Training Loss Comparisons for Base, Soft Attention, and Hard Attention Implementations

4.6.2 BLEU Scores

BLEU (Bilingual Evaluation Understudy) is a metric used to evaluate the quality of text generated by machines, such as translations or image captions. It works by comparing n-grams, which are sequences of words in the generated text, to those in a reference text, providing a quantitative measure of the model's performance. BLEU uses unigrams, bigrams, trigrams, and 4-grams to measure how many matching sequences appear in both the generated text and the reference text. Higher BLEU scores indicate better matches and more accurate text generation.

MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4
BASE	0.2788	0.1206	0.0617	0.0238
SOFT ATTENTION	0.3567	0.1869	0.0973	0.0457
HARD ATTENTION	0.3380	0.1727	0.0883	0.0391

Table 1: BLEU Scores for Different Models

4.7 CAPTIONS GENERATED

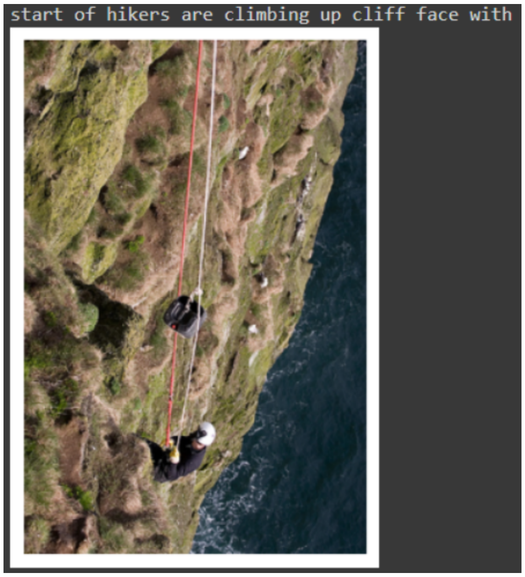
start of hikers are climbing up cliff face with

Figure 5: Enter Caption

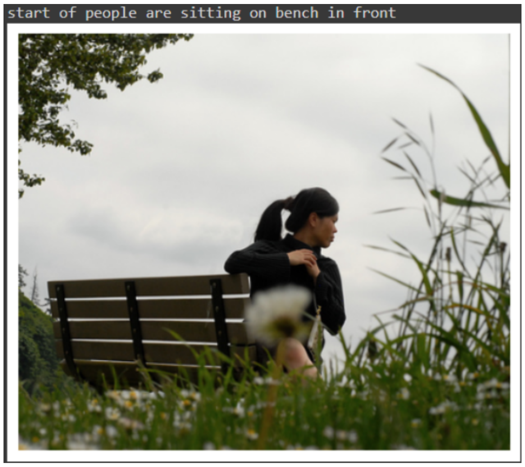
start of people are sitting on bench in front

Figure 6: Enter Caption

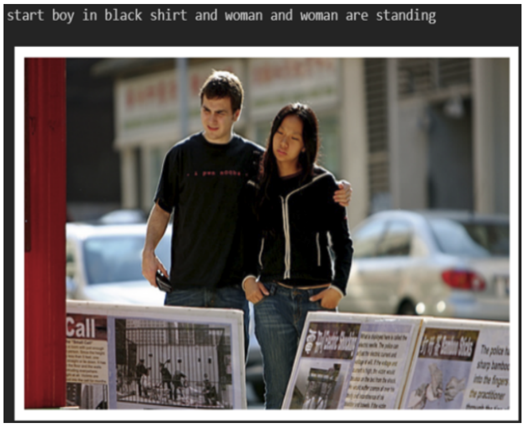
start boy in black shirt and woman and woman are standing

Figure 7: Enter Caption

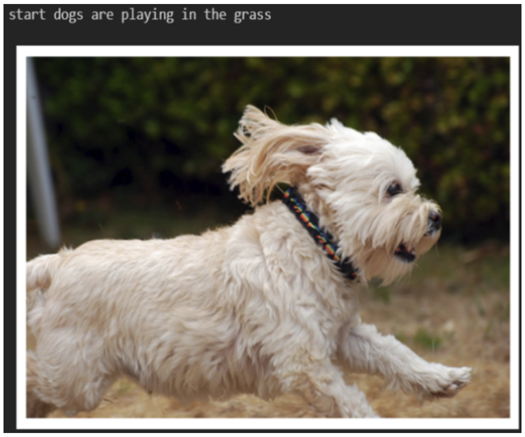
start dogs are playing in the grass

Figure 8: Enter Caption

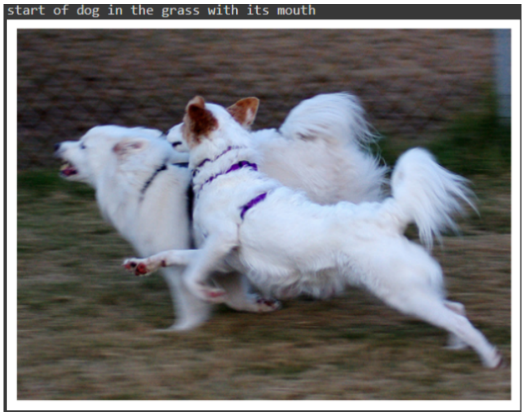
start of dog in the grass with its mouth

Figure 9: Enter Caption

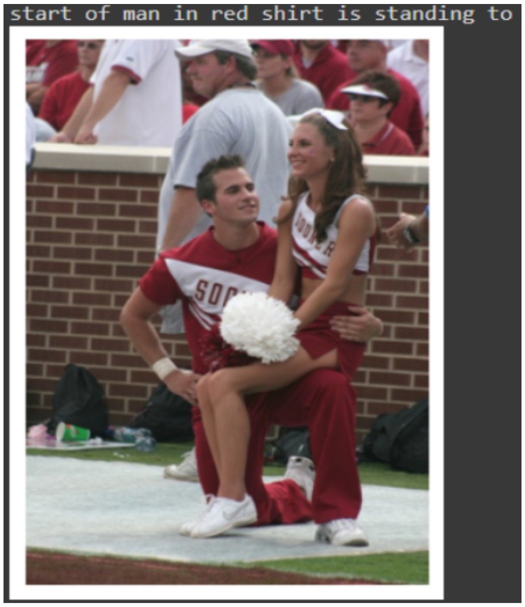
start of man in red shirt is standing to

Figure 10: Enter Caption

5 Contributions

Activity	Description	Kalyan	Vidul
Preprocessing and Helper Functions	Implementation of preprocessing steps to convert the images and texts to embeddings, splitting of data and other required helper functions.	75%	25%
Base Model and Soft Attention	Developed and coded the base model and soft attention mechanism.	60%	40%
Hard Attention, Optimization and Results	Developed and coded hard attention mechanism. Designed and coded experiments to test the different models based on BLEU scores and caption generation.	40%	60%
Proposal, Presentation and Report	Documentation and tabulation of findings. Literature reviews focusing on different aspects of the project : base implementation, soft and hard mechanisms and results.	25%	75%

Table 2: Contributions

6 Acknowledgements

We referred the below papers mentioned in the references to understand how to implement the CNN and RNN methodologies together. The models implemented were our own work. We refered the website of "Analyticsvidhya" to understand how an Image Captioning model can be built, especially how to preprocess the data and convert it to be compatible with our models such as data generation, conversion of texts to word embeddings. We studied and tried to implement the concept of soft and hard attention based on the paper "Show, attend and tell" by Xu, K., et al. (2015).The implementation of the simplified attention mechanisms was our own work. The implementation of BLEU scores and captions generators was also our own work .We referred the paper by Cornia, M, et al to understand what type of BLEU scores are optimal and desirable for image captioning tasks.

7 References

- Xu, K., et al. (2015). *Show, attend and tell: Neural image caption generation with visual attention*. International Conference on Machine Learning.
- Vinyals, O., et al. (2015). *Show and tell: A neural image caption generator*. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Anderson, P., et al. (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.
- Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Springer Science Business Media.
- Lin, T. Y., et al. (2014). *Microsoft COCO: Common Objects in Context*. European conference on computer vision. Springer, Cham.
- Plummer, B. A., et al. (2015). *Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models*. Proceedings of the IEEE international conference on computer vision.

- https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/
- https://www.kaggle.com/datasets/adityajn105/flickr8k
- Cornia, M., Baraldi, L., Serra, G., Cucchiara, R. (2017). *W*here to put the image in an image caption generator. In Proceedings of the CVPR. IEEE.