End-to-End Pre-training for Image Captioning

Artan Vafaei, Sowresh Mecheri-Senthil, Anthony Nguyen, Michael Tanjuakio
The University of Texas at Dallas
asv210003@utdallas.edu sxm220283@utdallas.edu ahn200000@utdallas.edu mat200000@utdallas.edu

1. Problem Statement

Most modern-day Vision-Language models [2] rely on combining separately pre-trained Vision and Language models, which generally leads to good performance with less computational power. However, this approach has several limitations: the models may not be optimal for specific tasks, are dependent on the availability of high-quality pre-trained data, and integrating independently trained components can result in inefficiencies. Additionally, the initial pre-training process can be computationally expensive.

2. Approach

For our approach, it is research-oriented. This approach will be novel since training a Resnet and GPT2-like model together is usually not done due to computational constraints [1, 3]. However, we aim to show that this approach, after undergoing some modifications, is still viable for image captioning tasks. This approach will allow the model to learn more specific features in the dataset, as the Resnet is training in conjunction with the GPT2 model. We introduce an end-to-end pipeline for image captioning tailored for specific tasks, eliminating the need for large datasets or complex models. The proposed model features a ResNet-like architecture that generates image embeddings, which are then passed through a linear projection into language embedding space. This embedding is subsequently fed into a GPT-2-like language model, enabling the generation of accurate captions. If time permits, we will implement cross-attention to condition the generated captions on the image embeddings, enhancing the model's ability to integrate visual context and improve caption relevance.

3. Data

We use image caption pair dataset generated by GPT4. NOTE: The dataset was generated by GPT4, however, we are using an existing dataset on the internet. Here is the link: https://huggingface.co/datasets/laion/gpt4v-dataset

4. Evaluation

We will use the BertScore metric to determine our model's performance [5]. BertScore will compare the semantic meaning between the generated text and the ground truth text.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. 1

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 1

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 1

[4] [2]