EasyChair Preprint

№ 13501

MedBLIP: Multimodal Medical Image Captioning Using BLIP

Van Thien Phan, Khanh Trinh Nguyen,
Anh Duc Dang Quang Hoang, Tien Quan Phan and
Bao Thien Nguyen Tat

May 31, 2024

MedBLIP: Multimodal medical image captioning using BLIP*

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 participated in the second task of the ImageCLEFmedical-Caption 2024 challenge using the ROCOv2 dataset with the BLIP approach.
Methods: Our approach leveraged the BLIP architecture for multimodal medical image captioning. This architecture employs a ViT (Vision Transformer) as the image encoder and a BERT (Bidirectional Encoder Representations from Transformers) as the text model.
Results: We ranked 5th according to BERTscore and placed 3rd with ROUGE, BLEURT, and RefCLIP scores. Additionally, we achieved 2nd place for BLEU-1, METEOR, and CIDEr scores. Notably, we obtained the top position with a CLIP score of 0.827074, demonstrating the effectiveness of our approach in medical image captioning.
Conclusion: Our participation in the ImageCLEFmedical-Caption 2024 challenge demonstrated the effectiveness of the BLIP architecture for medical image captioning, achieving a high CLIP score 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 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 for the ImageCLEFmedical 2024 Caption [6] task which consists of 2 subtask: Concept Detection, Caption

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); 21722717@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)

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 BLIP architecture with a Vision Transformer (ViT) image Encoder. We employed BLIP (base/large) [7] with pretrained weights from "Salesforce/blip-image-captioning-(base/large)".

2. Task and Dataset Descriptions

2.1. Task Description

ImageCLEFmedical-Caption is one of imageCLEF-medical'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 [8].

2.2. Dataset Descriptions

The dataset for this task is ROCOv2 [8]- an extended version of ROCO[9]. 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 9972 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.

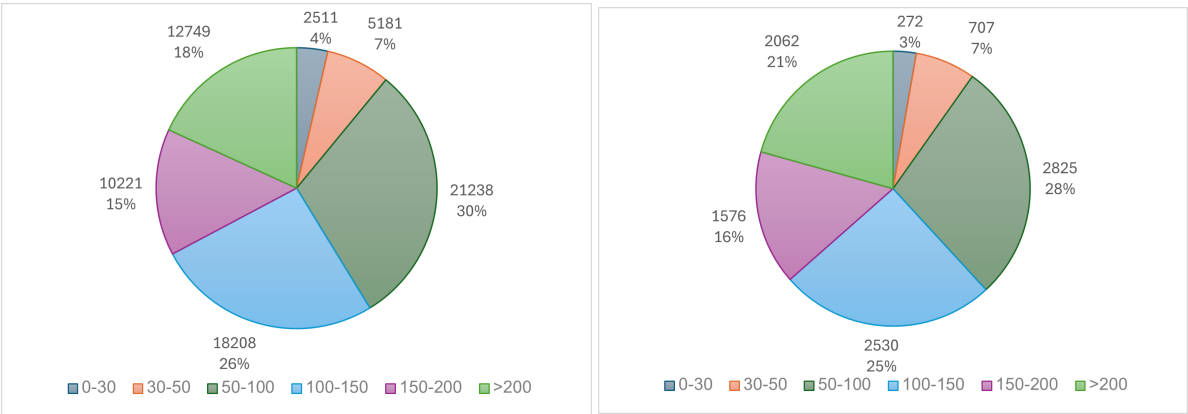


Figure 1: Distribution of caption lengths in the training set (left) and validation set (right)

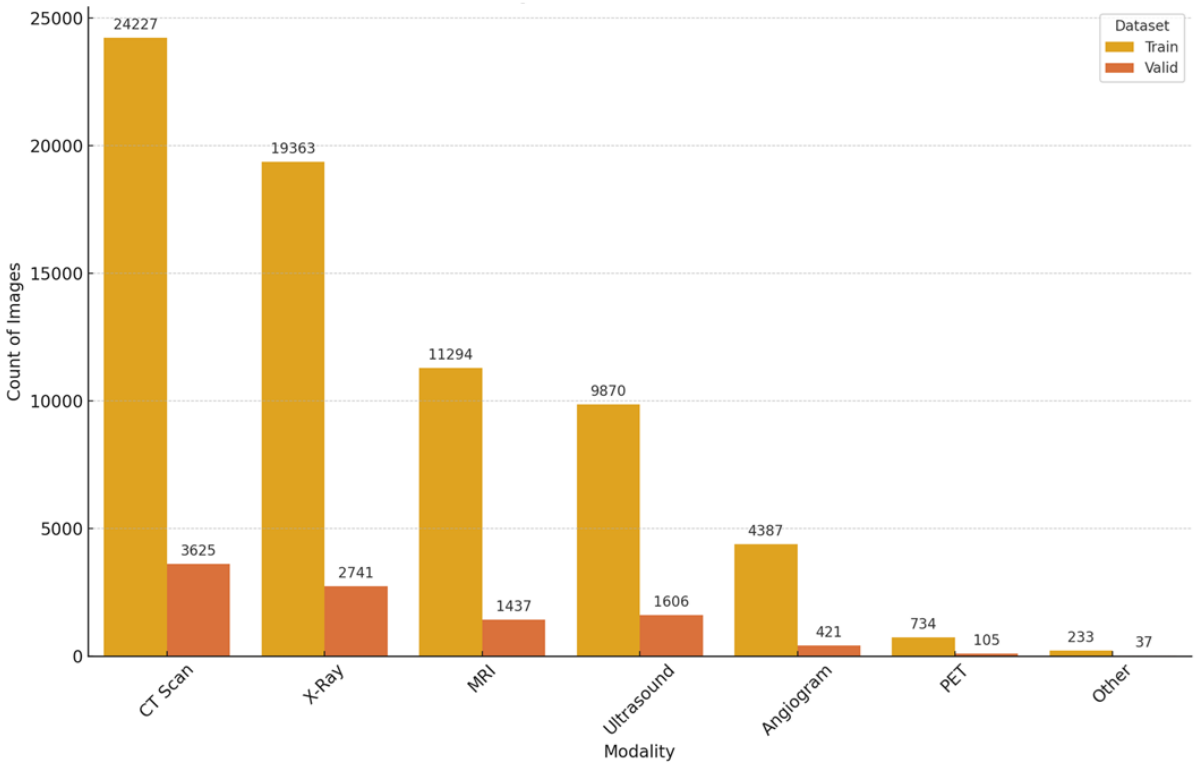


Figure 2: Distribution of image modalities in Train and Validation Sets.

3. Methods

3.1. Models

Bootstrapping Language-Image Pretraining (BLIP) [7] 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)[10] 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.

To be able to train or pretrain 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 a Unimodal Encoder, an Image-grounded Text Encoder, and an Image-grounded Text Decoder. The objectives are Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), and Language Modeling Loss (LM)

3.2. Evaluation Metrics

We employed two main metrics: BERT Score [11] and ROUGE score [12].To calculate BERTScore, we use the 'microsoft/deberta-xlarge-mnli' model, which can be found on the Hugging Face Model Hub. Additionally, other metrics such as BLEU-1 [13], BLEURT[14], METEOR[15], CIDEr[16], CLIPScore [17], RefCLIP score[18], ClinicalBLEURT score, and MedBERT score were also applied for evaluation Before evaluation, the text data underwent post-processing 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.

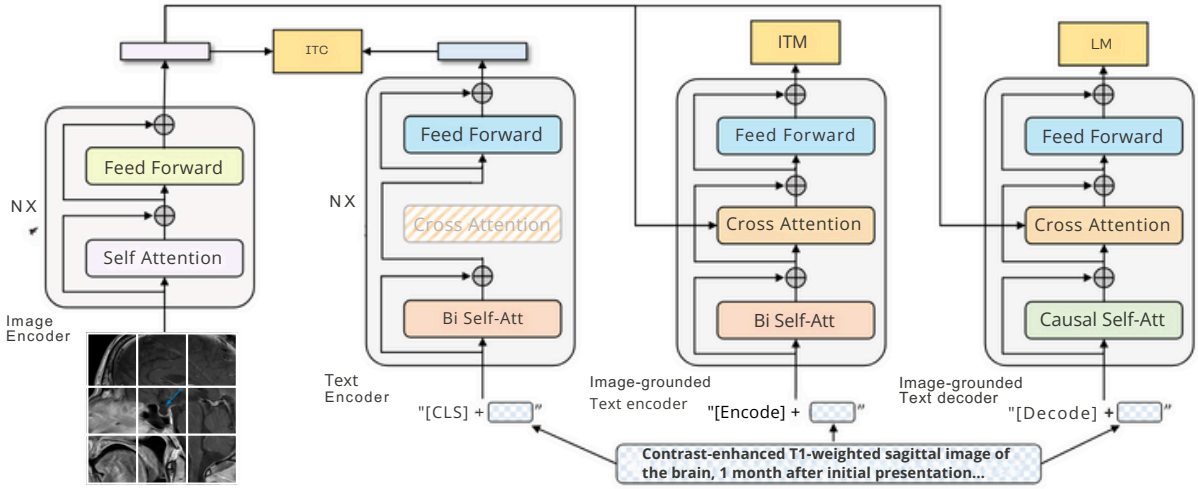


Figure 3: Pre-training model architecture and objectives of BLIP (same parameters have the same color).The multimodal mixture of encoder-decoder was proposed, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (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.

4. Experiments

4.1. Experimental Setup

In our experiments, we employed BLIP model (base/large) from pretrained checkpoints. For BLIP base, we utilized weights from checkpoint "Salesforce/blip-image-captioning-base". The training was conducted over 15 epochs with initial learning rate of 1e-5. We used a StepLR scheduler to decrease the learning rate by factor of 10 every 3 epochs. For the BLIP large model, we utilized weights from the checkpoint 'Salesforce/blip-image-captioning-large'. The 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 three experiments, we utilized the AdamW optimizer. Input image were resized to 224x224, while the maximum length of text input was set to 200 tokens. To facilitate model training, we used a single GPU A100 PCIE 40GB.

For each model, we experimented with 4 different generation settings with no_repeat_ngram_size = 3:

(1) Greedy Search
(2) Beam Search with beam_size = 3
(3) Beam Search with beam_size = 4
(4) Beam Search with beam_size = 5
(5) Beam Search with beam_size = 10

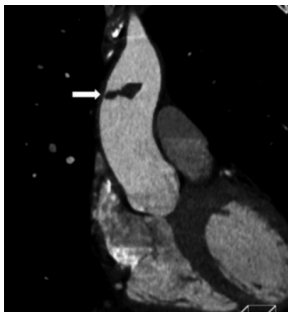
4.2. Experimental Results

4.2.1. Results on Validation Set

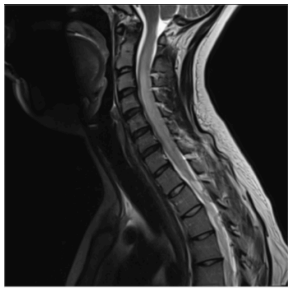
Based on the results in Table 1, the BLIP large model demonstrates better performance compared to the BLIP base model. Both models show significant generative capabilities, with beam search outperforming greedy search across ROUGE score, as well as BERTscore and BLEU score. Specifically, the BLIP base model achieves its highest BERTscore and ROUGE scores with a beam size of 5, and its best BLEU score

	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

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



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).



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.

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

Based on the Test Set results in table 2 announced by the organizing committee, our team ranked 5th according to BERTscore; With ROUGE score, BLEURT score, RefCLIP score we ranked 3rd. With BLEU-1, METOER and CIDEr score we achieved 2nd place. With CLIP score, we get the 1st result with 0.827074. These results demonstrate the model's expected performance. However, the model we tested still has a lot of room for further improvements, especially to optimize BERTscore.

Table 2
Results on Test Set

Team	ID	BERTScore	ROUGE	BLEU-1	BLEURT	METEOR
pclmed	634	0.629913	0.272626	0.268994	0.337626	0.113264
CS_Morgan	429	0.628059	0.250801	0.209298	0.317385	0.092682
DarkCow	220	0.626720	0.245228	0.195044	0.306005	0.088897
auebnlpgroup	630	0.621112	0.204883	0.111034	0.289907	0.068022
2Q2T	**643**	**0.617814**	**0.247755**	**0.221252**	**0.313942**	**0.098590**
MICLab	678	0.612850	0.213525	0.185269	0.306743	0.077181
DLNU_CCSE	674	0.606578	0.217857	0.151179	0.283133	0.070419
Kaprov	559	0.596362	0.190497	0.169726	0.295109	0.060896
DS@BioMed	571	0.579438	0.103095	0.012144	0.220211	0.035335
DBS-HHU	637	0.576891	0.153103	0.149275	0.270965	0.055929
KDE-medical-caption	557	0.567329	0.132496	0.106025	0.256576	0.038628

Team	ID	CIDEr	CLIPScore	RefCLIPScore	ClinicalBLEURT	MedBERT
pclmed	634	0.268133	0.823614	0.817610	0.466557	0.632318
CS_Morgan	429	0.245029	0.821262	0.815534	0.455942	0.632664
DarkCow	220	0.224250	0.818440	0.811700	0.456199	0.629189
auebnlpgroup	630	0.176923	0.804067	0.798684	0.486560	0.626134
2Q2T	**643**	**0.220037**	**0.827074**	**0.813756**	**0.475908**	**0.622447**
MICLab	678	0.158239	0.815925	0.804924	0.445257	0.617195
DLNU_CCSE	674	0.168765	0.796707	0.790424	0.475625	0.612954
Kaprov	559	0.107017	0.792183	0.787201	0.439971	0.608924
DS@BioMed	571	0.071529	0.775566	0.774823	0.529529	0.580388
DBS-HHU	637	0.064361	0.784199	0.774985	0.476634	0.582744
KDE-medical-caption	557	0.038404	0.765059	0.760958	0.502234	0.569659

5. Conclusion and Future work

In this paper, we implemented and experimented the BLIP model for the task of medical image captioning in the imageCLEFmedical-Caption 2024 challenge. Experimental results across various configurations showed promising outcomes. Specifically, the model achieved a CLIP score of 0.82707 on the test set of the ROCOv2 dataset. However, there is still significant room for improvement in our research. The primary weakness of the model is that it was pre-trained on a dataset quite different from the medical domain, resulting in considerable bias.

In the future research, we will focus on improving the model's accuracy by utilizing pretrained models with datasets that are more closely aligned with medical and diagnostic domains, as well as applying preprocessing methods tailored to different types of images.

Acknowledgment

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

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. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, D. V. Dylov, Medical image captioning via generative pretrained transformers, 2022. arXiv:2209.13983.

[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] 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.

[7] 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. doi:10.48550/ARXIV.2201.12086.

[8] 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, 2024. URL: https://arxiv.org/abs/2405.10004v1. arXiv:2405.10004.

[9] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. Friedrich, Radiology objects in context (roco): A multimodal image dataset, in: CVII-STENT/LABELS@MICCAI, 2018. URL: https://api.semanticscholar.org/CorpusID:53087891.

[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, CoRR abs/2010.11929 (2020). URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, CoRR abs/1904.09675 (2019). URL: http://arxiv.org/abs/1904.09675. arXiv:1904.09675.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] R. Vedantam, C. L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, CoRR abs/1411.5726 (2014). URL: http://arxiv.org/abs/1411.5726. arXiv:1411.5726.

[17] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, 2022. arXiv:2104.08718.

[18] 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.