UIT-2Q2T at ImageCLEFmedical 2024 Caption: Multimodal medical image captioning using Bootstrapping Language-Image Pre-training

Notebook for the UIT-2Q2T Team at CLEF 2024

Thien V. Phan[1,2], Trinh K. Nguyen[1,2], Quang A.D.D. Hoang[1,2], Quan T. Phan[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

Introduction: Medical image captioning is an important AI task in healthcare, automating the generation of text descriptions to support the management and interpretation of medical images. Our team, UIT-2Q2T, participated in the second task of the ImageCLEFmedical 2024 Caption challenge using the ROCOv2 dataset with the Bootstrapping Language-Image Pre-training (BLIP) approach.
Methods: Our approach leveraged the BLIP architecture for multimodal medical image captioning. This architecture employs a Vision Transformer (ViT) as the image encoder and a Bidirectional Encoder Representations from Transformers (BERT) as the text model.
Results: We ranked 5th according to BERTscore and placed 3rd with ROUGEScore, BLEURTScore, and RefCLIPScore. Additionally, we achieved 2nd place for BLEU-1, METEOR, and CIDEr scores. Notably, we obtained the top position with a CLIPScore of 0.827074, demonstrating the effectiveness of our approach in medical image captioning.
Conclusion: Our participation in the ImageCLEFmedical 2024 Caption challenge demonstrated the effectiveness of the BLIP architecture for medical image captioning, achieving a high CLIPScore of 0.82707. This result demonstrates the model's potential to generate accurate and informative textual descriptions from medical images, thereby aiding diagnosis and assisting non-experts in understanding medical images.

Keywords

CLEF 2024, Medical image processing, Image captioning, BERT, Pre-trained models, BLIP

1. Introduction

Image captioning, a well-established field in artificial intelligence (AI), finds applications across diverse domains. In healthcare, the increasing availability of medical imaging equipment and the efficiency of diagnosis based on visual data have fueled the popularity of image-based patient diagnosis. Medical image captioning models address this need by automatically analyzing and describing medical images. These models generate textual descriptions that assist doctors in diagnosing diseases, understanding physiological processes, and enabling non-experts to interpret medical imagery.

This field integrates computer vision and natural language processing, demanding an understanding of image components and their relationships [1]. Various models, such as the Show-Attend-Tell, GPT-3, and BioLinkBERT-Large, have been utilized to generate comprehensive and descriptive captions for medical images, including radiological scans and histopathological specimens [2] [3]. Transformer-based approaches, like the Global-Local Visual Extractor (GLVE) and Cross Encoder-Decoder Transformer (CEDT), have shown promise in capturing both global and local features of images, enhancing the accuracy of generated captions [4]. These advancements in medical image captioning not only facilitate

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
21522628@gm.uit.edu.vn (T. V. Phan); 21522717@gm.uit.edu.vn (T. K. Nguyen); 21522509@gm.uit.edu.vn (. Q. A.D.D. Hoang); 21522502@gm.uit.edu.vn (. Q. T. Phan); thienntb@uit.edu.vn (. T. B. Nguyen-Tat)

clinical workflows and decision-making but also contribute significantly to medical education by providing quantitative indicators and assessments for learning outcomes [5].

To successfully deploy image captioning in healthcare, it is essential to integrate effective algorithms and use a sufficiently large and diverse training dataset. Our team participated in ImageCLEF 2024 [6] for the ImageCLEFmedical 2024 Caption [7] task which consists of 2 subtask: Concept Detection, Caption Prediction. We mainly focus on the latter. Here, participants are required to automatically generate captions for given medical images, which could be of various modalities, such as ultrasound, X-Ray, Computer Tomography (CT), Magnetic Resonance Imaging (MRI), etc.

Our approach for the caption prediction subtask is based on Bootstrapping Language-Image Pre-training (BLIP) [8] architecture with a Vision Transformer (ViT) image Encoder [9]. In this paper, Section 2 outlines the task and dataset descriptions. Section 3 describes our proposed methodology. Section 4 details the implementation and results of the experiments. Finally, in Section 5, we conclude by summarizing the results, discussing the weaknesses, and outlining potential improvements for the future.

2. Task and Dataset Descriptions

At ImageCLEFmedical 2024, we participated in the image captioning task. This is the 8th edition of the ImageCLEFmedical caption task. In this section, we will introduce the task in the ImageCLEFmedical 2024 Caption and the dataset used for this challenge.

2.1. Task Description

ImageCLEFmedical 2024 Caption [7] is one of ImageCLEFmedical's tasks to create descriptive captions for visual content. The tasks in ImageCLEFmedical Caption include two sub-tasks:

1. Concept detection: Based on the visual image content, this subtask provides the foundation for the scene understanding step by identifying the individual elements from which the annotation is generated.
2. Captions prediction: The core task is to create descriptive captions for given images. Leveraging identified concepts and contextual understanding, the models are tasked with generating concise and informative textual descriptions that accurately reflect the visual content depicted in the image.

In this study, we focus on the second sub-task based on the provided dataset ROCOv2 [10].

2.2. Dataset Descriptions

The dataset for this task is ROCOv2 [10] - an extended version of ROCO [11]. It is a multimodal dataset consisting of radiological images and associated medical concepts and captions extracted from the PubMed Open Access subset. All images in the dataset were accompanied by a caption, which form the labels for the caption prediction task. Each caption was pre-processed by removing links from the captions. The splits for the dataset are as follows:

- Training Set: Consists of 70,108 radiology images.
- Validation Set: Consists of 9,972 radiology images.
- Test Set: Consists of 17,237 radiology images.

As shown in Figure 1, the majority of captions in the dataset range from 50 to 150 words in length. Similarly, Figure 2 illustrates that among the six imaging modalities represented in the dataset, CT scans and X-rays are predominant, accounting for 24,227 and 19,363 samples in the training set, respectively.

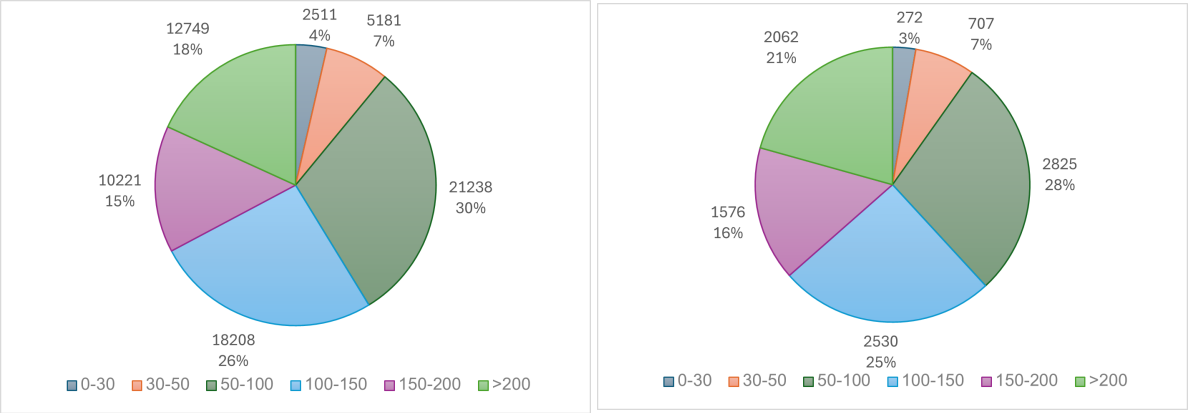


Figure 1: Distribution of caption lengths in the Training Set (left) and Validation Set (right).

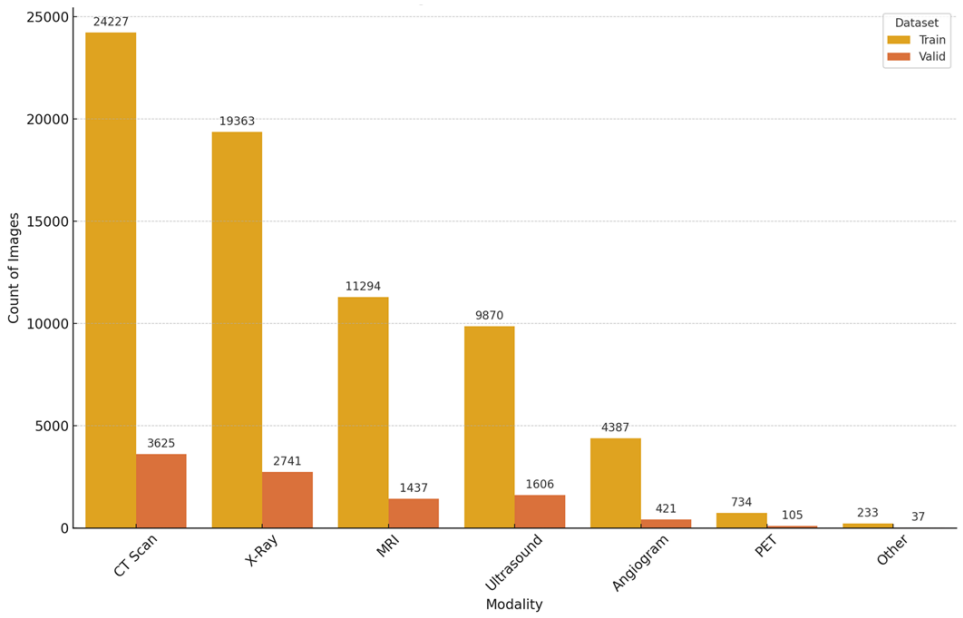


Figure 2: Distribution of image modalities in Train and Validation Sets.

3. Methods

In this study, the BLIP model was employed to tackle the image captioning task. This approach involves finetuning the BLIP model on the competition dataset, which consists of diverse and challenging image-caption pairs. The pipeline of our method is illustrated in Figure 3, showcasing the steps involved in adapting the BLIP model for our specific image captioning task.

3.1. Models

Bootstrapping Language-Image Pre-training (BLIP) [8] is a Vision-Language Pre-training (VLP) framework which transfers flexibly to both vision-language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones.

The model uses Vision Transformer (ViT) [9] which divides the input image into patches and encodes them as a sequence of embedding with the addition of [CLS] token to represent the globe image feature. As the authors mentioned ViT uses less computation cost and is a straightforward method, and is being adopted by recent methods.

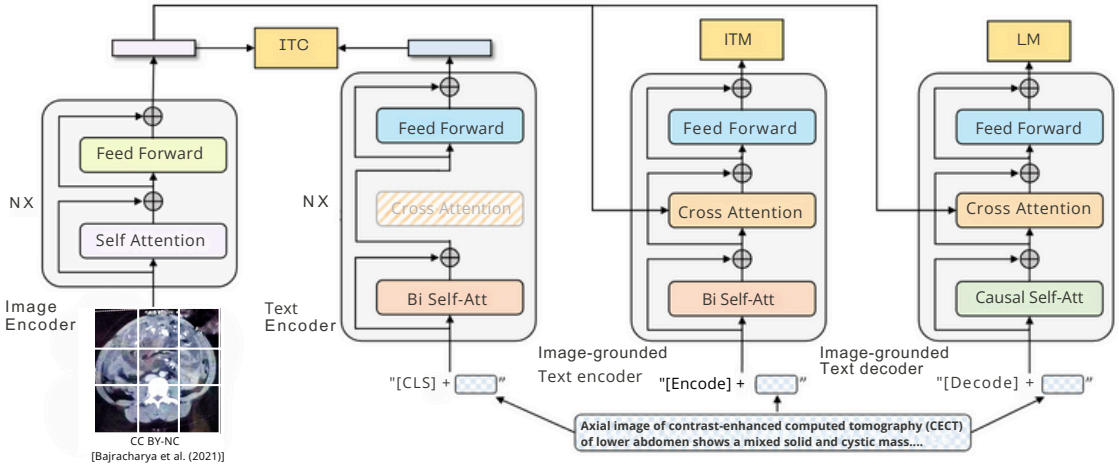


Figure 3: Pre-training model architecture and objectives of BLIP (same parameters have the same color). The multimodal mixture of encoder-decoder was proposed, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.

To be able to train or pre-train the model for understanding and generation tasks, a multimodal mixture of an encoder and decoder is used, integrating three functionalities and three objectives, as illustrated in Figure 3. The functionalities include:

- **Unimodal Encoder:** Encodes either the image or the text separately without considering the other modality. This helps in understanding individual representations.
- **Image-grounded Text Encoder:** Encodes text while being conditioned on the image, allowing the model to capture relationships between visual and textual information.
- **Image-grounded Text Decoder:** Generates text based on the given image, useful for tasks like image captioning where the output is text describing the input image.

The objectives are:

- **Image-Text Contrastive Loss (ITC):** Ensures that paired image and text representations are closer together in the embedding space compared to unpaired ones. This helps the model learn strong associations between images and their corresponding texts.
- **Image-Text Matching Loss (ITM):** Assesses whether a given image and text pair match or not, promoting accurate image-text alignment in the embedding space.
- **Language Modeling Loss (LM):** Focuses on generating coherent and contextually accurate text based on given inputs, improving the model's language generation capabilities.

These functionalities and objectives together enable the BLIP model to perform both vision-language understanding and generation tasks effectively.

3.2. Evaluation Metrics

Following the guidelines provided by the competition organizers, we employed two main metrics: BERTScore [12] and ROUGEScore [13]. To calculate BERTScore, we use the "microsoft/deberta-xlarge-mnli" model, which can be found on the Hugging Face Model Hub[1]. Additionally, other metrics

[1]https://huggingface.co/microsoft/deberta-xlarge-mnli (Last accessed: May 17, 2024)

such as BLEU-1 [14], BLEURT [15], METEOR [16], CIDEr [17], CLIPScore [18], RefCLIPScore [19], ClinicalBLEURTScore [20], and MedBERTScore [20] were also applied for evaluation.

As the organizers' instructions, the captions underwent preprocessing through three steps: conversion to lowercase, replacement of numbers with a special token, and removal of punctuation. This preprocessing aimed to standardize the text inputs and enhance the quality of evaluation result.

4. Experiments

In this section, we present our experimental setup and results for evaluating the BLIP model in the ImageCLEFmedical 2024 Caption challenge. The experiments were designed to test the model's performance across different configurations and metrics, aiming to generate accurate and informative captions for medical images. We describe the setup in detail and discuss the results obtained from various test scenarios.

4.1. Experimental Setup

In our experiments, we employed BLIP model (base/large) from pre-trained checkpoints. For BLIP base, weights were utilized from the checkpoint "Salesforce/blip-image-captioning-base"[2]. Training was conducted over 15 epochs with an initial learning rate of 1e-5. A StepLR scheduler was used to decrease the learning rate by a factor of 10 every 3 epochs. For the BLIP large model, weights were utilized from the checkpoint "Salesforce/blip-image-captioning-large"[3]. Training was conducted over 5 epochs with an initial learning rate of 1e-5 and was stopped when the loss ceased to decrease. Throughout all experiments, the AdamW optimizer [21] was used. Input images were resized to 224x224, and the maximum length of text input was set to 200 tokens. To facilitate model training, a single GPU A100 PCIE 40GB was used.

For each model, experiments were conducted with four different generation settings using no_repeat_ngram_size = 3 to prevent the model from repeating any n-gram of size 3 within the generated text. The generation settings are as follows:

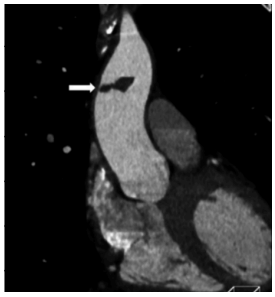
(1) Greedy Search: This setting selects the token with the highest probability at each step, ensuring a straightforward and fast generation process but potentially missing out on more diverse or optimal sequences.
(2) Beam Search with beam_size = 3: Beam search keeps the top 3 most probable sequences at each generation step, allowing for more exploration of potential sequences compared to greedy search.
(3) Beam Search with beam_size = 4: Similar to the previous setting, but with a beam size of 4, which balances between exploration and computational efficiency.
(4) Beam Search with beam_size = 5: This setting further increases the beam size to 5, allowing for more comprehensive exploration while balancing computational efficiency.
(5) Beam Search with beam_size = 10: With a beam size of 10, this setting aims for a broader exploration of possible sequences, potentially improving the quality of text generation at the cost of higher computational resources.

The source code for our experiments is available on GitHub[4].

4.2. Experimental Results

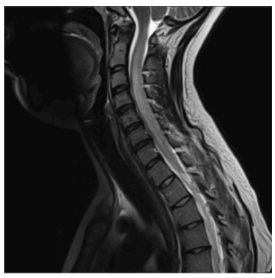
To evaluate the effectiveness of our approach using the BLIP model on the image captioning task, we conducted a series of experiments on the competition dataset. This section presents the results, highlighting the model's performance in generating captions. We compared the experimental results of the model using different configurations and also benchmark our model's performance against other teams'.

[2] https://huggingface.co/Salesforce/blip-image-captioning-base (Last accessed: May 17, 2024)
[3] https://huggingface.co/Salesforce/blip-image-captioning-large (Last accessed: May 17, 2024)
[4] https://github.com/QuangHoang059/DS312



CC BY [Sato et all. (2022)]

Ground truth: Computed tomography (CT) shows floating thrombosis (white arrow).
Prediction with greedy Search: contrast - enhanced computed tomography image of the aortic arch (white arrow).
Prediction with Beam Search (beam_size = 3): contrast - enhanced computed tomography image of the aortic arch (white arrow).
Prediction with Beam Search (beam_size = 4): contrast - enhanced computed tomography image of the aortic arch (white arrow).
Prediction with Beam Search (beam_size = 5): contrast - enhanced computed tomography image of the aortic arch (white arrow).
Prediction with Beam Search (beam_size = 10): contrast - enhanced computed tomography image of the aortic arch (white arrow).



CC BY-NC
[Trowbridge et all. (2022)]

Ground truth: Early sagittal T2-weighted MRI.
Prediction with greedy Search: sagittal t2 - weighted mri of the thoracic spine.
Prediction with Beam Search (beam_size = 3): sagittal t2 - weighted magnetic resonance image of the cervical spine.
Prediction with Beam Search (beam_size = 4): sagittal t2 - weighted magnetic resonance image of the cervical spine.
Prediction with Beam Search (beam_size = 5): sagittal t2 - weighted magnetic resonance image of the cervical spine.
Prediction with Beam Search (beam_size = 10): sagittal t2 - weighted mri of the thoracic spine.

Figure 4: Two examples of predicted results and ground truths in the validation set of the caption prediction task.

4.2.1. Results on Validation Set

Table 1
Evaluation results of BLIP base and large models on the validation set in 5 generation configurations.

	ROUGE	BERTscore	BLEU
BLIP base (1)	0.263178	0.659321	0.291905
BLIP base (2)	0.264012	0.658852	**0.300932**
BLIP base (3)	0.264665	0.659548	0.299855
BLIP base (4)	**0.264674**	**0.659648**	0.297638
BLIP base (5)	0.263178	0.659321	0.291905
BLIP large (1)	0.269548	0.666101	0.285273
BLIP large (2)	0.274387	0.667651	0.295454
BLIP large (3)	**0.274497**	**0.667971**	**0.295484**
BLIP large (4)	0.272249	0.667263	0.292144
BLIP large (5)	0.269548	0.666101	0.285273

As shown in Table 1, the BLIP large model outperforms the BLIP base model. Both models show generative capabilities, with beam search outperforming greedy search across ROUGEScore, as well as BERTScore and BLEUScore. Specifically, the BLIP base model achieves its highest BERTScore and ROUGE score with a beam size of 5, and its best BLEU score with a beam size of 3. The BLIP large model attains optimal results across all three metrics with a beam size of 4. Additionally, as illustrated by the two examples in Figure 4, the model accurately identifies objects and colors (white arrow), as well as different imaging modalities (CT and sagittal T2-weighted MRI).

4.2.2. Results on Test Set

According to the private test results announced by the organizing committee and partially presented in Table 2, our team ranked 5th based on the BERTScore metric. We achieved 3rd place with ROUGE, BLEURT, and RefCLIPScore metrics. For BLEU-1, METEOR, and CIDEr scores, we achieved 2nd place.

Table 2
Results table on the private test set for the top 5 teams based on the primary score.

Team	BERTScore	ROUGE	BLEU-1	BLEURT	METEOR	CIDEr	CLIPScore
pclmed	0.629913	0.272626	0.268994	0.337626	0.113264	0.268133	0.823614
CS_Morgan	0.628059	0.250801	0.209298	0.317385	0.092682	0.245029	0.821262
DarkCow	0.626720	0.245228	0.195044	0.306005	0.088897	0.224250	0.818440
auebnlpgroup	0.621112	0.204883	0.111034	0.289907	0.068022	0.176923	0.804067
2Q2T	**0.617814**	**0.247755**	**0.221252**	**0.313942**	**0.098590**	**0.220037**	**0.827074**

Notably, we attained 1st place with a CLIPScore of 0.827074. These results demonstrate the model's expected performance.

5. Conclusion and Future work

In this paper, we implemented and experimented with the BLIP model for the task of medical image captioning in the ImageCLEFmedical 2024 Caption challenge. The experimental results across various configurations showed promising outcomes, with the model achieving a CLIPScore of 0.82707 on the test set of the ROCOv2 dataset. Despite these achievements, there is still room for improvement in our research. The primary weakness of the model is its pre-training on a dataset significantly different from the medical domain, resulting in considerable bias.

Moving forward, we aim to enhance the model's accuracy by utilizing pre-trained models with datasets that are more closely aligned with medical and diagnostic domains. Additionally, we plan to apply preprocessing methods tailored to different types of medical images to further improve the performance. Exploring domain-specific augmentation techniques and integrating more diverse medical datasets could also provide substantial gains. By addressing these areas, we hope to develop a more robust and accurate medical image captioning model, which can be a valuable tool in clinical settings for aiding diagnosis and assisting non-experts in understanding medical imagery.

Furthermore, future research will involve a detailed analysis of the model's errors to understand the underlying reasons for its mispredictions. This analysis will guide the development of more effective strategies for fine-tuning and enhancing the model's capabilities. Our ultimate goal is to contribute to the advancement of AI in healthcare by providing reliable and interpretable models that can support medical professionals and improve patient outcomes.

Acknowledgment

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D4-2024-01.

References

[1] Y. Lin, K. Lai, W. Chang, Skin medical image captioning using multi-label classification and siamese network, IEEE Access 11 (2023) 23447–23454. doi:10.1109/ACCESS.2023.3249462.

[2] S. Elbedwehy, T. Medhat, T. Hamza, M. Alrahmawy, Enhanced descriptive captioning model for histopathological patches, Multimedia Tools and Applications 83 (2023) 1–20. doi:10.1007/s11042-023-15884-y.

[3] A. Selivanov, O. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, D. Dylov, Medical image captioning via generative pretrained transformers, Scientific Reports 13 (2023). doi:10.1038/s41598-023-31223-5.

[4] H. Lee, H. Cho, J. Park, J. Chae, J. Kim, Cross encoder-decoder transformer with global-local visual extractor for medical image captioning, Sensors 22 (2022) 1429. doi:10.3390/s22041429.

[5] D.-R. Beddiar, M. Oussalah, T. Seppänen, Automatic captioning for medical imaging (mic): a rapid review of literature, Artif. Intell. Rev. 56 (2022) 4019–4076. URL: https://doi.org/10.1007/s10462-022-10270-w. doi:10.1007/s10462-022-10270-w.

[6] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[7] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[8] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 12888–12900. URL: https://proceedings.mlr.press/v162/li22n.html.

[9] 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, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[10] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset, Scientific Data (2024). URL: https://arxiv.org/abs/2405.10004v1. doi:10.1038/s41597-024-03496-6.

[11] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset: 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, 2018, pp. 180–189. doi:10.1007/978-3-030-01364-6_20.

[12] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[13] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[14] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[15] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704.

[16] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72.

URL: https://aclanthology.org/W05-0909.

[17] R. Vedantam, C. L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575. doi:10.1109/CVPR.2015.7299087.

[18] J. Hessel, A. Holtzman, M. Forbes, R. Bras, C. Yejin, Clipscore: A reference-free evaluation metric for image captioning, 2021, pp. 7514–7528. doi:10.18653/v1/2021.emnlp-main.595.

[19] L. Jin, G. Luo, Y. Zhou, X. Sun, G. Jiang, A. Shu, R. Ji, Refclip: A universal teacher for weakly supervised referring expression comprehension, 2023, pp. 01–10. doi:10.1109/CVPR52729.2023.00263.

[20] A. Ben Abacha, W.-w. Yim, G. Michalopoulos, T. Lin, An investigation of evaluation methods in automatic medical note generation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2575–2588. URL: https://aclanthology.org/2023.findings-acl.161. doi:10.18653/v1/2023.findings-acl.161.

[21] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.