

ImageSem Group at ImageCLEFmedical Caption 2022 task: Generating Medical Image Descriptions based on Vision-Language Pre-training

Xuwen Wang¹, Jiao Li¹

¹*Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100020, China*

Abstract

This paper presents the work of ImageSem group in the ImageCLEFmedical Caption 2022 task. In the caption prediction subtask, we employed the bootstrapping language-image pre-training (BLIP) framework for generating medical image descriptions. We submitted 2 runs using exclusively the official training data and achieved the BLEU score of 0.2211, which ranked 9th among the participating teams. Despite the lower BLEU score, we achieved a relatively balanced performance across all other metrics, such as the ROUGE of 0.1847, the METEOR of 0.0675, the CIDEr of 0.2513, the SPICE of 0.0393 and the BERTScore of 0.6059, showing the potential of better verbal fluency via fine-tuning the medical caption generation task grounded by vision-language pre-training (VLP).

Keywords

Caption prediction, vision-language pre-training, bootstrapping

1. Introduction

As a classic track of ImageCLEF benchmark[1], the ImageCLEFmedical Caption task [2] consists of two subtasks, namely Concept Detection and Caption Prediction. On behalf of the Institute of Medical Information and Library, Chinese Academy of Medical Sciences, our Image Semantics group (MAI_ImageSem) participated in both of the two subtasks in the last 4 years[3], [4]. In this year, we focus on the Caption Prediction subtask, which asks participants to generate coherent captions for the entirety of an image, and requires higher accuracy and semantic interpretability of expression. For predicting fluent medical image captions, we employed BLIP[5], a recently released Bootstrapping Language-Image Pre-training framework that transfers flexibly to both vision-language understanding and generation tasks. This paper is organized as follows. Section 2 describes the dataset of the ImageCLEFmedical Caption 2022 task. Section 3 presents our methods for caption prediction. Section 4 lists our submitted runs. Section 5 makes a brief summarization.

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ li.jiao@imicams.ac.cn (J. Li)

🌐 <https://www.imicams.ac.cn> (J. Li)

🆔 0000-0003-3022-6513 (X. Wang); 0000-0001-6391-8343 (J. Li)



© 2022 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)

2. Dataset

The ImageCLEFmedical Caption 2022 track released an extended version of the ImageCLEF 2020 dataset, including 83,275 radiology images as training set, 7,645 radiology images as validation set, and 7,601 radiology images as test set. Each image is associated with UMLS (the Unified Medical Language System) [6] concepts and image captions originated from the PMC (PubMed Central) biomedical articles. For the caption prediction task, each caption is pre-processed, such as removing numbers and punctuation, lower-case, lemmatization, etc.

Table 1 shows the statistic of caption lengths (i.e. the number of words in a caption) in the official training set and validation set. It can be seen that the caption length of different images vary a lot, e.g., in training set, the caption length ranges from the minimum of 1 word to the max of 410, and in validation set, it ranges from 1 to 297. The mean caption length in training set is 16, while in validation set is 17. In addition to the influence of preprocessing operations, the various image context from original PMC articles lead to the significant differences in caption lengths, which is a challenge for accurate caption prediction. Clinically, captions containing only 1 word may provide poor semantic information, such as “angiography”, “radiograph”, “xray”, etc.

Table 1

Statistic of caption lengths in the training set and validation set of ImageCLEFmedical Caption 2022 task.

Dataset	Image Caption Pairs	Max Length	Min Length	Mean Length
Training	83,275	410	1	16
Validation	7,645	297	1	17

3. Methods

This section introduces our methods in the Caption Prediction subtask of ImageCLEFmedical Caption 2022 task. Figure 3 shows our workflow, which is described in detail in section 3.2.

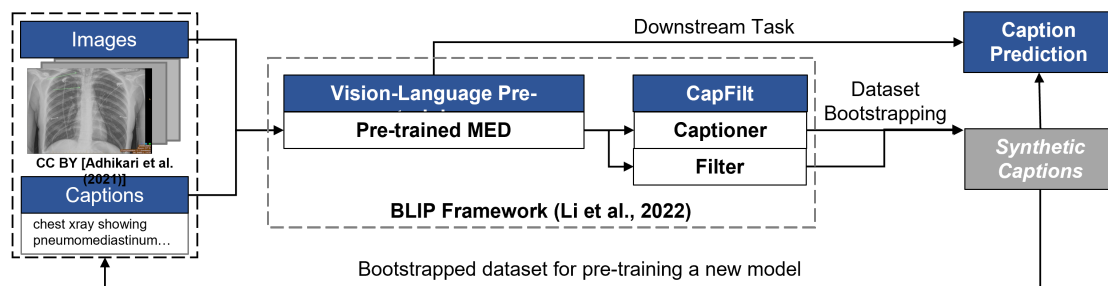


Figure 1: Workflow of ImageSem Group in the ImageCLEFmedical Caption 2022 task

3.1. Motivation

In the ImageCLEFmed Caption 2021 track, we employed two methods for caption prediction subtask. One was a pattern-based caption generation strategy, which combined the UMLS concepts identified from medical images with a predefined sentence pattern. The experimental results showed that the quality of generated sentences depended heavily on the accuracy of concept detection. The other method was an image-matching based model proposed by Zheng[7], which utilized two convolutional neural networks[8] to learn visual and textual representations simultaneously. It was based on an unsupervised assumption that every image or text group can be viewed as one class, so each category is equivalent to 1+m samples (i.e. 1 image vs m descriptions). The experimental results showed that the weakly matched medical images and captions may produce noisy or irrelevant descriptions.

The high-quality datasets of medical image-text pairs released by ImageCLEF provide opportunities for researchers to validate different ideas and methods. In the ImageCLEFmedical Caption 2022 track, we attempt to use a recently released vision-language pre-training (VLP) framework [5] to generate coherent and expressive descriptions for medical images.

3.2. Vision-Language Pre-training-based Caption Prediction

For generating fluent and reasonable medical image captions, we employed the BLIP (Bootstrapping Language-Image Pre-training) model proposed by Li et al. [5], which is a unified pre-training framework for vision-language understanding and generation downstream tasks. One contribution of BLIP is introducing multimodal Mixture of Encoder-Decoder (MED), a multi-task model which can operate either as a unimodal encoder, or an image-grounded text encoder, or an image-grounded text decoder. The model is jointly pre-trained with three vision-language objectives, including image-text contrastive learning, image-text matching, and image conditioned language modeling. Another highlight of BLIP is CapFilt (Captioning and Filtering) [5], a dataset bootstrapping method for learning from noisy image-text pairs. A pre-trained MED is fine-tuned into two modules, including a Captioner (based on the Image-grounded text decoder) to produce synthetic captions given images, and a Filter (based on the Image-grounded text encoder) to remove noisy captions. We also use this bootstrapping mechanism to expand our training set.

3.2.1. Data Preparation

We processed the official dataset according to the data format requirements of BLIP [5]. First, the file paths were filtered and image-caption samples were retained with complete image path. Second, the images from training set were resized to 224*224 pixels for pre-training, and images from validation set as well as test set were resized to 384*384 pixels for caption prediction. Third, to meet the requirement of BLIP's PyTorch framework that figure ID should be an integer without characters, we randomly converted original figure IDs to INT16 data type as the unique identification of images.

3.2.2. Experiments

In our work, we employed MED as the backbone framework. We tried two pre-training strategies. The first one performed vision-language pre-training from scratch based on the official training set (83,275 image-caption pairs) of ImageCLEFmedical Caption 2022, and then fine-tuned the caption prediction task based on the official validation set (7,645 image-caption pairs). The second one utilized the original BLIP model pre-trained on general datasets (with 14M images in total) [5] as initial parameters, and performed secondary pre-training based on the official training set. Then we also fine-tune the caption prediction task on the basis of validation set.

Due to limited computing resources, it took us a long time to complete all the pre-training processes. A preliminary analysis shows that the model pre-trained from scratch performed poorly on the downstream tasks, since it was hard to learn sufficient image-language association in the pre-training process, which subject to the limited scale of training data, as well as the diverse types of medical images and uneven captions. On the other hand, the secondary pre-training model with the BLIP initial parameters showed better performance on the downstream tasks, so we submitted the caption prediction results based on the secondary pre-training model to the subtask. We also applied the data bootstrapping process to predict synthetic captions for expanding training dataset. However, since the initial dataset contains a total of 83,275 image-text pairs, a limited number of high-quality captions were obtained by bootstrapping, which yield marginal improvement in the next round of caption prediction.

The experimental framework was implemented in PyTorch[9] and pre-trained on 8 NVIDIA V100 GPUs. We used ViT-B[10] as the image transformer, and the text transformer initialized from BERTbase[11]. Our model was pre-trained for 40 epochs using a batch size of 24. We used AdamW[12] optimizer with a weight decay of 0.05. The learning rate was warmed-up to $3e-5$ and decayed linearly with a rate of 0.85. With this experimental setting, we fine-tuned caption prediction task on the validation set and achieved a best BLEU score of 0.2344.

4. Submitted runs

Table 2 shows our submissions on the caption prediction subtask. Our best run is ID_182105, which use the pre-trained BLIP as initial parameter and perform secondary pre-training on the official training set. The hyper parameters of this submission were consistent with experimental settings in section 3.2.2, and achieved a BLEU score of 0.2211, ranked 9th in the leaderboard. We also achieved a relatively balanced performance across all other metrics, such as the ROUGE of 0.1847, the METEOR of 0.0675, the CIDEr of 0.2513, the SPICE of 0.0393 and the BERTScore of 0.6059, showing the potential of better verbal fluency. The submission ID_182348 refers to the secondary pre-training method with only 20 training epochs using a batch size of 24. It has achieved a BLEU score of 0.2179, the ROUGE of 0.1791, the METEOR of 0.0661, the CIDEr of 0.2304, the SPICE of 0.0371 and the BERTScore of 0.6013.

Figure 2 shows a few examples [13, 14, 15, 16] of our best run model on the validation set. It can be observed that, compared with Ground Truth, the predicted captions prefer to describe medical images from the global perspective. However, due to a lack of context or patient-related information, some descriptions are semantically incomplete or unclear, e.g. our model describes figure valid_084402 [16] as “computed tomography ct scan of the head and neck with contrast

show a soft tissue”, while the GT clearly state the nature of the lesion is “emphysema in the nasopharynx”.

Table 2

Submissions of MAI_ImageSem group in the caption prediction subtask

Submission ID	BLEU	ROUGE	METEOR	CIDEr	SPICE	BERTScore	Rank
182105	0.221136	0.184723	0.067541	0.251316	0.039311	0.605873	9
182348	0.217935	0.179142	0.066132	0.230447	0.037081	0.601299	–




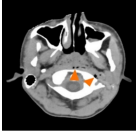
Figure ID	Image	Ground Truth	Prediction
ImageCLEFmedCaption_2022_valid_084382.jpg <small>CC BY [Adhikari et al. (2021)]</small>		chest xray showing pneumomediastinum	preoperative chest xray show pneumomediastinum and sub
ImageCLEFmedCaption_2022_valid_084427.jpg <small>CC BY [Isin et al. (2020)]</small>		diagram of the measurement plane on pelvic radiograph	a pelvic radiograph show the distance from the sacral bone to the
ImageCLEFmedCaption_2022_valid_084422.jpg <small>CC BY [Vijaywargiya et al. (2021)]</small>		mri axial view of the same	mri of the pelvis show a mass in the left adnexal region
ImageCLEFmedCaption_2022_valid_084402.jpg <small>CC BY [Urushidani et al. (2021)]</small>		head compute tomography show emphysema in the nasopharynx	computed tomography ct scan of the head and neck with contrast show a soft tissue

Figure 2: Examples from the official validation set with captions predicted by our best run model (using the pre-trained BLIP as initial parameter and perform secondary pre-training on the official training set)

5. Conclusions

This paper presents the work of ImageSem Group at the ImageCLEFmedical Caption 2022 task. We employed the bootstrapping language-image pre-training framework for the caption prediction subtask, and achieved a relatively balanced performance across all metrics. For further work, it would be helpful to multiple downstream tasks such as caption prediction by collecting more high-quality medical image-caption pairs, as well as introducing imaging diagnosis knowledge of specific diseases.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (Grant No. 61906214), the Beijing Natural Science Foundation (Grant No. Z200016).

References

- [1] B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022)*, LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.
- [2] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection, in: *CLEF2022 Working Notes, CEUR Workshop Proceedings*, CEUR-WS.org, Bologna, Italy, 2022.
- [3] Z. Guo, X. Wang, Y. Zhang, J. Li, Imagesem at imageclefmed caption 2019 task: a two-stage medical concept detection strategy, in: *ImageClef2019 working notes, CEUR Workshop Proceedings*, CEUR-WS.org, Lugano, Switzerland, 2019.
- [4] Y. Zhang, X. Wang, Z. Guo, J. Li, Imagesem at imageclef 2018 caption task: Image retrieval and transfer learning, in: *ImageClef2018 working notes, CEUR Workshop Proceedings*, CEUR-WS.org, Avignon, France, 2018.
- [5] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. (2022). URL: <https://arxiv.org/abs/2201.12086>.
- [6] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, *Nucleic Acids Research* 32 (2004) 267–270. doi:10.1093/nar/gkh061.
- [7] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embedding with instance loss, *Communications, and Applications ACM Transactions on Multimedia Computing* (2020) 1–23. doi:<https://doi.org/10.1145/3383184>.
- [8] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, *The handbook of brain theory and neural networks* 10 (1995) 3361.
- [9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, *NeurIPS* 32 (2019) 8026–8037.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, *ICLR* (2021).

- [11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, NAACL (2019) 4171–4186.
- [12] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, Computer Science (2017).
- [13] R. Adhikari, D. Manduva, S. V. Malayala, R. Singh, N. K. Jain, K. Deepika, T. Koritala, A rare case of vaping-induced spontaneous pneumomediastinum, 2021. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8435398/>.
- [14] Y. Isin, O. Hapa, Y. S. Kara, A. I. Kilic, , A. Balci, A ct study of the femoral and sciatic nerve periacetabular moving in different hip positions., 2021. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7488403/>.
- [15] K. Vijaywargiya, N. Kachhara, Q. Chahwala, A. Ruia, Carcinoma cervix leading to ichthyosis uteri: A rare case report, 2021. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8440718/>.
- [16] S. Urushidani, A. Kuriyama, A sudden decrease in voice volume: A rare manifestation of spontaneous pneumomediastinum, 2021. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8245747/>.