IMAGE CAPTIONING USING DEEP LEARNING

*Bhoomika Mathakala* *Sanjana Athena*

⋆ University of Central Florida

**1. INTRODUCTION**

Image captioning is the process of using a combination of visual understanding and language processing to generate accurate sentences that describe the contents of an image. This area of research is still relatively new in the field of Artificial Intelligence, and the goal is to find the most efficient way to process an image, represent its content, and convert it into words while maintaining fluency in language. Traditional methods used retrieval and templates, but with the introduction of deep learning and neural networks, particularly with convolutional neural networks and long-short-term memory, a better approach was found. The integration of these networks has allowed for more accurate and fluent image captioning. Recent advances in natural language processing have led to attention and transformer mechanisms becoming standard for resolving difficulties. These approaches have been applied to datasets like flickr-8k and flickr-30k, and results have been presented using metrics like BLEU score. The findings of these studies have been presented both numerically and qualitatively, and the models and code have been made public. In conclusion, deep learning is used to solve the problem of image captioning, and techniques such as attention, long-short term networks, and transformers are employed to achieve this goal. The results are presented using flickr datasets and various metrics from the literature, along with visual examples.

**KEYWORDS**

Image Captioning, Flickr-8k, Flickr-30k, Encoder, Decoder, Convolution Neural Networks, Recurrent Neural Networks, LSTM, Attention, Transformers, BLEU, METEOR, ROUGE-

L, CIDEr

**ACM Reference format:**

Sanjana Athena, Bhoomika Mathakala. 2023.Image Captioning Using Deep Learning. In *Proceedings of ACM Woodstock* conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1234567890

**2. PROBLEM STATEMENT**

we aim to address the challenge of Image Captioning using advanced deep learning techniques such as LSTM, CNN, and Attention-based models. To solve this problem, we will be employing supervised learning, where each dataset for this task consists of an image and its corresponding label. Our primary objective is to develop a deep learning model that can generate outcomes equivalent in quality and statistical performance to those achieved by existing models. By training such models, we can offer accurate image descriptions that can be utilized for a variety of business applications.

**3. TECHNIQUES**

In this section, we will initially outline our network structures and clarify the process of sequence processing in decoders that are linked with them.

**3.1 CNN-LSTM architecture**

It is based on a machine translation model, which we have adapted for the purpose of generating captions by using the hidden features of an input image.

**3.1.1 Description of Encoder:** We utilize the Resnet101 as the primary CNN architecture to extract high-dimensional feature vectors from an input image. Instead of flattening the features, we extract them from a lower convolutional layer, which is then used by the decoder to sequentially predict captions. Our encoder architecture generates a vector of size 2048 features to represent the convoluted image.

**3.1.2 Description of Decoder:** The encoded input image is fed to a decoder that uses LSTM RNNs for sequence processing. The attention-based decoder looks at various parts of the encoded image features to aid in sequence processing. Before predicting the next word, the decoder identifies important areas in the encoder output using an attention network, and at each time step, combines the encoding with weights calculated by the attention network. The attention mechanism focuses on the section of the image that will be used to predict the next word in the sequence.

**3.2 End-to-End Transformers Architecture**

The Transformer architecture is used to convert one sequence to another without using Recurrent Networks. Attention is a critical part of Transformer, and it uses positional encoding to process input before being processed by the network. The experiment uses images as input, and the positional encoding is considered more like spatial encoding. The formula for positional encoding is provided for both the encoder and decoder. The number of encoder and decoder layers is determined by hyperparameters, and keys, values, and queries are pre-designed dimensions. Transformers can extract context from different parts of the same sequence, and the output dimension should match the word embedding size. Pre-trained Glove embeddings of 300 dimensions were used in this case.

**3.3 Loss Function**

To train our models, we typically utilized cross-entropy loss, unless otherwise specified. Our loss function can be summarized as follows: for two-class classification, we used the below equation.

-(y log(p) + (1 - y) log(1 - p))

for multi-class loss calculation, we used the generalized cross-entropy loss, which is represented as

-Σc=1Θo,c log(po,c)

This loss function was specifically used for our architecture.

**3.4 Beam Search**

Beam search is a decision-making algorithm used in NLP and voice recognition models to find the best output based on variables like maximum probability or next output character. This algorithm selects various tokens at a given point in a sequence based on conditional probability and a hyper parameter known as Beam width, which determines the number of optimal choices. Beam search extends the search beyond greedy search by considering alternative words that may match better. It constructs a search tree using breadth-first search and uses the beam width to control the memory usage. The steps for beam search involve setting the beam width, passing tokens, and predicting, and predicting the final output.

**4. RESULTS**

The performance of our CNN+LSTM model on the test set is shown in Table 1, where we have presented the results in terms of score metrics such as BLEU1-4, ROUGE\_L, CIDEr, and METEOR.

| Beam Size | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
| --- | --- | --- | --- | --- |
| 1 | 0.623 | 0.444 | 0.309 | 0.212 |
| 3 | 0.651 | 0.470 | 0.334 | 0.234 |
| 5 | 0.652 | 0.473 | 0.338 | 0.237 |

| Beam Size | METEOR | ROUGE\_L | CIDEr |
| --- | --- | --- | --- |
| 1 | 0.221 | 0.471 | 0.541 |
| 3 | 0.218 | 0.475 | 0.563 |
| 5 | 0.216 | 0.474 | 0.561 |

**Table 1: Performance of CNN + LSTM on Flickr8k**

**5. DISCUSSION**

The experiment evaluated the performance of a machine translation system using different evaluation metrics such as BLEU 1-4, METEOR, CIDEr, and ROUGE\_L. The training was initiated with 50 epochs, and the BLEU-4 score was used to monitor progress. The experiment used a CNN+LSTM model and trained it with datasets Flickr8k and Flickr30k. The evaluation was conducted using different beam search sizes, and it was observed that beam size 3 was the optimal choice for caption generation, based on other scores such as METEOR, ROUGE\_L, and CIDEr. The experiment further experimented with a transformer model, which had a large number of learning parameters, and the performance was poor compared to the CNN+LSTM model. The number of attention heads was found to influence the performance of transformers, and minimizing the encoder and decoder with more attention heads was deemed an optimal choice

6. References

[1] Hossain, MD Zakir and Sohel, Ferrous and Shiratuddin, Mohd Fairuz and Laga, Hamid. A comprehensive survey of deep learning for image captioning, ACM Computing survey (CsUR), 51-6, 2019.

[2] Stefanini, Matteo and Coreia, Marcella and Baraldi, Lorenzo and Cascianelli, Silvia and Fiameni, Giuseppe and Cucchiara, Rita: From show to tell: A survey on image captioning, arXivpreprint arXiv:2017.12114.

[3]“Show and Tell: A Neural Image Caption Generator" by Oriol Vinyals et al. This paper presents an end-to-end system for image captioning that uses a CNN to encode the image and an LSTM to generate the caption.