Enhanced Image Captioning via Dual Encoder

Guangjie Xu[1]
2024312767
Shenzhen International Graduate School
Tsinghua University
Lishui Road, 2279
xu-gj24@mails.tsinghua.edu.cn

Yangchi Gao[1]
2024312766
Shenzhen International Graduate School
Tsinghua University
Lishui Road, 2279
gaoyangc24@mails.tsinghua.edu.cn

1 Introduction

1.1 Background

In recent years, advancements in deep learning have significantly propelled the development of text-image multimodality. These models (e.g. SimCLR, MOCO, CLIP, ALBEF, BLIP) have become increasingly versatile and capable of tackling diverse tasks that bridge the gap between textual and visual data.

Contrastive learning leads to state of the art performance in the unsupervised training of deep image models(Khosla et al., 2020), which enhances the model's ability to align text and image representations effectively. For instance, models like CLIP have demonstrated exceptional performance in image classification, image description, and cross-modal retrieval tasks by learning from large-scale datasets of paired image-text data(Radford et al., 2021). The BLIP-2 model exemplifies an innovative approach to visual-language learning by leveraging a combination of pre-trained and frozen components. Specifically, it integrates a frozen CLIP image encoder, a frozen language model, and a lightweight Q-Former module. This architecture is designed to strike an optimal balance between efficiency and performance, enabling the model to excel across a variety of vision-language tasks while using significantly fewer parameters compared to its counterparts(Li et al., 2023).

1.2 Related Work

Text-image multimodality has become a rapidly evolving field, focusing on the integration of visual and textual information to enhance machine understanding and generation across diverse tasks. Li et al.(2021) proposed a visual and language representation learning framework called ALBEF (Align Before Fuse) to improve the joint representation of large-scale image and text data. ALBEF aligns image and text representations by introducing contrastive loss and then fuses them through cross-modal attention, thereby achieving more specific visual and language representation learning. CLIP(Contrastive Language-Image Pre-training) is a framework for learning visual models from natural language supervision, which is pre-trained by predicting whether an image and a text description match(Radford et al., 2021). Different from previous pre-training models only in either understanding-based tasks or generation-based tasks, BLIP(Bootstrapping Language-Image Pre-training) is a pre-training framework for vision-language understanding and generation that achieves state-of-the-art performance on multiple vision-language tasks by guiding the synthesis of diverse captions from noisy image-text pairs and removing noisy captions, different from previous pre-training models(Li et al., 2022). Li et al.(2023) proposed BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models.

Preprint. Under review.

1.3 Problem

Our project aims to propose an image captioning model based on BLIP-2 model with a dual-image encoder(CLIP encoder and SAM encoder) and mixed Semantic learning to simultaneously capture both the overall information and finer details within images.

2 Motivation

While significant progress has been made through deep learning and multimodal models, existing approaches often struggle to achieve a comprehensive understanding of images. Traditional models either emphasize high-level global features or focus on localized object details, but fail to capture both of these perspectives effectively.

However, in real world, meaningful captions rely not only on recognizing individual objects but also on understanding their relationships and the broader scene context. A model that only focuses on the overall scene might miss critical details, while one that fixates on objects could lose the overall meaning. Therefore, it is significant to balance global and local information.

By leveraging dual-image encoder to capture global and local features and introducing mixed semantic learning to integrate these levels of understanding, we aim to design an image captioning model that generates precise, coherent, and detailed captions.

3 Methodology

3.1 Model Structure

Our main idea is to use different encoders to capture different information based on BLIP2 Model. On the one hand, we tend to use a ViT image encoder pre-trained on CLIP to be responsible for extracting the overall information of images. On the other hand, we choose a different encoder to supplement the fine-grained information of medical images. The overview of our model structure is shown in fig1.

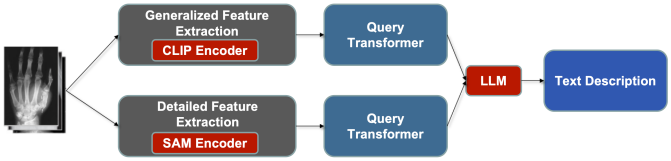

Figure 1: Workflow

First, based on CLIP and SAM, we train two Vision Transformer Encoders separately to encode image features, generating two sets of distinct image embedding vectors $\mathbf{v}_{CLIP}$ and $\mathbf{v}_{SAM}$:

$$\mathbf{v}_{CLIP} = CLIP(X),$$
$$\mathbf{v}_{SAM} = SAM(X),$$

where X denotes the image.

Second, we use two Q-formers respectively to gain the aligned features after processing the output of dual encoders through cross attention, which are denoted as follows:

$$\bar{\mathbf{v}}_{CLIP} = f_{CLIP}(\mathbf{q}_{CLIP}, \mathbf{v}_{CLIP}),$$
$$\bar{\mathbf{v}}_{SAM} = f_{SAM}(\mathbf{q}_{SAM}, \mathbf{v}_{SAM}).$$

$\mathbf{q}_{CLIP}$ and $\mathbf{q}_{SAM}$ represent the learnable query vectors in Q-formers. The overview of Q-former is shown in fig2.(Li et al, 2023)

Last, we use LLM model to generate text descriptions with outout of Q-former and text prompt embedded as input:

$$r = LLM(\bar{\mathbf{v}}_{CLIP}, \bar{\mathbf{v}}_{SAM}, \mathbf{v}_{text}).$$

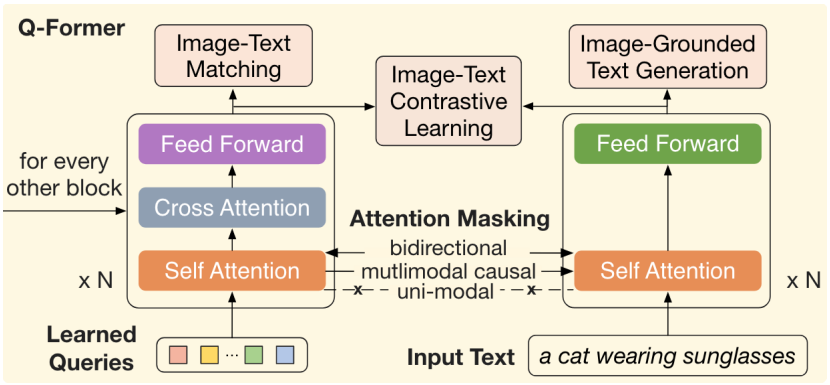

Figure 2: Q-former

3.2 Mixed Semantic Pre-training

We follow the workflow in BLIP2(Li et al, 2023), train two Q-formers seperatively with frozen image encoders. We jointly optimize three pre-training objectives that share the same input format and model parameters: Image-Text Contrastive Learning (ITC), Image-grounded Text Generation (ITG) and Image-Text Matching (ITM).

Each objective employs a different attention masking strategy between queries and text to control their interaction. ITC aims to reduce the distance between image and corresponding text features, ITG aims to generate text descriptions, and ITM aims to determin the relevance of images and text.

Based on our needs, we train Q-former of CLIP with general dataset COCO(Lin et al, 2014), and to gain more fine-grained details, we train Q-former of SAM with general and medical datasets COCO(Lin et al, 2014), ROCO(Obioma et al, 2018), and MedICaT(Subramanian et al, 2020).

3.3 Captioning with Frozen LLM

Based on the pre-training from the previous stage, the outputs of the dual encoders and Q-Formers are concatenated and fed into the LLM (OPT model). In this stage, the image encoders and LLM are frozen, while only the Q-Formers and linear projection layers are trained.

4 Experiments

We employed BLEU(Papineni et al, 2002) and METEOR(Banerjee et al, 2005)as the evaluation metrics. The results are shown in table1.

Table 1: Result

Models	Bleu1 (10^3)	Bleu2 (10^3)	Bleu3 (10^3)	METEOR (10^3)
BLIP2 (G)	7.4	3.0	1.2	26.1
BLIP2 (G+M)	53.2	20.6	7.5	28.3
SAM-BLIP2 (G+M)	83.4	29.4	10.0	35.5
Ours(G, G)	101.6	45.4	22.2	59.1
Ours(G+M, G)	107.2	45.8	22.3	53.0
Ours(G+M, G+M)	48.1	22.8	10.9	49.4
Ours(G, G+M)	**108.9**	**48.1**	**23.1**	**62.6**

The content in parentheses represents the training datasets, where G denotes the general dataset and M denotes the medical dataset.Through the result, it seems that the cooperation of the CLIP encoder and SAM encoder performs better than either of them alone. The CLIP encoder helps capture general semantic information and the SAM encoder helps capture detailed semantic information.

Their cooperation significantly improves the quality of the generated text. Additionally, results of different training strategies show that the SAM encoder is more proficient at capturing fine-grained image details compared to the CLIP encoder.

5 Conclusion

Our project improve the performance of BLIP2 by introducing dual image encoders, with CLIP capturing general information and SAM capturing fin-grained details.

In furture research, we aim to use a larger LLM and more professional datasets to improve the accuracy. Also, we are considering applications in other fields like restoring blurred photos.

References

[1] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan, "Supervised contrastive learning." *Advances in neural information processing systems* 33 (2020): 18661-18673.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning transferable visual models from natural language supervision." in *International conference on machine learning*. PMLR, 2021.

[3] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, Steven Chu Hong Hoi, "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in neural information processing systems* 34 (2021): 9694-9705.

[4] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." in *International conference on machine learning*. PMLR, 2022.

[5] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." in *International conference on machine learning*. PMLR, 2023.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll ar, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014, pp. 740–755.

[7] P. Obioma, S. Koitka, J. R uckert, F. Nensa, and Christoph M F. Friedrich, "Radiology objects in context (roco): a multimodal image dataset," in *Intravascular Imaging and Computer Assisted Stenting and LargeScale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVIISTENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI*, 2018.

[8] S. Subramanian, L. L. Wang, S. Mehta, B. Bogin, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, and H. Hajishirzi, "Medicat: A dataset of medical images, captions, and textual references," *arXiv preprint arXiv*:2010.06000, 2020.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[10] S. Banerjee and A. Lavie, "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*, 2005.