CS 7643 - Deep Learning
Project Proposal

1. Project Information

1.1. Team Name

Caption Captains

1.2. Project Title

Image Captioning with Deep Learning

1.3. Members

- Christoper Dugan
- Mouhcine Aitounejjar
- Walker Stevens
- Xuan Xu

2. Project Summary

Image captioning, in the context of deep learning, is the process of generating natural language descriptions from images. It has applications across the fields of medicine, human-computer interaction, marketing and social media to name a few. It also improves the lives of the visually impaired by enabling them to interact with images.

This is an interesting topic because it is a relatively new domain that combines the fields of computer vision and natural language processing, which are both studied in this course. Further, it is a useful foundation for understanding how machines can interpret images.

Our team is excited to work on this project that has many applications and also offers an opportunity to learn about combining CNN and NLP architectures.

3. Approach

After reading through the current papers on captioning techniques, we will run state of the art implementations in order to understand the fundamentals of captioning architectures for the model we will construct. Once we vectorize our language data, we will decide on the architecture that goes into the CNN, encoder, and decoder. This will consist of deciding the orientation/number of layers as well as the method of normalization (batch vs. dropout). Once we validate our results through hyperparameter choice and architecture optimization, we will use the BLEU-4 metric to score our model on the MS COCO dataset, as well as an observable sanity check by detecting its zero-shot capability. As a bonus, we may also generate captions on our own personal images and possibly see if AI-generated art reproduces similar images to the caption description.

4. Resources / Related Work & Papers

After the release of BERT [2] and the adaption of transformers [13] into the realm of computer vision [9], we witnessed a boom of *Vision-and-Language Pre-training* (VLP) using multimodal models that process and align image and natural language text using transformer/BERT layer and cross-attention layer (e.g.: VilBERT [8], Pixel-BERT [5], VD-BERT [14], Unicoder-VL [4], VL-BERT [11], ImageBERT [10], Fashion-BERT [3] , VisualBERT [7], Uniter [1], LXMERT [12], etc.)

As of March 2023, mPLUG [6] achieved the highest BLEU-4 scores on image captioning on COCO Captions. mPLUG targeted the bottlenecks of computational efficiency and image-text information asymmetry by introducing cross-modal skip connections in the architecture. Thus enabling the multi-modal fusion to occur at disparate levels in the abstraction hierarchy across the modalities. Compared to connected-attention networks (concatenate visual and linguistic features as input) and co-attention networks (separate transformers for visual and text with fusion happening via cross-attention layers), mPLUG's proposed cross-modal skip-connected network was a top performer in downstream tasks (including image captioning) and runtime efficiency.

5. Datasets

- https://cocodataset.org/#captions-2015
- https://pytorch.org/vision/stable/generated/torchvision.datasets.CocoCaptions.html

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[3] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260, 2020. 1

[4] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 1

[5] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 1

[6] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 1

[7] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[9] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 1

[10] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 1

[11] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1

[12] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[14] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. *arXiv preprint arXiv:2004.13278*, 2020. 1