CVIT Workshop 2023

Automated Image Captioning

Roja Sahoo[1]

[1]CVIT, Centre for Visual Information Technology, IIIT Hyderabad

May 29, 2023

Abstract

A key function of vision-language comprehension is picture captioning, in which the model predicts a textual summary of an input image. By using a straightforward mapping network and CLIP encoding as a prefix to the caption, we are able to produce the image captions. The most effective model for vision-language perception is the recently proposed CLIP model, which has rich semantic features that were trained using textual context. Our main argument is that by using a pre-trained language model (GPT2), we can comprehend a wide range of visual and textual input. Therefore, our method just needs a little period of training to develop a reliable captioning model. For large-scale and varied datasets, it effectively creates meaningful captions without the need for extra annotations or pre-training. Unexpectedly, our approach still performs well when only the mapping network is taught, leaving CLIP and the language model fixed. This enables a simpler architecture with fewer trainable parameters. By quantitatively comparing our model against state-of-the-art techniques on the tough Conceptual Captions and nocaps datasets, we show that our model outperforms them while being easier, quicker, and lighter.

Contents

1 Introduction

Automated image caption generation is a fascinating field at the intersection of computer vision and natural language processing (NLP) that aims to bridge the gap between visual content and textual descriptions. With the rapid advancement of deep learning models, particularly in the form of convolutional neural networks (CNNs) for image understanding and transformer-based models for language generation, we have witnessed significant progress in the development of automated systems capable of generating descriptive captions for images.

By integrating the strengths of models like CLIP and GPT-2, we have been able to create automated image captioning systems that effectively bridge the gap between visual content and textual descriptions. These systems hold promise for a wide range of applications, from enriching visual content understanding to improving accessibility and searchability in various domains.

CLIP (Contrastive Language-Image Pretraining) excels at learning image-text associations and encoding images and text into a shared latent space. GPT-2 (Generative Pre-trained Transformer 2) is a highly versatile language model capable of generating coherent and contextually relevant text.

We extract the visual prefix of an input image using the CLIP encoder and the mapping network. We start generating the caption conditioned on the visual prefix, and predict the next tokens one by one, guided by the language model output. For each token, the language model outputs probabilities for all vocabulary tokens, which are used to determine the next one by employing a greedy approach or beam search.

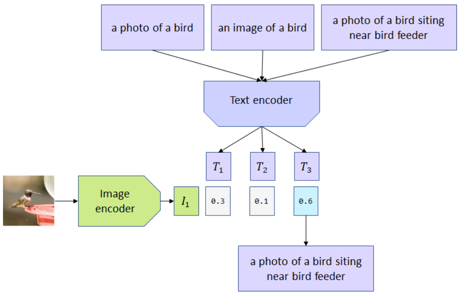

Figure 1: CLIP encoding

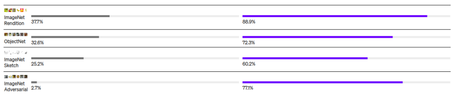

Figure 2: CLIP performs 4 times more efficient at zero-shot ImageNet accuracy when compared to other methods

2 Methodology

- Pretraining CLIP: CLIP is initially pretrained on a large dataset of images and their associated text descriptions. During pretraining, CLIP learns to encode images and texts into a shared latent space, where similar images and captions are closer together.

- Fine-tuning CLIP: The pretrained CLIP model is then fine-tuned on a specific image-caption dataset to further improve its performance on the captioning task. This fine-tuning involves training CLIP to predict the correct captions given the corresponding images.

- Feature Extraction: To generate captions for new images, the integrated system first uses CLIP's image encoder to extract visual features from the input image. These features capture the important visual information within the image.

- Context Generation: The extracted image features are then passed through the GPT-2 model, which is a powerful language model capable of generating coherent and contextually relevant text. The GPT-2 model takes the image features as input and generates a context for the caption generation.

- Caption Generation: Finally, the integrated system combines the generated context from GPT-2 with the image features and feeds them into a decoder module. The decoder generates the final caption by generating words sequentially based on the combined image and textual context. This process continues until an appropriate caption length or a predefined stopping criterion is reached.

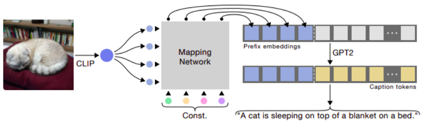

Figure 3: CLIP and GPT Mapping

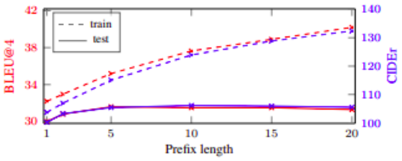

Figure 4: MLP mapping network with fine-tuning of the language model.

3 Results and Analysis

As can be seen, our generated captions are meaningful and depict the image successfully for both datasets Conceptual Captions and COCO. As can be seen, our Conceptual Captions model generalizes well to arbitrary unseen images as it was trained over a sizable and diverse set of images. Moreover, our model successfully identifies uncommon objects even when trained only over COCO.

We observed that increasing the size of the prefix length, up to a certain value, improves the performance of the model in an underlying task. We constrained it to 10 to abide by the memory space. The smaller learning rate = 0.00002 and batch size = 40 allows for optimal set of weights to be determined. AdamW allows for quicker training of the mapping.

Training usually takes a lot of time, this was resolved doing it only for the mapping and using pretrained models. High quality/detailed images took longer to process. Sometimes the extrapolation led to blank/simple images being described wrongly.

4 Conclusion

Overall, our CLIP-based method for captioning images is easy to use, doesn't need any extra annotations, and trains more quickly. We suggest a more straightforward approach, but as the dataset gets richer and more varied, it shows more promise. We see our method as a

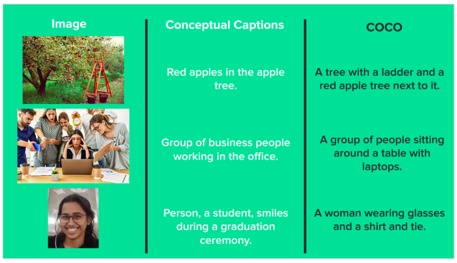

Image	Conceptual Captions	COCO
	Red apples in the apple tree.	A tree with a ladder and a red apple tree next to it.
	Group of business people working in the office.	A group of people sitting around a table with laptops.
	Person, a student, smiles during a graduation ceremony.	A woman wearing glasses and a shirt and tie.

Figure 5: Some example images and their captions on the Conceptual Captions and COCO dataset.

new paradigm for image captioning that focuses on utilising existing models and only trains a small mapping network. Instead of learning new semantic entities, this technique simply learns to adapt the pre-trained models' current semantic understanding to the format of the target dataset. We predict that in the near future, use of these potent pre-trained models will increase. Therefore, it is really interesting to learn how to use these components. It can also be modified to perform other challenging tasks, such as visual question answering or image to 3D translation.

5 Appendix

CLIP is a multimodal model that learns to associate images and their textual descriptions. It combines a vision encoder and a language encoder into a joint embedding space, where the embeddings of corresponding images and text are close together. The vision encoder is based on a convolutional neural network (CNN), typically using architectures like ResNet or Vision Transformer (ViT). The language encoder is based on a Transformer architecture, similar to GPT-2.

GPT-2 is a state-of-the-art language model that uses a Transformer architecture to generate coherent and contextually relevant text. It consists of a stack of Transformer encoder layers, which encode the input text and capture its semantic and syntactic information. GPT-2 can be fine-tuned on various downstream tasks, including image captioning.

The textual descriptions associated with the images are preprocessed by tokenizing the sentences into words or subword units, converting the tokens to numerical representations, and adding special tokens (e.g., [CLS] and [SEP]) to mark the beginning and end of the text.

generate_caption: This is the main function that generates the captions. It weighs the best probabilities of the recognized tokens and gives the most probable output.

generate_beam: Beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set. This function generates a list of possible captions from the best one.

6 Acknowledgements

I'd like to thank all the CVIT professors and research students who introduced me to the field of NLP and Computer Vision. With their guidance and input I have been able to apply my understanding to produce this project report.

References

[1] https://arxiv.org/pdf/2111.09734.pdf

[2] https://www.louisbouchard.ai/clipcap/