A Neural Network Approach to Image Captioning Using Multimodal Data on Flickr8k

Melike Çolak
Faculty of Science
Hacettepe University
Ankara, Türkiye
n22239753@cs.hacettepe.edu.tr

I. Abstract

Image captioning, a critical task in computer vision and natural language processing, involves generating descriptive text for images. This study introduces a neural network model that integrates image and sequence features for this purpose. The model's architecture includes two branches: one processes image features using a pre-trained CNN, and the other handles sequence data through an embedding layer and LSTM network. These features are merged and passed through dense layers to generate captions. Trained over 100 epochs with categorical cross-entropy loss and the Adam optimizer, the model achieved a training loss of 1.446. BLEU scores of 0.516880 (BLEU-1) and 0.293009 (BLEU-2) reflect its ability to produce coherent and contextually relevant captions. Qualitative evaluations show that while the model generates accurate descriptions, it sometimes generalizes specific details. This model demonstrates effective multimodal integration for image captioning, suggesting its potential for automated image description tasks. Further improvements could enhance detail accuracy and overall performance. Code and data are available on my GitHub page.(https://github.com/melikecolak/image-captioning)

II. Introduction

Image Captioning is the task of translating an input image into a textual description. As such, it connects Vision and Language in a generative fashion, with applications that range from multi-modal search engines to help visually impaired people. Although recent years have witnessed an increase in accuracy in such models, this has also brought increasing complexity and challenges in interpretability and visualization. In this work, the primary objective of this project is to improve the accuracy and diversity of image captioning systems using the Flickr8 dataset [1]. Image captioning involves generating natural language descriptions for images, which presents several challenges including understanding image content, and context, and generating coherent and relevant captions. While existing image captioning models have shown promising results, they often lack diversity and may generate generic captions that fail to capture nuanced details in the images.

Accurate and diverse image captioning is crucial for various applications such as assistive technologies for visually impaired individuals, content retrieval, and human-computer interaction. By improving the quality and diversity of image captions, we can enhance these applications' user experience and usability. Additionally, advancing image captioning techniques contribute to the broader field of computer vision and natural language processing, pushing the boundaries of AI research. This project aims to address these limitations by leveraging the Flickr8 dataset to enhance the quality and diversity of image captions.

III. Literature Review

Existing image captioning approaches typically employ encoder-decoder architectures, often enhanced with attention mechanisms, to generate captions based on image features. These methods have shown promising results but may suffer from issues such as lack of diversity and relevance in generated captions. This project builds upon these existing methods by leveraging the Flickr8 dataset and exploring novel techniques to address these limitations.

Image captioning is the task of generating a short text description of the content of an image. It has various applications, such as making visual content accessible to the visually impaired and improving visual search and categorization tasks. Image captioning is an active research area, and new methods are constantly being developed. The state-of-the-art can generate captions that are both accurate and fluent, in terms of the content of the image and the grammar and semantics of the language. Image captioning has two approaches, knowledge-based and learning-based. Knowledge-based methods use knowledge bases, such as WordNet, to define objects and concepts. An image processing method is used to identify objects in the image, and then these objects are matched with the knowledge base to generate a caption. Learning-based approaches: These methods use models trained on large datasets of image-caption pairs. These models learn to generate new captions by considering both the content of the image and the grammar and semantics of the caption.

When the literature was examined, Vinyals et al. [2] presented one of the first deep-learning models for image captioning. The model uses an encoder-decoder architecture and learns to match the image and the caption. It is quite successful in generating captions that are consistent with the content of the image and grammatically correct. Xu et al.

[3] propose a model that learns to attend to different parts of the image. This allows the model to focus on the most important parts of the image and generate more accurate and detailed captions. Mao et al. [2] propose a model that uses two recurrent neural networks (RNNs) to encode the image and the caption together. This allows the model to better learn the temporal relationships between the image and the caption and generate more consistent captions.

Several academic papers have utilized the Flickr [1], [4] datasets to advance the field of image captioning. For instance, a study by Mao et al. [5] introduced Flickr30k Entities, augmenting the original Flickr30k dataset with detailed region-to-phrase correspondences, improving the granularity of image-to-sentence models. Another paper [6] explored using deep learning with the Flickr8k dataset to generate descriptive captions, leveraging a combined CNN and LSTM architecture to enhance caption accuracy and contextual relevance. Additionally, research on using the ResNet-50 model pre-trained on ImageNet with Flickr8k demonstrated significant improvements in generating meaningful captions due to effective feature extraction capabilities [7]. Chen and Zitnick [8] developed a recurrent visual representation model for image caption generation, which leverages both visual and semantic features to produce human-like descriptions, demonstrating significant improvements on the Flickr30k dataset. Further research by Ghneim et al. [9] explored the use of Visual Attention Prediction Networks (VAPN) and Contextual Spatial Relation Extraction (CSRE) for image caption generation, achieving improved performance on the Flickr8k and Flickr30k datasets. Yu et al. [10] introduced a dual attention mechanism on pyramid feature maps, which better localizes visually indicative regions and improves the quality of generated captions, as validated on the Flickr8k, Flickr30k, and MS COCO datasets.

These studies collectively highlight the advancements in image captioning through the integration of sophisticated neural network architectures, attention mechanisms, and comprehensive datasets, paving the way for more accurate and contextually relevant automated image descriptions.

IV. Proposed Methodology

The proposed model architecture utilizes an encoder-decoder framework with multimodal fusion for image captioning and it uses the Flickr8 image captioning dataset. This methodology outlines a comprehensive approach to developing an image captioning system using deep learning techniques. By leveraging the powerful feature extraction capabilities of the VGG16 model and the sequence generation strengths of LSTM networks, the proposed method aims to produce high-quality captions that accurately describe the content of images.

A. Dataset

The project utilized on the Flickr8 dataset, which consists of over 8,000 images with multiple captions per image. This dataset offers a diverse range of images across different categories, providing sample training data for developing robust image captioning models. In this project, some data

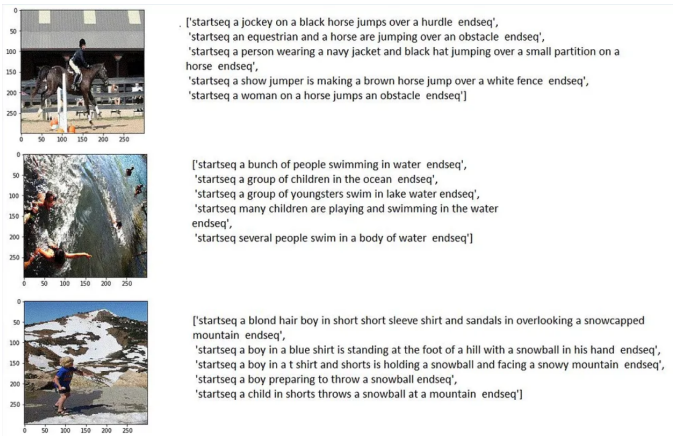

Fig. 1. An example of an image from the Flickr8 dataset and its description.

preprocessing techniques were applied to the dataset. Such as preparing the Flickr8 dataset by preprocessing images (e.g., resizing, normalization) and tokenizing captions. Data preprocessing and cleaning is an important part of the whole model-building process. Understanding the data helps us to build more accurate models.

The Flickr8k dataset contains a total of 8092 images in JPEG format with different shapes and sizes. Of which 6000 are used for training, 1000 for test and 1000 for evaluation. Also, it contains text files describing train_set and test_set. Flickr8k.token.txt contains 5 captions for each image i.e. total of 40460 captions. In Figure 1, an example of an image and it is description is shown.

B. Model Architecture

The core of the proposed image captioning system is based on a Long Short-Term Memory (LSTM) network, designed to handle sequential data and capture long-term dependencies. The model architecture integrates two main components: the image feature extractor and the sequence generator.

The image feature extractor processes the input images and generates fixed-size feature vectors. This component utilizes the VGG16 model, effectively transferring knowledge from the ImageNet dataset to our specific task. The extracted feature vectors are then fed into the sequence generator, an LSTM network, which generates captions word by word. The sequence generator includes an embedding layer to convert words into dense vectors, a dropout layer to prevent overfitting, and dense layers to transform the combined features into a probability distribution over the vocabulary. The LSTM network is designed to take the image features and the previous words of the caption as input, predicting the next word in the sequence. This recursive process continues until a special end-of-sequence token is generated, indicating the completion of the caption. The architecture of the model is detailed in Figure 2. Below, we provide an in-depth description of each component in the architecture.

The image feature extraction begins with an input layer that accepts a feature vector of size 4096. This vector typically

represents precomputed features from an image, extracted using a pre-trained convolutional neural network (CNN). The model accepts a 4096-dimensional input vector representing image features. To prevent overfitting, a Dropout layer with a dropout rate of 0.4 is applied. This randomly sets 40% of the input units to zero at each update during training time, which helps regularize the model. A fully connected Dense layer with 256 units and ReLU activation is used to transform the input features into a more compact and useful representation.

The sequence feature extraction involves embedding textual input and processing it through an LSTM network. The model accepts a sequence of integers of a specified maximum length (max_length), which corresponds to the tokenized and padded text input. The input sequence is passed through an Embedding layer that maps each integer (word index) to a dense vector of fixed size 256. The mask_zero=True parameter ensures that zero-padding is ignored during the training. Similar to the image feature branch, a Dropout layer with a 0.4 rate is applied to the embedded sequence to avoid overfitting. The sequence is then processed by an LSTM (Long Short-Term Memory) layer with 256 units. This layer captures the temporal dependencies in the sequence data and outputs a fixed-size vector.

The decoder model combines the processed image and sequence features to generate the final output. The outputs from the image and sequence processing branches (both 256-dimensional vectors) are combined using an element-wise addition operation. This effectively merges the two modalities into a single representation. The combined feature vector is then passed through another Dense layer with 256 units and ReLU activation to refine the representation further. Finally, a Dense layer with a softmax activation function produces the output. The size of this layer corresponds to the vocabulary size (vocab_size), which determines the probability distribution over possible output classes.

The model is compiled using the categorical cross-entropy loss function, suitable for multi-class classification problems. The Adam optimizer updates the model weights during training, providing efficient and adaptive learning.

C. Training and Evaluation

The training process for the model involved several key parameters and steps, ensuring that the model learned to generate accurate and contextually relevant sequences from the given data. The following parameters were used during training:
-Number of epochs: 100
-Batch size: 32
-Vocabulary size: 8485
-Maximum sequence length: 35

The training process was executed over 100 epochs, wherein the model was trained using a data generator for each epoch. The data generator dynamically provided batches of training data to the model, ensuring efficient utilization of memory and computational resources. The model was trained using

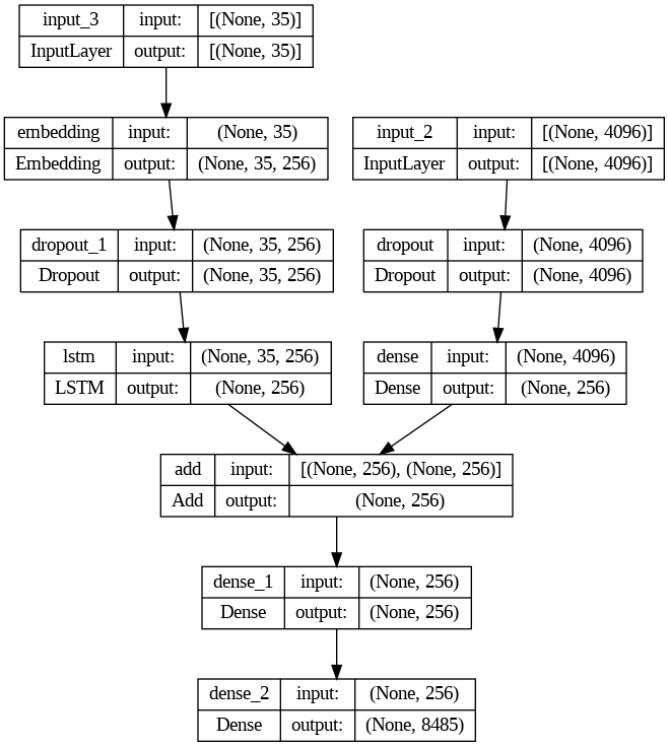

Fig. 2. Image captioning model architecture.

the categorical cross-entropy loss function, which is well-suited for multi-class classification tasks. The Adam optimizer updated the model weights, offering an adaptive learning rate for faster convergence.

In each epoch, the data generator supplied the model with batches of data consisting of image features, sequence inputs, and the corresponding target sequences. This approach ensured that the model was exposed to the entire training dataset multiple times, allowing it to learn and refine the relationships between the image and sequence inputs effectively. After 100 epochs of training, the model achieved a training loss of 1.446. This indicates that the model has successfully minimized the loss function, learning to predict sequences that are closer to the actual sequences in the training data. The training loss value reflects the model's ability to fit the training data, with lower values indicating better performance.

To evaluate the model's performance, the BLEU (Bilingual Evaluation Understudy) score is used. The BLEU score assesses the quality of the generated captions by comparing them to human-generated reference captions, considering factors such as precision and n-gram overlap. This metric provides a quantitative measure of how closely the generated captions match the ground truth, with higher scores indicating better performance. Evaluation is conducted on the test set, ensuring that the results reflect the model's generalization capability and its effectiveness in generating accurate and coherent image descriptions.

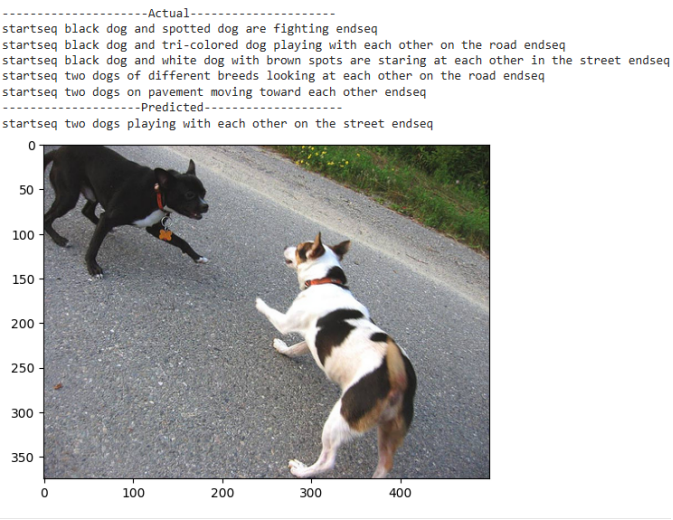

Fig. 3. Example from the test set, comparing the actual and predicted sequences generated by the model.

V. Results

Upon training the model, we evaluated its performance using BLEU scores, a common metric for assessing the quality of text generated by machine learning models. The BLEU (Bilingual Evaluation Understudy) score ranges from 0 to 1, with higher values indicating better performance. For our trained model, we achieved the following BLEU scores:

BLEU-1: 0.516880 BLEU-2: 0.293009

These scores suggest that the model performs reasonably well in generating text sequences that are similar to the reference sequences, particularly for unigram matches (BLEU-1). However, the score decreases for bigram matches (BLEU-2), indicating some limitations in capturing more complex dependencies within the sequences. To provide a more qualitative understanding of the model's performance, we present an example from the test set in Figure 3, comparing the actual and predicted sequences generated by the model.

The predicted sequence " two dogs playing with each other on the street " demonstrates that the model can capture the general theme and context of the actual sequences, focusing on the interaction between the dogs. However, it lacks some of the specific details present in the actual sequences, such as the colors of the dogs and their specific actions (e.g., "fighting" vs. "playing"). This indicates that while the model is effective at generating coherent and contextually relevant descriptions, it sometimes generalizes the information, leading to less detailed outputs. Overall, the results highlight the model's ability to integrate and process multimodal data, generating meaningful and contextually appropriate text sequences. However, there is room for improvement, particularly in enhancing the model's ability to capture and reproduce more detailed and specific information from the input data.

VI. Conclusion

This study introduced a neural network model for image captioning that integrates image and sequence features. The model architecture includes two branches: one for image features using a pre-trained CNN, and another for sequence data through an embedding layer and LSTM network. These features are merged and processed through dense layers to generate captions. Trained on the Flickr8k dataset for 100 epochs, the model achieved a training loss of 1.446 and BLEU scores of 0.516880 (BLEU-1) and 0.293009 (BLEU-2). These results indicate that the model can generate coherent and contextually relevant captions, although it sometimes generalizes specific details.

Overall, the model demonstrates effective multimodal integration for image captioning and holds promise for automated image description tasks. Future enhancements could focus on improving the model's ability to capture finer details for even better performance.

References

[1] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," in *Journal of Artificial Intelligence Research*, vol. 47, 2013, pp. 853–899.

[2] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[4] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "Flickr30k: Collecting image annotations from the crowd," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1961–1972.

[5] J. Mao and et al., "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2641–2649.

[6] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[7] J. Donahue and et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[8] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422–2431.

[9] N. Ghneim and et al., "Image caption generation using visual attention prediction and contextual spatial relation extraction," *Journal of Big Data*, vol. 7, no. 1, pp. 1–18, 2020.

[10] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020, pp. 5805–5813.