First Submit – Sept 28, 2024

Project Proposal: Image Captioning with Attention Mechanism

Idea: Develop a model that generates descriptive captions for images using a combination of CNN and LSTM with an attention mechanism.

1. APPLICATION CONTEXT

Image captioning is a task at the intersection of Computer Vision and Natural Language Processing, where the aim is to generate a textual description of an image (Xu et al., 2023). The process involves translating an image (visual input) into a sequence of words (text output) that accurately describe the content of the image.

The dominant framework for image captioning is the Encoder-Decoder architecture, inspired by sequence-to-sequence models in machine translation (see Fig. 1). In this framework, the encoder (typically a pre-trained CNN) extracts feature maps from the input image into a fixed-size feature vector, which are then passed to the decoder (an LSTM or other RNN variant) that outputs a sequence of words forming the caption (Vinyals et al., 2015).

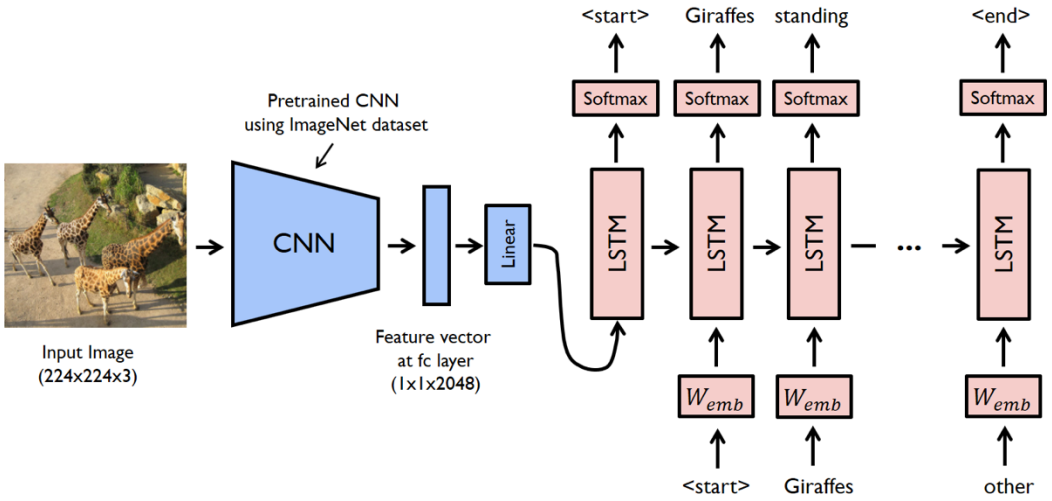

Figure 1. Encoder-Decoder Framework for Image captioning (Gaurav & Mathur, 2021).

However, this approach has limitations, as it condenses the entire image into a single vector, potentially losing spatial information that might be critical for generating accurate descriptions.

The attention mechanism addresses the limitation of fixed-size representations, it works by dynamically weights different spatial regions of the image, allowing the model to attend to the most relevant areas as each word of the caption is generated (Xu et al., 2015). That way, the model learns an "attention map" over the image's spatial features, guiding the decoder to focus on relevant regions while generating each word.

Image captioning has a wide range of real-world applications, such as generating descriptions for visually impaired individuals, improving image search engines, and automatically generating annotations for large-scale datasets (Xu et al., 2023).

2. MACHINE LEARNING OBJECTIVE

The primary objective of this project is to predict a descriptive caption for an image, given its visual features. To achieve this, a model will be created that, given an image, can generate captions that are contextually accurate and semantically meaningful using both CNN for feature extraction and LSTM networks with an attention mechanism.

Formally, we aim to predict a sequence of words $Y = \{y_1, y_2, \dots, y_T\}$, where y_T is the word at time step t and T is the length of the caption, given an input image $\mathcal{X}$, using the image feature map extracted by the CNN I.

3. DATASET

The dataset purposed in this project is the popular **Flickr8k**. It consists of images with a diverse distribution of objects and scenes, which helps generalize the model's ability to describe different kinds of visuals.

It is made up of 8,000 images (see Fig. 2), each image is paired with five different captions describing the content, providing varied descriptions which adds richness for training models. Each image in the dataset is an RGB image stored in JPEG format, while the captions are plain text sentences stored in TXT format.

The dataset takes up approximately 1 GB of storage in total. The data will be split into training, validation, and testing sets using an 80-10-10 ratio.

Figure 2. Data sample from Flickr8k dataset.

4. PERFORMANCE METRICS

For evaluating the performance of the image captioning model, the following combination of machine learning metrics and business metrics will be used:

Machine Learning Metrics

- **BLEU score**: A common metric in NLP tasks used to measure the similarity between the generated text and the reference text. It measures how many words or sequences of words (n-grams) overlap. A higher BLEU score indicates a closer match to the reference text, and higher-order n-gram captures more context, and matching longer sequences of words.

Business/Impact Metrics

This system automatically generates descriptions from product's images for large online retailer. To assess the system's performance the metrics purposed are Time-to-Market (Time Savings %), Cost Savings on Content Creation (%) and Sales Increment (%). The model deployment will be conducted via an A/B test on products with captions generated by the model and others written manually.

5. REFERENCES AND PREVIOUS RESULTS

Recent studies on image captioning have reported varying performance across different architectures. A simple CNN as the encoder and an LSTM as the decoder, combined with an attention mechanism, achieved a BLEU-1 score of 61% and a BLEU-4 score of 19.5%, demonstrating modest performance in caption generation (Gaurav & Mathur, 2021). In contrast, transformer-based models have set new benchmarks in image captioning, achieving significantly higher performance, with BLEU-1 scores reaching 96% and BLEU-4 scores as high as 72.8% (Xu et al., 2023).

Gaurav and Mathur, P. (2021). A survey on various deep learning models for automatic image captioning. *Journal of Physics: Conference Series*, 1950(1), p. 012045. doi:10.1088/1742-6596/1950/1/012045.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
This paper introduces a similar encoder-decoder architecture using a CNN and an LSTM, showing that such models can produce high-quality captions.

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. *International Conference on Machine Learning (ICML)*.
This paper adds an attention mechanism to the image captioning model, allowing the model to focus on different parts of the image when generating each word in the caption. The attention mechanism is shown to significantly improve the quality of generated captions.

Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X., & Li, W. (2023). Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing, 546,* 126287. https://doi.org/10.1016/j.neucom.2023.126287
This paper covers a wide range of methods, architectures, and challenges in the field, providing a deep dive into various models and datasets used in image captioning.