COSE474-2024F: Final Project Proposal
Image Captioning using CLIP and GPT

Taha El Bouzidi

1. Introduction

Automatically generating descriptive captions for images has significant applications in accessibility, image indexing, and human-computer interaction. This project aims to leverage pre-trained models such as CLIP (Radford et al., 2021) (for visual features) and GPT (for text generation) to develop an improved image captioning system. Recent works like "Show, Attend, and Tell" (Xu et al., 2015) have made progress in this area, but they often require extensive training on domain-specific data, which is a limitation.

2. Problem Definition & Challenges

The task is to generate natural language descriptions for images using pre-trained models. The main challenge is to align CLIP's image features with GPT's text generation in a way that produces accurate and contextually relevant captions. The models should also generalize well to unseen images without overfitting.

3. Datasets

The MS-COCO dataset (Chen et al., 2015) will be used for evaluation, as it contains over 330,000 images with human-written captions. It offers a suitable benchmark for testing the quality and fluency of generated captions.

4. Goals

The primary goal is to develop an image captioning system combining CLIP and GPT to generate high-quality, fluent, and accurate captions. A secondary goal is to experiment with fine-tuning and attention mechanisms to further improve the system's performance.

5. Schedule

- Week 1: Literature review and dataset preparation.
- Week 2: Implement baseline model with CLIP and GPT.
- Week 3: Fine-tune and evaluate model on MS-COCO dataset.
- Week 4: Compare with SOTA models and finalize the report.

6. Comparison with SOTA

The model will be compared to existing state-of-the-art methods such as "Show, Attend, and Tell" (Xu et al., 2015) and transformer-based captioning models (**?**). Evaluation metrics will include BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and CIDEr (Vedantam et al., 2015) to measure caption accuracy and fluency.

References

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Radford, A., Kim, J. W., Hallacy, K., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 2021.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

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