Project Title: Fine-grained image captioning

Team members: Ting-Chih Chen(tingchih@vt.edu), Xiao Guo(kevinguo2003@vt.edu), and Yi Lu (yilu38@vt.edu)

Introduction of the problem:
Generating accurate and thorough textual descriptions of images is a difficult task in computer vision. A particular focus is placed on catching small visual distinctions and features that people may easily overlook. Deep learning-based approaches have gained much attention in recent years because to their potential to learn complicated features and patterns in picture data automatically.

The objective of this research is to create an image captioning system based on deep learning that can produce captions for photos of a high caliber. Using large-scale image-captioning datasets, several cutting-edge deep-learning models will be trained and fine-tuned with a focus on capturing fine-grained details and subtleties in images. Our initiatives concentrate on removing additional information from photos, like "a little boy in a red shirt and yellow shoes is kicking a broken football in the rain" rather than "a little boy is kicking a ball."

Related work: While lots of effort has been put into improving the performance of traditional image captioning, including accuracy. Some representative works include combing object detection and language models.[6] One representative work is *"Knowing when to look: Adaptive attention via a visual sentinel for image captioning."* by *Lu et al.* [4] which uses Resnet for image encoder and LSTM for language model. This group also proposed a simple yet effective way of filling slots in a sentence template using the detected visual objects.[3] While they achieve a relatively high accuracy using traditional NLP metrics like Bleu, Rogue, and
, a rising research field now focuses on producing even more detailed captioning for the image than simply increasing the accuracy.

One opening work is to provide dense captioning, which tries to caption every object in the image in detail. Justin *et al.* came up with DenseCap[5] which utilized a CNN, a novel dense localization layer, and an RNN to generate label sequences. Yet, it only provides captioning for a single object instead of the whole image. To solve that, Zheng Jun *et al.* push the frontier to generate longer, richer, and more fine-grained sentences and paragraphs as image descriptions.[7] In our project, we strive to learn and improve on this existing work in this field to provide even more interesting captioning for the image.

Dataset:
Our experiments will be conducted on three datasets that are widely used for image captioning tasks. The first dataset we would like to use is MS COCO Captions. This dataset has 330, 000 image samples, each having 5 short sentences of natural language captioning [8]. The other dataset is Flickr30k, with 31, 000 images from Flickr, each with 5 reference human annotated captions[9]. In addition to MS COCO and Flickr30k, the third dataset we are planning to use is Visual Genome, with 101,174 images adopted from the MS COCO dataset with enhanced annotations that are aware of relations between objects in the image[10].

MS COCO was collected from a wide range of different sources, which include different image types, object types, and image purposes. At the same time, Flickr30k was collected from an online image hosting platform where users share their images of different contents and purposes. And since Visual Genome was created based on MS COCO, we can ensure the diversity of our data samples. Besides that, training and testing our models on different datasets can alleviate the potential biases in the annotation styles and selection of contents of each dataset, which could improve the generalizability of our models.

The plan of method development:
In this project, we mainly focus on how to improve the performance of image captioning. We want the captionings has many detailed information not only output the simple and ambiguous captionings. So, we will use LAVIS[1], ExpansionNet v2[2] and one traditional image captioning model to compare the results. Also, we will fine-tune ExpansionNet v2 to achieve the best performance. LAVIS is a robust library for language-vision intelligence. LAVIS can solve many vision-language tasks, and LAVIS includes many language-vision models, for example, ALBEF, BLIP, CLIP, and ALPRO. Although LAVIS has many pretrained checkpoints, we cannot prove LAVIS has a great performance in some datasets that it did not pretrain before. ExpansionNet v2 explores the possibility of performance bottlenecks in the input length. In this project, we mainly use ExpansionNet v2 to observe the results on different sequences length. We think we can fine-tune ExpansionNet v2 in different datasets and different sequences length to obtain the detailed captionings. After the comparison, we will try to buid up a new model based on the previous work. We hope the performance can approach their performance.

The summary of our contributions in this project:

- We apply LAVIS, ExpansionNet v2 and one traditional image captioning model on different datasets. Also, we will show some ablation study to discuss why the results are like this.
- We fine-tune ExpansionNet v2 to achieve the best performance.
- We will try to buid up a new model based on the previous experience. Then, we will try our best to approach their performance.

REFERENCE

[1] Li, Dongxu, et al. "Lavis: A library for language-vision intelligence." arXiv preprint arXiv:2209.09019 (2022).
[2] Hu, Jia Cheng, Roberto Cavicchioli, and Alessandro Capotondi. "ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning." arXiv preprint arXiv:2208.06551 (2022).
[3] Lu, Jiasen, et al. "Neural baby talk." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[4] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[5] Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[6] Hossain, MD Zakir, et al. "A comprehensive survey of deep learning for image captioning." ACM Computing Surveys (CsUR) 51.6 (2019): 1-36.
[7] Zha, Zheng-Jun, et al. "Context-aware visual policy network for fine-grained image captioning." IEEE transactions on pattern analysis and machine intelligence 44.2 (2019): 710-722.
[8] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.
[9] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47, 853-899.
[10] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123, 32-73.