Fine-grained image captioning

Ting-Chih Chen, Xiao Guo and Yi Lu
Department of Computer Science
Virginia Tech

Abstract

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 explore which parts of the model affect the captioning generation. In this paper, we compare the performance of LAVIS and ExpansionNet v2. In addition, we research the results of different visual embedding methods.

Keywords: Image captioning, visual embedding, and NLP
Github repo link: https://github.com/ting-chih/CS5814-final-project

1 Introduction

Image captioning, a common deep learning task to automatically generate text captions to image data, has become a popular topic in recent years, drawing public interest Herdade et al. [2019]. While the current state-of-the-art models can yield high performances in detecting objects in the images and generating text descriptions that make sense to the objects, new challenges have popped up for image-language models to detect the minuscule and easy-to-ignore components in the image to generate more detailed captioningTran et al. [2016]. In this research, we compared the performance of two state-of-the-art models published in the past few years to study which part of the deep-learning model can help enrich image captioning to include as many details as possible. By diving into the workings of these well-performing models tested to perform well in image captioning tasks, we aspire to provide insights to future developers and researchers to improve the current state-of-the-art models.

2 Related works

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 modelsHossain et al. [2019]. One representative work is "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." by Lu et alLu et al. [2017]. 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 objectsLu et al. [2018]. While they achieve a relatively high accuracy using traditional NLP metrics like Bleu, Rogue, and Meteor, 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 which utilized a CNN, a novel dense localization layer,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

and an RNN to generate label sequencesJohnson et al. [2016]. 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 descriptionsZha et al. [2019]. 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.

3 Datasets

In this paper, we mainly use two datasets, MS COCO Lin et al. [2015] and Flicker 30K Young et al. [2014]. We use the models trained on MS COCO and test on the Flicker 30K to explore the results. The MS COCO is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The annotations in this dataset include about 80 object categories, NLP captioning, panoptic, and dense pose. The training/validation/testing split is 118K/5K/41K in the MS COCO. In our testing time, we use Flicker 30K. The Flicker 30K contains over 30,000 images from Flicker and each image has five reference sentences provided by human annotators. We calculate the metrics with these five sentences and our generated image captioning.

4 Methodology

4.1 LAVIS

For the baseline comparison, we used models from the Lavis library published by Salesforce Li et al. [2022a]. Davis provides an open-source deep learning library that provides advanced, pertaining models for vision and language tasks. Among the numerous deep learning models Lavis provides, we chose BLIP, which was trained on the MS COCO dataset aiming for image captioning tasksLi et al. [2022b]Lin et al. [2015]. The architecture of BLIP can is in Figure 1.

The author of BLIP aimed to solve two pain points in vision-language pre-training models. First is that most models that existed were not able to adapt to both generation and understanding tasks. This happened because the encoder-based pre-training models are not performing well in transferring image generation tasks like image captioning. On the other hand, the encoder-decoder-based models are not performing in image-text retrieval tasks. BLIP introduced the multi-modal mixture of encoder-decoder(MED) architecture. For computational efficiency, BLIP uses a visual transformer (ViT) to flatten the image into lower-dimensional vectors of smaller patches of the image source and corresponding positional information embeddings Dosovitskiy et al. [2020]. Then, the embedding was passed to the MED architectures, which utilizes a unimodal encoder, an image-grounded text encoder, or an image-grounded text decoder based on the tasks. By combining the three different architectures, BLIP has an advanced performance on numerous image-language tasks. The second gap in research BLIP solved was that most of the state-of-art pre-training models existing was trained based on image-text pairs retrieved from the internet. Due to the labor cost of finely captioning images for large datasets, noises in image-text data were unavoidable. To solve this, BLIP introduced Captioning and Filtering(CapFilt), which includes a captioner and filter to improve the quality of web-scrapped data. The captioner is based on a pre-trained image-ground text decoder to generate captioning for the web image. The filter is based on an image-ground text encoder to detect the mismatch between the image and text pairs. By employing these two functionalities, BLIP improves the quality of the web-scrapped image-text dataset by generating new labels and removing noisy labels in the dataset at the same time. The architecture of the CapFilt can be found in Figure 2

4.2 ExpansionNet v2

ExpansionNet v2 Hu et al. [2022] is a expansion method to explore the possibility of performance bottlenecks in the input length in Deep Learning methods. The special thing in this model is that ExpansionNet v2 uses the block static expansion which distributes and processes the input over a heterogeneous and arbitrarily big collection of sequences characterized by a different length

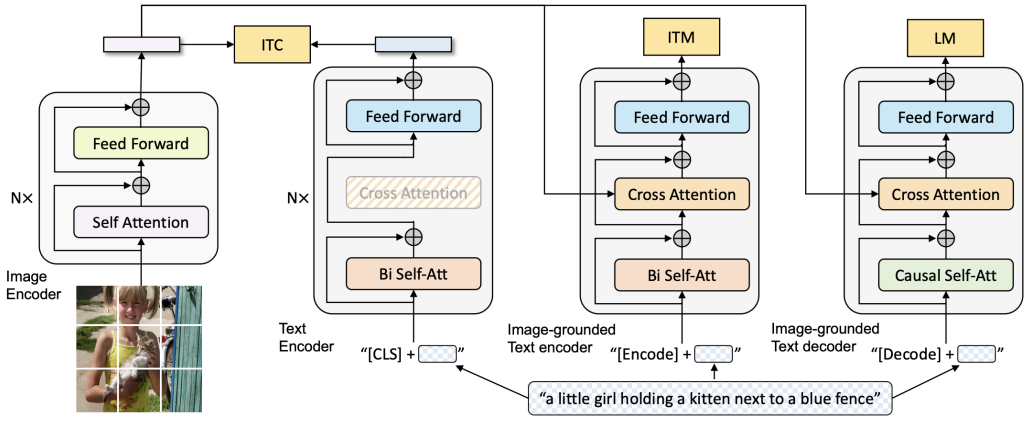

Figure 1: BLIP architecture.

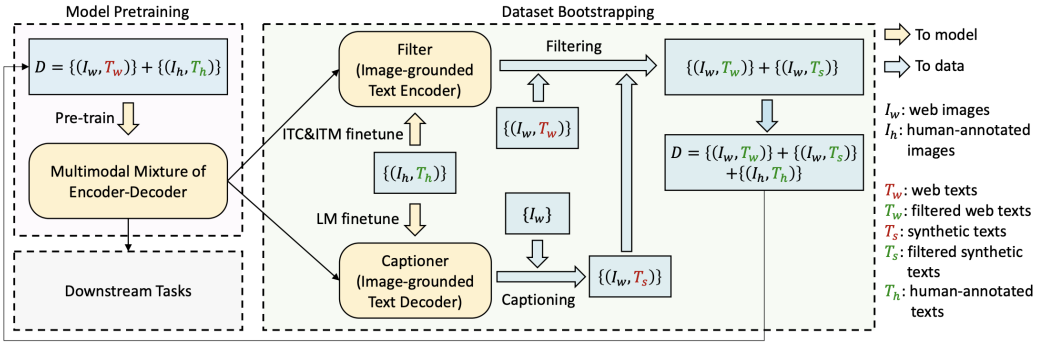

Figure 2: CapFilt architecture.

compared to the input one. This method is proven not only effective but also faster than the existing approaches of image captioning. The architecture is in Figure 3.

The ExpansionNet v2 architecture is to obtain visual features using Swin-Transformer Liu et al. [2021]. The benefit of Swin-Transformer is that Swin-Transformer builds hierarchical feature maps by merging image patches and has linear computation complexity because self-attention only focuses on the local windows. This embedding method is good at image classification and dense recognition tasks. After obtaining visual features, ExpansionNet v2 will feed the results into the encoder which is made of Static Expansion, FeedForward blocks, skip connection and pre-layer normalization. In addition, the decoder is made of Dynamic Expansion, Cross-Attention, FeedForward blocks, skip connection, and normalization applied on each component. ExpansionNet v2 mainly uses this architecture to generate image captioning. The contribution of this model is Block Static and Dynamic Expansion Layers. Block Static Layer is operated in both bidirectional processing and decoding stage. Dynamic Expansion Layer is designed to preserve the auto-regression property. In our work, we mainly use ExpansionNet v2 to explore different input sequence lengths and to observe the performance of the different visual embedding methods.

5 Results

Generated Captions

The following graphs contain the images we gathered from both the Flicker30K dataset and the internet. So none of the images are from our training set. We can observe from 4 that the two models we used provided pretty similar results that clearly describe the objects in the image. However,

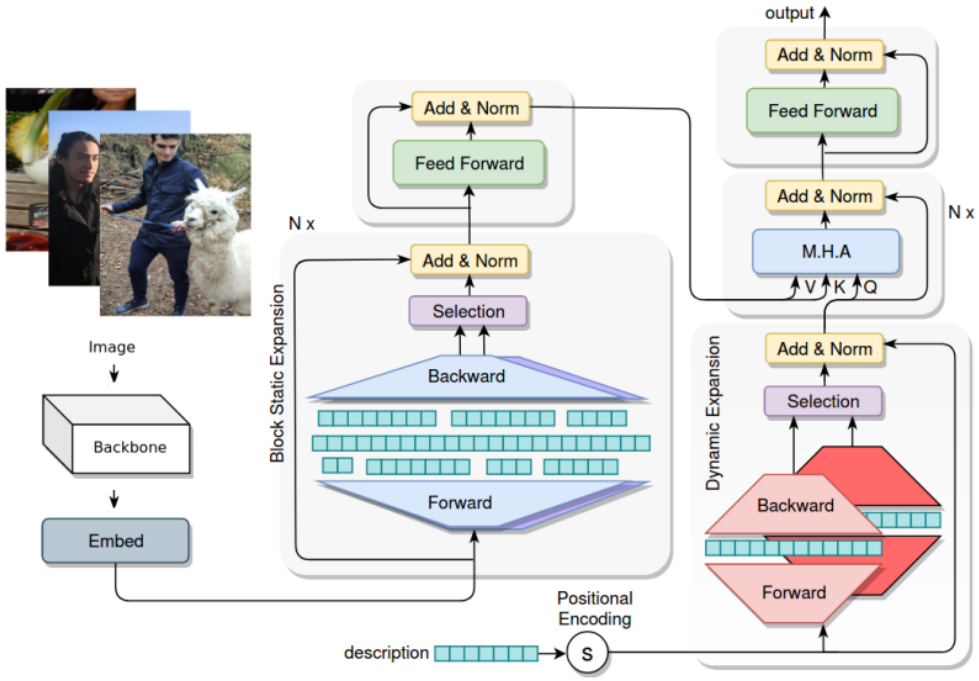

Figure 3: ExpansionNet v2 architecture.

Table 1: Comparison of two models over different evaluation metrics. We could see that the two models have close performance on metrics like Spice, Meteor, and Cider. However, the expansion net v2 model outperforms Lavis specifically on BLeU and Rouge.

Methodology	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Spice	Meteor	Rouge	Cider
ExpansionNet v2	0.6776	0.4848	0.3303	0.2236	0.1499	0.2192	0.4606	0.3834
Lavis	0.6386	0.4302	0.2922	0.1966	0.1565	0.2201	0.3887	0.4290

due to the nature that these two models are trained on the MS-COCO dataset, which is relatively simple and lacks details, we found that the captions fail to detect certain things. For example, the ExpansionNet_v2 has described a raccoon as "An animal with white and brown color." in 6. Also, while the caption is accurate and does not provide wrong information, it did not deliver a higher level of understanding of the relationship between the objects in the images. For example, instead of climbing a mountain, the ExpansionNet_v2 again describes the action of climbing to be "walking on the side of the mountain." in 5. Lastly, none of these two models could identify the emotions of the people inside the images.

Numerical Results

To quantitatively examine the performance of our models, we have adopted several major metrics used by the major language-related tasks, including BLeU, Spice, Meteor, Rouge, and Cider. We present our results in Table1. We could see that our image captioning performed reasonably on most metrics due to the nature of transferring the model trained on a relatively simple dataset to test on the Flicker30K dataset. While the two models provided similar results on most metrics, it's clear that ExpansionNet_v2 has outperformed Lavis on all the BLeU metrics and Rouge.

6 Conclusion

Limitation and Future Work

E: A sail boat is sitting in the ocean.
L: A sailboat in the middle of the ocean

E: A group of children sitting at a table.
L: a group of young children sitting around a table

E: A group of men standing next to an airplane.
L: a group of men standing next to an airplane

Figure 4: Generated captions for three random-picked Flicker 30K photos

E: A woman walking on the side of a mountain.
L: a woman climbing up the side of a mountain

E: A group of men walking down a city street.
L: a group of men in suits walking down a sidewalk

E: A crowd of people standing in front of a tower
L: A person holding a key to the eiffel tower

Figure 5: Generated captions for three random-picked internet photos

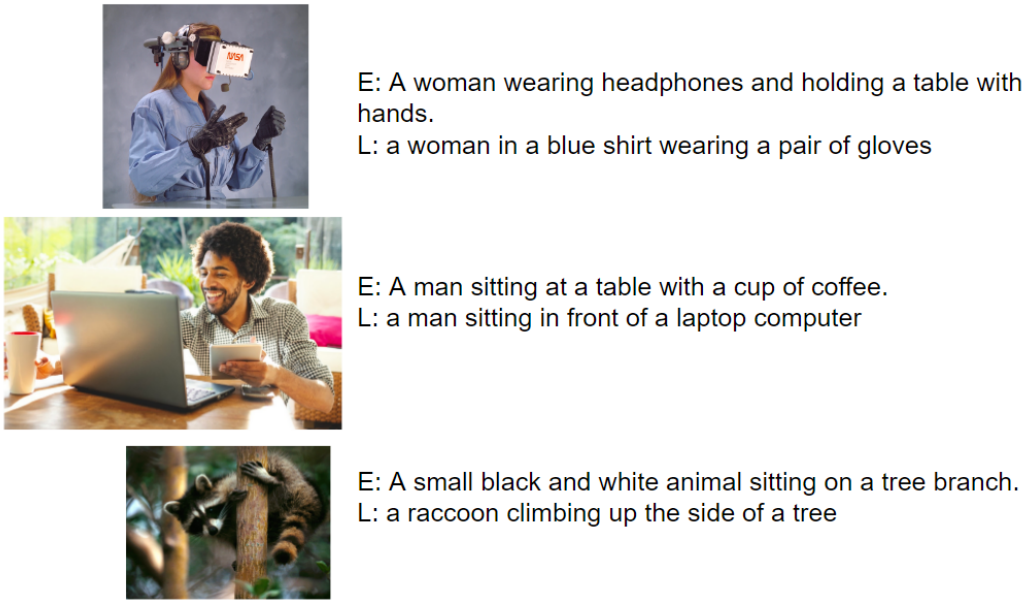

Figure 6: Generated captions for three high-diffuculty random-picked internet photos

For future work, we will address some of the limitations and challenges we faced when approaching our research problem. A first-and-foremost problem is how we could define the quality of the caption. As traditional metrics like BLeU only compare for similarity and fluency, there is no good way to really determine how detailed and accurate our caption can describe the attributes and relations of the objects in the images. We propose to develop a new metric that can measure the degree of fine-graininess of the captions by introducing creative yet consistent human-developed metrics to solve this problem.

Another direction of future work is to continue improving the architecture of the ExpansionNet$\check{2}$, which is responsible for expanding the initial captions with more attributes. We aim to make it more robust and flexible to handle different types of images and attributes by adjusting the internal network layers and hyperparameters. Finally, we intend to investigate other works that strive to provide more attributes of the images, such as emotions, actions, or relations, and explore how to integrate them with our approach. One exciting work we noted is the Segment Anything Model (SAM) proposed by Meta, which could finish object detection tasks in a surprisingly accurate and fast way. With the help of this model, we could build a better image-captioning model with more focus on language generating for more detailed, more accurate, and more natural paragraph-long image descriptions.

Summary

To summarize our work, we have presented a performance comparison for fine-grained image captioning based on the ExpansionNet_v2 and Lavis library in this paper. We have shown that the Swin transformer, as the underlying model used by ExpansionNet_v2, outperforms ViT, the core model supporting the Lavis library's functionalities. This proves that the Swin transformer is more suitable to serve as the backbone for extracting visual features from images due to its sliding window feature. We have also investigated BLIP, a caption-filtering method used by the Lavis library that can effectively denoise the generated captions by removing irrelevant or redundant words. Our experiments on two fine-grained image captioning datasets demonstrate the effectiveness of our approach in generating accurate and detailed captions compared to the baseline from both the Flicker30K dataset and human-generated captions.

References

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019.

Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich image captioning in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 49–56, 2016.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning, 2017.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk, 2018.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.

Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):710–722, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL `https://aclanthology.org/Q14-1006`.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. Lavis: A library for language-vision intelligence. *arXiv preprint arXiv:2209.09019*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. Expansionnet v2: Block static expansion in fast end to end training for image captioning, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.