Transferring Pre-Trained Large Language-Image Model for Medical Image Captioning

Notebook for the Baidu Intelligent Health Unit and Peng Cheng Laboratory Joint Team at CLEF 2023

Wenshuo Zhou^{1,†}, Zhiyu Ye^{2,†}, Yehui Yang¹, Siqi Wang¹, Haifeng Huang¹, Rongjie Wang^{2,*} and Dalu Yang^{1,*}

¹Baidu Intelligent Health Unit, Beijing 100085, China ²Peng Cheng Laboratory, Shenzhen 518055, China

Abstract

This paper introduces the work conducted by the team "closeAI2023" in the ImageCLEFmedical Caption 2023 Image Caption sub-task. Medical image captioning poses unique difficulties due to the specialized nature of the medical domain. It requires the generation of accurate and coherent captions that not only describe the visual content but also capture the essential medical information conveyed by the images. To leverage the abilities of pre-trained Large-Image Models, we utilise the state-of-the-art BLIP-2 with a giant vision transformer (ViT-g) and Open Pre-trained Transformer Language Models ($OPT_{2.7B}$) as the foundation of our caption prediction sub-task. To adapt the model to the medical domain, we employed a two-stage fine-tuning process. The pre-trained $OPT_{2.7B}$ was fixed during the whole training process. A step-wise fine-tuning of the ViT-g and the Q-Former modules was conducted to better align with the characteristics of medical data. Our team's approach yielded promising results, as we achieved a second-place ranking among all participating teams with a BERTScore of 0.6281. Additionally, our model performed well across various evaluation metrics: ROUGE of 0.2401 (4th), BLEURT of 0.3209 (1st), BLEU of 0.1846 (3rd), METEOR of 0.0873 (3rd), CIDEr of 0.2377 (1st), and CLIPScore of 0.8074 (3rd).

Keywords

ImageCLEF, Medical Image Caption, Caption Prediction, BLIP-2, Domain Adaptation

1. Introduction

ImageCLEF 2023 is a Multimodal Challenge organized as part of CLEF Initiative Labs[1]. Since 2003, ImageCLEF has been dedicated to investigating solutions for challenges involving multimodal data across diverse domains. Medical image captioning, recognized as a significant and demanding task within the medical field, has been featured in ImageCLEF for the seventh consecutive year. ImageCLEFmedical Caption 2023[2] encompasses two subtasks, including Concept Detection and Caption Prediction Task. Our team mainly focused on the latter task.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece *Corresponding author.

[†]These authors contributed equally.

^{© 0 2023} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CEUR Workshop Proceedings (CEUR-WS.org)

 Table 1

 Number of images per modality in ImageCLEFmedical Caption 2023 training dataset.

Overall		Modalities									
	X-Ray	Plain X-Ray	MRI	Ultrasonography	Angiogram	PET	PET/CT	Radionuclide Imaging	Radiographic Imaging	Other	
60,918	20,955	16,838	9,482	8,355	3,954	472	136	37	21	668	

The Caption Prediction Task necessitates the generation of accurate and coherent descriptions based on medical images. To accomplish this, the model must effectively recognize and extract the semantic information embedded within the medical images, capture the inherent relationships between these semantics, and proficiently express them using medical terms. BERTScore[3] and ROUGE[4] are the primary and secondary evaluation metrics for this task.

The dataset used for ImageCLEFmedical Caption 2023 is an updated and extended version of the Radiology Objects in COntext (ROCO) dataset[5]. ROCO dataset contains over 80,000 radiology images with various modalities including ultrasound, X-Ray, Computer Tomography (CT), Magnetic Resonance Imaging (MRI) and so on. The statistics of images modalities in the training set is shown in Table 1. All images in ROCO have corresponding caption, keywords, Unified Medical Language Systems (UMLS) Concept Unique Identifiers (CUIs) and Semantic Type. For caption prediction task, the ground truth captions were pre-processed by removing all the links contained in original captions.

We employed BLIP-2[6], a vision-language pre-training method that bootstraps from frozen pre-trained unimodal models, for the caption prediction task. BLIP-2 is a recently proposed vision-language pre-training method by Li et al. building upon their previous work of BLIP[7] and it has demonstrated superior performance compared to various other vision-language pre-training methods, including Flamingo[8], across a range of vision-language tasks such as visual question answering, image captioning, and image-text retrieval.

In this paper, our method is specifically introduced in Section 2, the experiments and results are demonstrated in Section 3 and a brief summary is given in Section 4.

2. Method

The pipeline of our method is shown in Figure 1. Our method adopted the fine-tuned BLIP-2 ViT-g $OPT_{2.7B}$ model in [6] and a two-stage fine-tuning, *i.e.* concept-based fine-tuning and overall fine-tuning, was performed to the model on the competition dataset.

2.1. Architecture

The framework of BLIP-2 consists of three main components: an image encoder, a lightweight Querying Transformer (Q-Former), and a large language model (LLM). The pre-training process of BLIP-2 comprises two stages: vision-language representation learning from a frozen image encoder stage and vision-to-language generative learning from a frozen LLM stage. In deep learning, the term "frozen" refers to a state where specific layers or parameters of a neural network are set to be untrainable or unmodifiable. This implies that, during the training process, the weights associated with these frozen layers or parameters remain constant and do not get



Figure 1: The training pipeline of our method. The image encoder and Q-Former are in yellow boxes, indicating that their parameters were updated during training, while the LLM in the green box had its parameters frozen. In concept-based fine-tuning stage, both concept loss and caption loss were utilized, but only caption loss was employed in overall fine-tuning stage.

updated. During the two pre-training stages of BLIP-2, image models and language models were frozen, preserving their initial image understanding and text generation capabilities. In contrast, Q-Former was trained exclusively in the pre-training to extract visual representations that effectively corresponded to textual information and provided this information to the LLM. When applying BLIP-2 models to downstream tasks such as image captioning, the LLM remained frozen during the fine-tuning process, while the parameters of the image encoder and Q-Former were updated[6].

For our specific task, we adopted the BLIP-2 model for image captioning in [6] and chose the ViT-g/14 from EVA-CLIP[9] as the image encoder. Regarding the LLM selection, while the encoder-decoder-based FlanT5XL[8] exhibited superior performance in the zero-shot image captioning task compared to decoder-based OPT models, the $OPT_{2.7B}$ [10] demonstrated a slightly stronger ability to generate normal captions[6]. As a result, we adopted the $OPT_{2.7B}$ as the LLM.

2.2. Training Strategy

In the pre-training and fine-tuning phase of ViT-g and BLIP-2 ViT-g $OPT_{2.7B}$, standard natural image datasets such as ImageNet[11] and COCO[12] were employed. However, our task is based on medical images, which exhibit a substantial domain shift from natural images, thus we performed a two-stage fine-tuning process on the competition dataset with the LLM $OPT_{2.7B}$ frozen and simultaneously the Q-Former together with the image encoder ViT-g were updated. The language modeling loss[7] was utilized during the fine-tuning, which acts on image-grounded text to optimize the model's ability to generate coherent captions according to visual information.

Stage 1 - Concept-based Fine-tuning. In this stage, we used the concepts and captions of

images to jointly optimize the model. We designated the loss between the output of Q-Former and image concepts as the concept loss, while the loss between the final output of $OPT_{2.7B}$ and the image caption was termed the caption loss. Both of these losses were essentially language modeling losses, albeit with distinct optimization objectives.

- Concept loss aims to encourage Q-Former to generate expressions that align closely with professional medical terminology, leveraging the representative features extracted from medical images. Since each image corresponds to specific concepts represented by Concept Unique Identifiers (CUIs), we initially map these CUIs to English and then organize them into language descriptions using the sentence structure: "The image shows [concept 1], [concept 2], ..., [concept n]." These descriptions contain valuable and accurate information. Our intention is to make the output of Q-Former as close as possible to these descriptions.
- Caption loss calculates the difference between the final outputs of OPT_{2.7B} and the ground truth image caption, ensuring that Q-Former can generate informative and professional prompts while also allowing OPT_{2.7B} to generate accurate captions.

Stage 2 - Overall Fine-tuning. In this stage, the model was exclusively trained by minimizing the caption loss, prioritizing overall optimization during fine-tuning.

3. Experiments

3.1. Implementation Details

Our framework was developed using PaddlePaddle¹ version 2.4.2 and trained on 8 Ascend 910 NPUs. The adapter plug-in PaddleCustomDevice² was utilized in order to be compatible with the Ascend NPU. The size of image input were resized to 364×364 and the batch size was 16 in both fine-tuning stage. The model was fine-tuned for 50 epochs and 20 epochs in the first and second stage respectively. We use the AdamW optimizer with the weight decay of 0.05. The initial learning rate was set to 10^{-5} with a warm-up of 1000 steps to gradually adjust the learning rate.

In addition, based on the statistics depicted in Figure 2, the maximum output length of the model was set to 128. It is evident from the data that the majority of sentence lengths fall below this threshold. Our aim is to ensure the conciseness and adherence of the generated captions to the sentence length distribution in the dataset by setting an appropriate maximum output length.

3.2. Post-processing

Generative language models frequently face the problem of degradation, specifically, the quality of the generated text gradually declines as its length increases. This degradation often leads

¹https://github.com/PaddlePaddle/Paddle

²https://github.com/PaddlePaddle/PaddleCustomDevice



Figure 2: The frequency histogram regarding the sentence length of captions in the ImageCLEFmedical Caption 2023 dataset. The horizontal axis represents the length of sentences.

Table 2

Comparisons of some predicted captions before and after the post-processing.

Data ID	Before Deduplication	After Deduplication
000034	transthoracic echocardiography view of a small aneurysm in the left atrial sep- tum ao aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa aneurysm aoa	Transthoracic echocardiography view of a small aneurysm in the left atrial septum.
000047	anteroposterior radiograph of the left ankle at the end of the first year of treatment showing the thickening of the tibia and fibula and the growth of the distal end of the tibia into the distal end of the fibula note the thickening of the distal end of the tibia into the distal end of the fibula and the growth of the distal end of the tibia into the distal end of the fibula at the end of the first year of treatment	Anteroposterior radiograph of the left an- kle at the end of the first year of treat- ment showing the thickening of the tibia and fibula and the growth of the distal end of the tibia into the distal end of the fibula.
000080	ultrasonographic view of the patellar tendon of the patellofemoral joint of a 4-year-old domestic dog the patella patellar tendon lv patellar ligament lp patellar tendon ap patella ap patellar ligament ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella ap patella	Ultrasonographic view of the patellar ten- don of the patellofemoral joint of a 4-year- old domestic dog

to the emergence of repetitive patterns at various levels, including characters, phrases, and sentences.

To address the problem of duplicate vocabularies in our model's output, we implemented a post-processing. Specifically, we provide a predicted caption which contains repeated words as a prompt for ChatGPT ³, requesting it to generate a Python code snippet capable of removing repetitive content. This generated code was then employed to perform text deduplication on all predicted captions. Following the post-processing step, the BERTScore of our results on validation data increased from 0.608 to 0.628. Table 2 demonstrates a comparison of some predicted captions before and after deduplication.

³https://openai.com/blog/chatgpt

Table 3

Ablation studies on validation data of caption prediction task. BLIP-2 ViT-g Q-Former in the first row indicates that this model is only composed of ViT-g and Q-Former, without LLM. * represents the results of the model have been post-processed.

Model	Image Size	Trainable Params	Total Params	BERTScore	ROUGE-1
BLIP-2 ViT-g Q-Former	224	1.1B	1.1B	0.541	0.160
BLIP-2 ViT-g $OPT_{2.7B}$	224	1.1B	3.8B	0.593	0.249
BLIP-2 ViT-g OPT _{2.7B}	364	1.1B	3.8B	0.608	0.255
BLIP-2 ViT-g $OPT_{2.7B}^{*}$	364	1.1B	3.8B	0.628	0.253



Ground Truth: Chest X-ray showing enlarged cardiac silhouette with cardiothoracic ratio of 70%, and mild pulmonary congestion.

Prediction: chest x-ray showing bilateral infiltrates



Ground Truth: A giant retroperitoneal tumor.

Prediction: computed tomography ct scan of the chest showing a large lesion in the thoracic aorta

Figure 3: Two examples of predicted results and ground truths in the validation set of caption prediction task.

3.3. Results

3.3.1. Ablation Studies on Validation Set

Table 3 displays the results of ablation studies during our model selection process. We tested different model compositions, input image sizes, and the effect of post-processing via these studies. The result of BLIP-2 ViT-g $OPT_{2.7B}$ after post-processing (the last row of Table 3) was ultimately chosen and submitted.

Figure 3 provides two examples of the validation data to visually demonstrate the performance of our chosen model. Both examples showcase the input medical images alongside their corresponding predicted captions provided by our model. In the two presented examples, our prediction results successfully recognized the modality of the medical images, but fell short in accurately detecting anomalies. In the first example, the description of "bilateral infiltrates" relatively matched the presence of "pulmonary congestion"; however, it failed to identify cardiac enlargement. The second example involved an incorrect diagnosis, misclassifying an abdominal CT as a chest CT. Although a sizable lesion was identified, its localization was inaccurate.

Rank	Team Name	Run ID	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
1	CSIRO	4	0.642519	0.244618	0.313707	0.161486	0.079775	0.202512	0.814717
2	closeAI2023	7	0.628106	0.240061	0.320915	0.184624	0.087254	0.237704	0.807454
3	AUEB-NLP-Group	2	0.617034	0.213014	0.295011	0.169212	0.071982	0.146601	0.803888
4	PCLmed	5	0.615190	0.252756	0.316561	0.217150	0.092063	0.231535	0.802123
5	VCMI	5	0.614736	0.217545	0.308386	0.165322	0.073449	0.172042	0.808184

 Table 4

 The results of the top five teams for the caption prediction task in ImageCLEF 2023[2].

3.3.2. Results on Test Set

The ranking for the caption prediction task is determined based on the BERTScore. A complete list of all runs for the caption prediction are now available in the results folder[2] and on the official website⁴. Table 4 displays the best runs' results of the top five teams. It shows that our team "closeAI2023" achieved a second-place ranking with a BERTScore of 0.6281, which is only 0.0144 lower than that of the first-ranked team. Among the seven listed metrics, we have surpassed the first-ranked team in four of them, *i.e.* BLEURT, BLEU, METEOR and CIDEr. Furthermore, our BLEURT and CIDEr metrics achieved top positions with scores of 0.3209 and 0.2377, respectively. These results demonstrate the consistent and strong performance of our method across various evaluation criteria in the competition. However, despite the commendable results attained by our model, Figure 3 suggests that there still remains room for improvement.

4. Summary

This paper introduces the work of team "closeAI2023" in Caption Prediction Task of Image-CLEFmedical Caption 2023. The model we used was obtained through a two-stages fine-tuning based on BLIP-2 ViT-g $OPT_{2.7B}$. To eliminate the impact of duplicate statements, we also performed post-processing on the outputs of the model. Our team ultimately achieved second place in this task, with a BERTScore of 0.6281. This points out the effectiveness of our approach in generating high-quality captions for medical images. Codes and models will be open-sourced at OpenMedIA⁵[13].

Acknowledgments

The computing resources of Pengcheng Laboratory Cloudbrain II are used in this research. We acknowledge the support provided by OpenI Community (https://git.openi.org.cn).

⁴https://www.imageclef.org/2023/medical/caption
⁵https://openi.pcl.ac.cn/OpenMedIA

References

- [1] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.
- [2] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.
- [3] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
- [4] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
- [5] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, 2018, pp. 180–189.
- [6] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023).
- [7] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.
- [8] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
- [9] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, Eva: Exploring the limits of masked visual representation learning at scale, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19358–19369.
- [10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, arXiv preprint arXiv:2205.01068 (2022).
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

- [12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. 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, Springer, 2014, pp. 740–755.
- [13] J.-X. Zhuang, X. Huang, Y. Yang, J. Chen, Y. Yu, W. Gao, G. Li, J. Chen, T. Zhang, Openmedia: Open-source medical image analysis toolbox and benchmark under heterogeneous ai computing platforms, in: Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, Proceedings, Part I, Springer, 2022, pp. 356–367.