PCLmed at ImageCLEFmedical 2023: Customizing General-Purpose Foundation Models for Medical Report Generation

Bang Yang[1,2], Asif Raza[2], Yuexian Zou[2] and Tong Zhang[1,*]

[1] Peng Cheng Laboratory, Shenzhen 518055, China

[2] ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China

Abstract

Medical caption prediction which can be regarded as a task of medical report generation (MRG), requires the automatic generation of coherent and accurate captions for the given medical images. However, the scarcity of labelled medical image-report pairs presents great challenges in the development of deep and large-scale neural networks capable of harnessing the potential artificial general intelligence power like large language models (LLMs). In this work, we propose customizing off-the-shelf general-purpose large-scale pre-trained models, i.e., foundation models (FMs), in computer vision and natural language processing with a specific focus on medical report generation. Specifically, following BLIP-2, a state-of-the-art vision-language pre-training approach, we introduce our encoder-decoder-based MRG model. This model utilizes a lightweight query Transformer to connect two FMs: the giant vision Transformer EVA-ViT-g and a bilingual LLM trained to align with human intentions (referred to as ChatGLM-6B). Furthermore, we conduct ablative experiments on the trainable components of the model to identify the crucial factors for effective transfer learning. Our findings demonstrate that unfreezing EVA-ViT-g to learn medical image representations, followed by parameter-efficient training of ChatGLM-6B to capture the writing styles of medical reports, is essential for achieving optimal results. Our best attempt (PCLmed Team) achieved the 4^{th} and the 2^{nd}, respectively, out of 13 participating teams, based on the BERTScore and ROUGE-1 metrics, at the ImageCLEFmedical 2023 caption prediction task.

Keywords

Medical Image Captioning, Foundation Model, Vision-Language Pre-Trained Models, Bilingual Language Model, Transfer Learning

1. Introduction

Medical report generation (MRG) requires the automatic generation of accurate and fluent reports that describe the impressions and findings of medical images, making it promising to address the well-documented "physician burnout" phenomenon [1, 2]. As one of generative medical tasks [3, 4, 5, 6] and a vision-language (VL) task, MRG needs high-quality cross-modal annotations, i.e., medical image-report pairs. Although a few open-source datasets have been

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

*Corresponding author.

yangbang@pku.edu.cn (B. Yang); asifraza151@pku.edu.cn (A. Raza); zouyx@pku.edu.cn (Y. Zou); zhangt02@pcl.ac.cn (T. Zhang)

https://web.pkusz.edu.cn/adsp/ (Y. Zou)

0000-0003-2019-0377 (B. Yang); 0000-0002-8838-4963 (T. Zhang)

introduced to foster research in this direction [7, 8, 9], their data scale, i.e., up to hundreds of thousand of pairs, pose great challenges for developing deep and giant neural networks.

Recently, an increasing number of large-scaled pre-trained models, i.e., foundation models (FMs), are emerging in computer vision (CV) [10, 11, 12], natural language processing (NLP) [13, 14, 15], and their intersection [16, 17, 18]. These foundation models are built for general purposes and can handle a wide range of tasks by prompt engineering [19, 20, 21] or parameter-efficient transfer learning [22, 23, 24, 25]. More importantly, due to model and data scaling, FMs may acquire *emergent abilities* to solve tasks that are difficult for small models [26]. Given the shortage of labeled medical image-report pairs and the impressive powers of FMs, an intermediate problem rises up: can we effectively utilize off-the-shelf general-purpose FMs and adapt them for medical report generation?

In this work, inspired by the state-of-the-art VL pre-training approach BLIP-2 [16], we present our encoder-decoder-based MRG model, where a lightweight query Transformer (Q-Former) was used to bridge two FMs: EVA ViT-g [11] and ChatGLM-6B [14], a bilingual LLM trained to align to human intentions. Different from BLIP-2, where the Q-Former was pre-trained with image-text pairs to bridge the modality gap of two FMs, we directly fine-tune the model on the target dataset. Furthermore, we perform ablative experiments on the trainable components of our MRG model to identify the crucial factors for effective transfer learning. Through experimentation on ImageCLEFmedical 2023 caption prediction [9], we discover the significance of unfreezing EVA-ViT-g to learn medical image representations and parameter-efficient fine-tuning of ChatGLM-6B to capture the writing styles of medical reports. As a result, our best solution achieved the 4^{th} and 2^{nd} positions among 13 participating teams in the competition, as evaluated by BERTScore and ROUGE-1 metrics.

2. Related Works

Medical Report Generation Inspired by the success of neural-network-based image captioning [27, 28, 29, 30, 31], medical report generation has attracted enormous research interest in recent years. Mainstream approaches for medical report generation adopt the *de facto* encoder-decoder framework and focus on two aspects: 1) improving the cross-modal alignment between medical images and reports via reinforcement learning [32, 33], architecture design [34, 35, 36], or explicit loss constraints [37, 38] and 2) exploring retrieval- and knowledge-augmented report generation [32, 39, 38, 40, 41]. These methods use dataset-related annotations to train their model, whose power, however, is restricted by the small number of labeled pairs.

Adaptation of Foundation Models With more and more foundation models (FMs) springing up in CV, NLP, and their intersection, how to efficiently adapt FMs to a specific task becomes a research hotspot. One effective technique is *prompt engineering* [42], which aims to affect the behaviors of language FMs by providing them with a textual template filled with task-related priors [43, 23], demonstrations of several examples [19, 44], or a chain of thoughts [20, 21, 45]. Although such a technique usually does not require model optimization, it could be sensitive to the selection of templates and few-shot examples [46, 47]. Another popular technique is known as parameter-efficient transfer learning, which introduces lightweight components, e.g.,

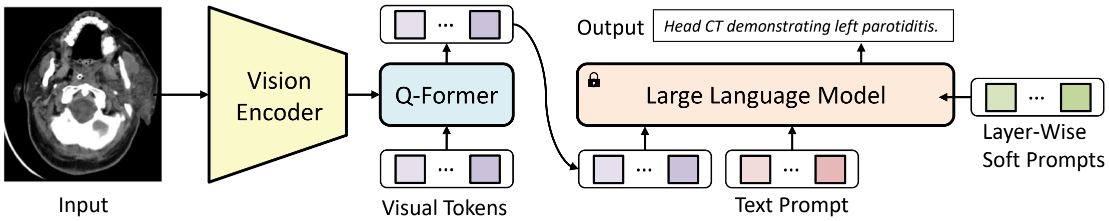


Figure 1: Overview of our model for medical report generation.

Adapter [22], continuous prompts [23, 25], and LoRA [24], into FMs to influence their activation and responses. With the aforementioned two techniques, recent practices have demonstrated the effectiveness of adapting general-purpose FMs to the medical domain [48, 49, 50]. Different from these practices, we focus on medical report generation in this work.

3. Approach

Overview As shown in Figure 1, our model for medical report generation comprises three networks following BLIP-2 [16]: a vision encoder, a query Transformer (Q-Former), and a LLM. To effectively transfer the knowledge of FMs, we unfreeze the vision encoder and parameter-efficiently fine-tune the language model with P-Tuning [25]. Given the medical image I, the text prompt $\mathbf{x} = \{x_1, x_2, \ldots, x_M\}$, and the target report $\mathbf{y} = \{y_1, y_2, \ldots, v_T\}$, our model minimize the following negative log likelihood:

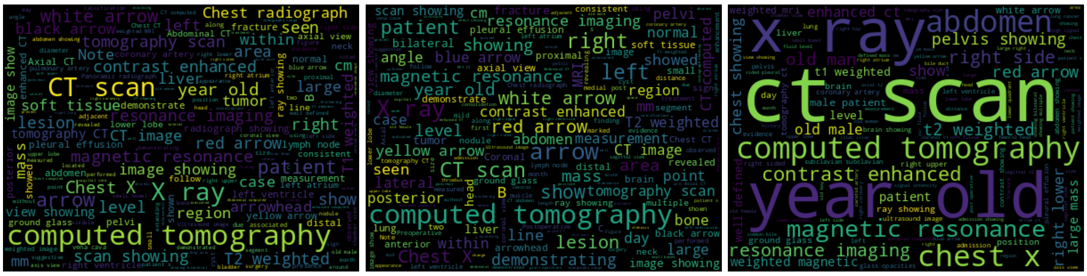
$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p(y_t|\mathbf{y}_{<\mathbf{t}}, \mathbf{x}, I; \theta), \tag{1}$$

where θ denotes all parameters of the model, and $\mathbf{y}_{<\mathbf{t}}$ means previously generated tokens before the t-th time step. Next, we introduce each component of our model.

Vision Encoder We adopt EVA-ViT-g [11] as the vision encoder to extract patch-based features $\boldsymbol{R} \in \mathbb{R}^{N \times d_v}$ of the medical image I. Since EVA-ViT-g uses a patch size of 14, we will obtain a long feature sequence for a high-resolution image, e.g., $N = 677$ for a 364×364 image.

Query Transformer To reduce the computation costs of the subsequent modeling process, we adopt a Q-Former with K ($K < N$) learnable visual tokens to obtain aggregated visual features. In the implementation, Q-Former is a BERT-like encoder [51] with cross-attention layers inserted at a certain frequency.

Large Language Model We adopt ChatGLM-6B [14] as the decoder to generate texts. Specifically, ChatGLM-6B is a decoder-only Transformer. Hence, we need to prefix the decoder input sequence with the aggregated visual features to achieve vision-grounded medical report generation. It is noted that ChatGLM-6B generates texts in Chinese or in English adaptively, i.e., no explicit signal is fed into the model to indicate which language to be generated.



(a)

(b)

(c)

Figure 2: Word clouds of ground-truth reports from the training set (a) and the validation set (b). In (c), we show the word cloud of reports generated by our model on the validation set.

P-Tuning To preserve the ability of the LLM, we adopt the P-Tuning [25] technique to fine-tune ChatGLM-6B. Specifically, the original self-attention layers of ChatGLM-6B follow the formulation below:

$$Attn(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are queries, keys, and values mapped from the same input sequence. P-Tuning introduces layer-wise soft prompts ($\boldsymbol{P_K}$ and $\boldsymbol{P_V}$) into the self-attention layers and modifies the keys and values. Thus, Equation. 2 is re-formulated as follows:

$$Attn(\boldsymbol{Q}, \mathrm{Concat}(\boldsymbol{P_K}, \boldsymbol{K}), \mathrm{Concat}(\boldsymbol{P_V}, \boldsymbol{V})). \tag{3}$$

4. Experiments

4.1. Brief Introduction of the Competition and Task

ImageCLEFmedical have been a part of ImageCLEF every year since 2004 and pays specific attention to the multimedia retrieval applications in medical tasks [52]. In 2023, ImageCLEFmedical consists of four main tasks: automatic image captioning, controlling the quality of synthetic medical images, visual question answering for colonoscopy images, and medical dialogue summarization, with several subtasks within each main task. We participated the captioning task of ImageCLEFmedical 2023, which has been the 7th edition since ImageCLEF 2017, with a specific focus on the caption prediction subtask.

Specifically, ImageCLEFmedical 2023 caption prediction uses an updated and extended version of the Radiology Objects in COntext (ROCO) dataset [53]. It provides 60,918, 10,437, and 10,473 radiology images for training, validation, and testing, respectively. Images are originated from biomedical articles and each image is annotated with one medical report. In Figure 2 (a) and (b), we visualize the word clouds of ground-truth reports from the training and validation sets. As we can observe, there are many computed tomography (CT) and X-ray images in the dataset. Besides, there are some common expressions like “white arrow”, indicating that a large portion of images have been marked by humans.

4.2. Experimental Setups

Dataset We use the official dataset and splits provided by ImageCLEFmedical 2023 caption prediction, as introduced in Section 4.1.

Metrics We report two metrics following the guidelines of the competition[1], namely BERTScore [54] and ROUGE-1 [55]. Specifically, BERTScore is a model-based metric that calculates the semantic similarity of two sentences. ROUGE-1 measures the number of matching unigrams between a model-generated text and a reference.

Image Processing During training, we process images with random resized cropping, horizontal flipping, and RandomAugment [56]. During inference, we directly resize images to a specific resolution. We consider two image sizes in this paper: 224 and 364.

Model Settings We use EVA-ViT-g [11] as the vision encoder and v1.0 ChatGLM-6B[2] [14] as the decoder by default. We always prompt ChatGLM-6B with the text "Question: What is the radiology report for this image? Answer:" during training and inference. Following BLIP-2's official code[3], we implement BERT-like Q-Former, where cross-attention layers are inserted every two Transformer blocks and there are $K = 32$ visual tokens to be learned. For P-tuning, we append 4 soft prompt tokens to the key and value sequences of each self-attention layer of ChatGLM-6B. Given that the number of layers and the model dimension of ChatGLM-6B is 28 and 4,096 respectively, P-tuning will introduce 0.9 M additional parameters. Besides ChatGLM-6B, we consider two more variants of language models: OPT-2.7B [57] and a randomly initialized Transformer-based decoder [58] (abbr. TF-Base).

Hyper-Parameters In the training phase, we truncate reports into a maximum length of 64. We use AdamW [59] and L2 weight decay of 0.05 to train models with 12 samples per batch for 10 epochs. The learning rate is increased to 1e-4 after 2,000 warm-up steps and then follows a cosine annealing scheduler. After training, we use the beam search decoding algorithm with a beam size of 5 to generate medical reports. We set the repetition penalty to 2 to avoid duplication and limit the length of generated reports to at least 8.

4.3. Results

We conducted five runs at ImageCLEFmedical 2023 caption prediction, with the second and third runs being the intermediate results of the fourth and fifth runs, respectively. Therefore, there are only three valid results on the test set. Table 1 summarizes the performance of different model variants. We have the following observations.

- Comparing #{1, 2, 3}, we can see that using a frozen language FM is superior to training a small language model from scratch. This suggests that the syntactic knowledge of language FMs learned from common corpora can benefit medical report generation.

[1] https://www.imageclef.org/2023/medical/caption
[2] https://github.com/THUDM/ChatGLM-6B
[3] https://github.com/salesforce/LAVIS

Table 1
Performance on ImageCLEFmedical 2023 caption prediction. †: Trainable components. ∗: P-tuning. Q-Former is omitted here, which is trainable and remains the same in all experiments.

#	Image Size	Vision Encoder	Language Model	#Parameters		Validation Set		Test Set	
				Total	Trainable	BERTScore	ROUGE-1	BERTScore	ROUGE-1
1	224	EVA-ViT-G	TF-Base†	1.1B	140.7M	0.524705	0.181943	-	-
2	224	EVA-ViT-G	OPT-2.7B	2.8B	105.2M	0.599041	0.230394	-	-
3	224	EVA-ViT-G	ChatGLM-6B	7.3B	108.3M	0.606204	0.231744	-	-
4	224	EVA-ViT-G	ChatGLM-6B∗	7.3B	109.2M	0.606173	0.240584	0.607918	0.242249
5	224	EVA-ViT-G†	ChatGLM-6B∗	7.3B	1.1B	0.617632	0.263236	-	-
6	364	EVA-ViT-G†	ChatGLM-6B∗	7.3B	1.1B	**0.622668**	**0.271068**	**0.614836**	**0.253279**

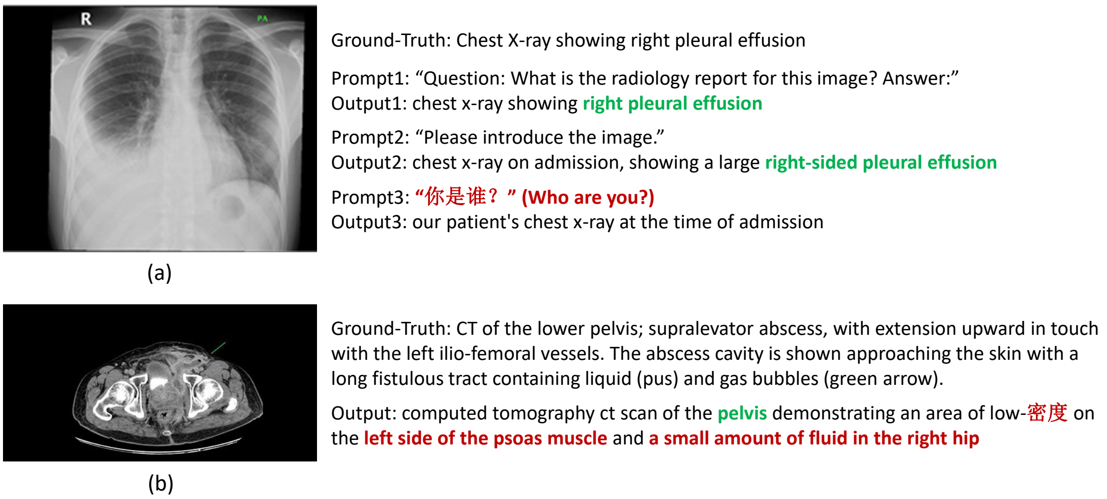
- Comparing #{3, 4}, we can observe that using the P-tuning technique enables ChatGLM-6B to generate more faithful medical reports, leading to 0.9% absolute improvements under the ROUGE-1 metric.
- Comparing #{4, 5}, we can see obvious improvements in both metrics after training the vision encoder together. This indicates the great differences between open-domain images and medical images and the necessity of medical image representation learning.
- Comparing #{5, 6}, we can conclude that a higher image resolution can preserve more details and thus benefits accurate medical report generation.

In short, customizing general-purpose FMs for medical report generation is feasible and can boost performance by a large margin (e.g., #6 vs. #1). Nevertheless, there still exist many potential improvement directions of our FM-based model, e.g., 1) designing a multi-stage training procedure to learn from unlabeled medical images and texts and 2) speeding up the inference process via model compression and knowledge distillation. Additional scores (BLEU, BLEURT, METEOR, CIDEr, CLIPScore) for the competition can be found in the results folder [52] and on the website [9].

4.4. Discussion

We here give two qualitative examples in Figure 3 to show three problems of our model.

Catastrophic Forgetting of the Question-Answering (QA) Ability Our model is built upon ChatGLM-6B, a bilingual LLM with the QA ability. Considering that we keep ChatGLM-6B frozen all the time, it would be interesting to see how the QA ability is affected before and after training. As shown in Figure 3 (a), when we change the prompt text from the default one (i.e., Prompt1) to the one irrelevant to the image content (i.e., Prompt3), the model's response does not meet with our expectation. In other words, we can not observe the ability of zero-shot visual QA as in the BLIP-2 paper [16]. We suspect such catastrophic forgetting is caused by two reasons: 1) model training does not include visual QA data, and 2) the prompt text fed into the model is fixed during training.



Ground-Truth: Chest X-ray showing right pleural effusion

Prompt1: “Question: What is the radiology report for this image? Answer:”
Output1: chest x-ray showing **right pleural effusion**

Prompt2: “Please introduce the image.”
Output2: chest x-ray on admission, showing a large **right-sided pleural effusion**

Prompt3: **“你是谁？ ” (Who are you?)**
Output3: our patient's chest x-ray at the time of admission

(a)

Ground-Truth: CT of the lower pelvis; supralevator abscess, with extension upward in touch with the left ilio-femoral vessels. The abscess cavity is shown approaching the skin with a long fistulous tract containing liquid (pus) and gas bubbles (green arrow).

Output: computed tomography ct scan of the **pelvis** demonstrating an area of low-密度 on the **left side of the psoas muscle** and **a small amount of fluid in the right hip**

(b)

Figure 3: Two qualitative examples on the validation set of ImageCLEFmedical 2023 caption prediction. Although our FM-based model may generate accurate keywords, it suffers from catastrophic forgetting of the question-answering ability, hallucination, and code-switching generation problems.

Hallucination As shown in Figure 3 (b), our model may produce hallucinate unintended text. The word cloud in Figure 2 (c) also shows that our model generates age information excessively. Such a hallucination problem is common among various natural language generation tasks [60, 61, 62]. One reason for this problem is that the cross-entropy loss for language modeling leads to the problem of exposure bias [63]. To solve this problem, reinforcement learning is one of the effective techniques [64, 13].

Code-Switching Generation As shown in Figure 3 (b), our model may generate Chinese tokens unexpectedly though it is fine-tuned on English-only data. On one hand, it indicates the risks to inherit the drawbacks of pre-trained models. On the other hand, it is a common phenomenon in multilingual machine translation systems [65, 66]. To overcome this problem, we may feed an explicit signal into the model to indicate which language to be generated or substitute the multilingual vocabulary with a monolingual one. Furthermore, we verify if the sporadic non-English tokens are semantically correct in Table 2, where we utilize ChatGPT to polish the generated reports with non-English tokens[4]. As we can see, performance does not change basically. Therefore, there is still a huge room to improve our model.

5. Conclusion

In this work, we presented a simple yet effective attempt to customize general-purpose foundation models (FMs) for medical report generation. Our study not only shows the significant per-

[4] The prompt for ChatGPT is “You are a writing assistance. You should polish the requested sentence following the below rules: (1) ensure that the output sentence is in English; (2) remove any special characters; (3) remove repetitive and redundant expressions; Please follow the rules to re-write this sentence:”.

Table 2
Effect of whether using ChatGPT to polish the generated reports with non-English tokens. The performance does not change basically after using ChatGPT, indicating that non-English tokens are semantically wrong.

#	Use ChatGPT?	# polished reports	# total reports	Test Set	
				BERTScore	ROUGE-1
6	×	0	10,473	0.614836	**0.253279**
7	√	120	10,473	**0.615190**	0.252756

formance boost brought by FMs but also reveals the challenges of fine-tuning LLMs: catastrophic forgetting of the question-answering ability, hallucination, and code-switching generation. All these demerits should be addressed to develop a multilingual trustworthy chat-bot in a specific medical scenario. We leave this research to our future work. Source codes with model weights will be available at OpenMedIA[5] [67] after the conference.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (grant No. U21A20523 and 62176008). The computing resources of Pengcheng Cloudbrain are used in this research. We acknowledge the support provided by OpenI Community[6].

References

[1] D. S. Tawfik, J. Profit, T. I. Morgenthaler, D. V. Satele, C. A. Sinsky, L. N. Dyrbye, M. A. Tutty, C. P. West, T. D. Shanafelt, Physician burnout, well-being, and work unit safety grades in relationship to reported medical errors, in: Mayo Clinic Proceedings, 2018, pp. 1571–1580.

[2] C. P. West, L. N. Dyrbye, T. D. Shanafelt, Physician burnout: Contributors, consequences and solutions, Journal of Internal Medicine 283 (2018) 516–529.

[3] A. Chaves, C. Kesiku, B. Garcia-Zapirain, Automatic text summarization of biomedical text data: A systematic review, Information 13 (2022) 393.

[4] F. Liu, B. Yang, C. You, X. Wu, S. Ge, Z. Liu, X. Sun, Y. Yang, D. Clifton, Retrieve, reason, and refine: Generating accurate and faithful patient instructions, Advances in Neural Information Processing Systems 35 (2022) 18864–18877.

[5] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, D. Capurro, A survey on deep learning and explainability for automatic report generation from medical images, ACM Computing Surveys (CSUR) 54 (2022) 1–40.

[6] F. Liu, B. Yang, C. You, X. Wu, S. Ge, A. Woicik, S. Wang, Graph-in-graph network for

[5] https://openi.pcl.ac.cn/OpenMedIA
[6] https://git.openi.org.cn

automatic gene ontology description generation, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 1060–1068.

[7] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, Journal of the American Medical Informatics Association 23 (2016) 304–310.

[8] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific data 6 (2019) 317.

[9] 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.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, arXiv preprint arXiv:2304.02643v1 (2023).

[11] 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.

[12] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al., InternImage: Exploring large-scale vision foundation models with deformable convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14408–14419.

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[14] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., GLM-130B: An open bilingual pre-trained model, International Conference on Learning Representations (2023).

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMA: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[16] 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:2301.12597v3 (2023).

[17] OpenAI, GPT-4 technical report, arXiv preprint arXiv:2303.08774v3 (2023).

[18] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, N. Duan, Visual ChatGPT: Talking, drawing and editing with visual foundation models, arXiv preprint arXiv:2303.04671v1 (2023).

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: NeurIPS, 2020.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, in: NeurIPS, 2022.

[21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot

reasoners, in: NeurIPS, 2022.

[22] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.

[23] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4582–4597.

[24] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[25] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, J. Tang, P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 61–68.

[26] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, Transactions on Machine Learning Research (2022).

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057.

[28] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys 51 (2019) 1–36.

[29] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, L. Wang, Scaling up vision-language pre-training for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17980–17989.

[30] B. Yang, T. Zhang, Y. Zou, Clip meets video captioning: Concept-aware representation learning does matter, in: Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, Proceedings, Part I, Springer, 2022, pp. 368–381.

[31] B. Yang, F. Liu, Y. Zou, X. Wu, Y. Wang, D. A. Clifton, ZeroNLG: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation, arXiv preprint arXiv:2303.06458v1 (2023).

[32] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, Advances in neural information processing systems 31 (2018).

[33] H. Qin, Y. Song, Reinforced cross-modal alignment for radiology report generation, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 448–458.

[34] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven Transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1439–1449.

[35] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer, 2021,

pp. 72–82.

[36] Z. Chen, Y. Shen, Y. Song, X. Wan, Cross-modal memory networks for radiology report generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5904–5914.

[37] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, Y. Zou, Unify, align and refine: Multi-level semantic alignment for radiology report generation, arXiv preprint arXiv:2303.15932v4 (2023).

[38] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, X. Chang, Dynamic graph enhanced contrastive learning for chest x-ray report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3334–3343.

[39] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13753–13762.

[40] C. Y. Li, X. Liang, Z. Hu, E. P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6666–6673.

[41] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, D. Xu, When radiology report generation meets knowledge graph, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 12910–12917.

[42] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text Transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[44] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, in: NeurIPS, 2022.

[45] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., PaLM: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311v5 (2022).

[46] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8086–8098.

[47] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, International Journal of Computer Vision 130 (2022) 2337–2348.

[48] V. Liévin, C. E. Hother, O. Winther, Can large language models reason about medical questions?, arXiv preprint arXiv:2207.08143v3 (2022).

[49] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, arXiv preprint arXiv:2212.13138v1 (2022).

[50] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on medical challenge problems, arXiv preprint arXiv:2303.13375v2 (2023).

[51] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional Transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[52] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcia 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.

[53] 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.

[54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2019.

[55] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[56] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 702–703.

[57] 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:2205.01068v4 (2022).

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[59] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR, 2019.

[60] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, Object hallucination in image captioning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4035–4045.

[61] B. Yang, Y. Zou, F. Liu, C. Zhang, Non-autoregressive coarse-to-fine video captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 3119–3127.

[62] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[63] M. Ranzato, S. Chopra, M. Auli, W. Zaremba, Sequence level training with recurrent neural networks, in: International Conference on Learning Representations, 2016.

[64] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7008–7024.

[65] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, arXiv preprint arXiv:2008.00401v1 (2020).

[66] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672v3 (2022).

[67] 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.