KU-DMIS-MSRA at RadSum23: Pre-trained Vision-Language Model for Radiology Report Summarization

Gangwoo Kim[1◇] **Hajung Kim**[1] **Lei Ji**[2] **Seongsu Bae**[3] **Chanhwi Kim**[1] **Mujeen Sung**[1]
Hyunjae Kim[1] **Kun Yan**[4] **Eric Chang**[5] **Jaewoo Kang**[1,6†]
[1]Korea University* [2]Microsoft Research Asia [3]KAIST AI [4]Beihang University
[5]Centre for Perceptual and Interactive Intelligence, Hong Kong [6]AIGEN Sciences Inc.
{gangwoo_kim, kangj}@korea.ac.kr

Abstract

In this paper, we introduce CheXOFA, a new pre-trained vision-language model (VLM) for the chest X-ray domain. Our model is initially pre-trained on various multimodal datasets within the general domain before being transferred to the chest X-ray domain. Following a prominent VLM, we unify various domain-specific tasks into a simple sequence-to-sequence schema. It enables the model to effectively learn the required knowledge and skills from limited resources in the domain. Demonstrating superior performance on the benchmark datasets provided by the BioNLP shared task (Delbrouck et al., 2023), our model benefits from its training across multiple tasks and domains. With subtle techniques including ensemble and factual calibration, our system achieves first place on the RadSum23 leaderboard for the hidden test set.

1 Introduction

Chest radiography is a widely used imaging modality for assessing the thorax and diagnosing cardiopulmonary conditions. However, there is a significant shortage of clinical doctors in several under-resourced regions, delaying diagnosis and treatment and reducing the quality of care. Developing an automated system for analyzing radiographs can improve radiologist workflow efficiency and expand healthcare services to these regions.

To promote research in this direction, the BioNLP 2023 workshop opens a new shared task called RadSum23 (Delbrouck et al., 2023) for the radiology report summarization.[1] In this task, participants are asked to build a model that takes a *findings* section as input, which is a generic radiology report of a given X-ray image, and then outputs an *impression* section, which is a summary of key observations in the given report. High-resolution X-ray images can be used as input along with findings sections. The radiology report summarization task aims to effectively distill complex clinical observations from chest X-ray images into concise and coherent summaries.

In this paper, we introduce CheXOFA (One For All tasks with Chest X-ray), a novel vision-language model designed for the chest X-ray domain. Specifically, we initialize our model using OFA (Wang et al., 2022), a Transformer model pre-trained with a unified sequence-to-sequence schema on diverse uni- or cross-modal tasks such as image classification, language modeling, and image captioning. We further train the model to generate full-text reports from the chest X-ray image using the MIMIC-CXR dataset (Johnson et al., 2019). Then, we fine-tune the model on the radiology report summarization task. When summarizing the report, our model jointly encodes visual information from the chest X-ray image with the full-text report, taking advantage of its multimodal nature. Additionally, we employ subtle techniques such as task-specific ensemble (Dai et al., 2021) and factual calibration to further improve the model performance.

Our experiments demonstrate that our proposed methods largely enhance model performances on two test sets of the shared task. On the official leaderboard [2] for MIMIC-CXR hidden test set, our system ranked first place, achieving the state-of-the-art performances in most evaluation metrics. Especially, it surpasses the second-best model by 2.3 and 2.9 in BLEU and F1-CheXbert score. In the ablation study, we demonstrate how much each method contributes to the improvement.

◇ Work done while interning at Microsoft Research Asia
† Corresponding author
* Department of Computer Science and Engineering
[1]vilmedic.app/misc/bionlp23/sharedtask
[2]vilmedic.app/misc/bionlp23/leaderboard/

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 567–573
July 13, 2023

2 Related Works

2.1 Automated Radiology Report Generation

Utilizing radiographic datasets such as MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019), various methodologies for automated report generation have recently attracted attention. Notably, these datasets often include both chest X-ray images and free-text reports, enabling the use of automated rule-based labelers (Irvin et al., 2019) or neural models (Smit et al., 2020) to extract disease labels from the reports. It can be categorized into two tasks: 1) radiology report generation, which is similar to medical image captioning and aims to describe radiology images in detail (*findings* section), having seen significant progress in recent years (Chen et al., 2020; Zhang et al., 2020b; Liu et al., 2021a,b; Miura et al., 2021; Chen et al., 2022); 2) radiology report summarization (Zhang et al., 2018), which focuses on summarizing *findings* section into *impressions* section in radiology reports. Most existing research (Zhang et al., 2020c; Hu et al., 2021, 2022b; Karn et al., 2022) tends to focus on text-based summarization while some image-incorporating studies (Delbrouck et al., 2021; Hu et al., 2022a) use suboptimal methods lacking appropriate multi-modal pre-training objectives for the generative task.

2.2 Multimodal Foundation Models

Vision-language pretraining models are becoming the foundation models effective for multimodal tasks including Vilbert (Lu et al., 2019), OFA (Wang et al., 2022), Flamingo (Alayrac et al., 2022) for open-domain, and MedViLL (Moon et al., 2022), Clinical-BERT (Yan and Pei, 2022), BioViL (Boecking et al., 2022) and M3AE (Chen et al., 2022) for biomedical domain. However, most of the existing works for clinical domain mainly focus on pretraining the encoder for understanding tasks like medical VQA, classification and so on, and rarely take this radiology report summarization task as downstream task. In this paper, we propose CheXOFA, a pre-trained generative VLM that learns the required capabilities for the radiology report summarization task.

3 CheXOFA

We propose a novel vision-language model, CheXOFA, specifically designed for the radiology report summarization. We first initialize our model parameters with OFA (Wang et al., 2022), which has been shown to be effective in the general domain. Then, we pre-train and fine-tune the model with the various tasks in the medical imaging domain. In addition, we newly introduce a factual calibration technique to further improve the model performance.

3.1 Multimodal Architecture

The backbone of the CheXOFA model is Transformer (Vaswani et al., 2017) architecture with a sequence-to-sequence framework specialized for generative tasks. Following BART (Lewis et al., 2020) and GPT (Radford et al.), we utilize byte-pair encoding (BPE) (Sennrich et al., 2016) to transform text sequences into linguistic features of subword sequences. On the vision side, we use a visual extractor to encode an image into a sequence of hidden representations. Specifically, we divide an input image into fixed size of patches. Then, ResNet (He et al., 2016) modules are used to convolve the visual information into visual features $x^v \in \mathcal{R}^{|P| \times d}$, where $|P|$ is the number of patches and d is the dimension of hidden representation. Overall, the linguistic features and the visual features are concatenated into one sequence for feeding into the encoder-decoder Transformer for modality fusion and sequence generation.

3.2 Training and Inference

Our model is optimized with the cross-entropy loss to the ground-truth sequences. Given an input x and an output y, we train model parameters θ by minimizing $\mathcal{L} = -\sum_{t=1}^{|y|} \log P_\theta(y_t|y_{<t}, x)$, where x can be composed of instruction x^i, linguistic and visual features x^l, x^v. In the inference phase, we choose the beam search as the decoding strategy to obtain better text sequences. Additionally, we adjust the length penalty that assigns weights according to the length of each beam. We control the output sequence length and explore the optimal value for it.

3.3 Pre-training and Downstream Tasks

Following Wang et al. (2022), we unify cross-modal tasks into a simple sequence-to-sequence format. We design a unified learning framework for pre-training and downstream tasks, which require multimodal reasoning ability. CheXOFA's versatile design allows it to tackle a wide range of tasks using a single model. Furthermore, the model is designed with multitasking capabilities, allowing it to simultaneously handle multiple tasks across various modalities. To this end, the model shares its parameters and schema across all tasks. Meanwhile, we employ task-specific instructions manually crafted for each task.

We pre-train CheXOFA with the report generation (RGen) task and then fine-tune it with report summarization (RSum) on MIMIC-CXR dataset (Johnson et al., 2019). In the RGen task, the model learns to generate *findings* section of the report, based on the chest X-ray image. We use the same instruction x^i with that of the image captioning task, "What does the image describe?". For our target task, RSum, the model is trained to generate *impression* section, given *findings* section of the report. We also exploit the corresponding chest X-ray image to jointly leverage the visual information. Hence, an input x is composed of visual features, subword tokens of *findings* section, and instruction, "what is the summary of the following article?". Furthermore, we newly design the classification-supported RSum task (cls-RSum) to enhance the factual correctness of the summary. In the task, the model additionally performs a classification task for observed disease from the X-ray image or report. Then, it generates summaries based on the identified category, ensuring relevance and coherence.

3.4 Ensemble with Factual Calibration

To improve the overall performance, we utilize an ensemble method that combines various predictions from multiple models. Following Dai et al. (2021), we select the best prediction based on the mutual similarity score. In particular, we calculate similarity scores for every possible pair of predictions, creating a mutual similarity matrix. Subsequently, we aggregate the matrix in a row-wise fashion, averaging value for each row. The prediction with the highest score is selected as the final output. If multiple outputs have the same highest score, we randomly select the final output, ensuring that it is chosen in an unbiased manner. To compute the similarity, we use F1-RadGraph as a scoring function. Through the combination of diverse predictions, we are able to obtain an optimal summary and mitigate the failure of individual models.

We perform a calibration to improve the factual correctness of the ensembled prediction. We first extract medical observations from the prediction by using cheXbert labeler (Smit et al., 2020). Then, we check whether they are matched with identified labels by the cls-RSum model. If not, the cls-RSum result is chosen as the final prediction. We find the factual calibration performs effectively when the ensemble process yields a low aggregated similarity score, which is in proportion to the model's confidence.

4 Experiment

4.1 Experimental Setup

Evaluation Metric For evaluation, we used five metrics: BLEU4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), Bertscore (Zhang et al., 2020a), F1-cheXbert (Zhang et al., 2020d), and F1-RadGraph (Delbrouck et al., 2022a). BLUE4 and ROUGE-L measure syntactical similarity based on n-gram overlap between reference and generated summaries, while Bertscore measures semantic similarity to handle synonyms and paraphrasing. To evaluate the factual correctness of generated summaries, F1-cheXbert and F1-RadGraph are used. Out of these five metrics, F1-RadGraph is selected as the primary metric for ranking participating systems, as it uses RadGraph (Jain et al., 2021) annotations to better consider the consistency and correctness of extracted entities and relations.

Benchmark Datasets Our model was trained on MIMIC-CXR (Johnson et al., 2019) and evaluated on both the test set of MIMIC-CXR and a newly-collected hidden test set. MIMIC-CXR is a publicly available large dataset consisting of 128,032 report-image pairs with 227,835 multi-view images[3]. The dataset is split into training, validation, and test sets, which comprise 125,417, 991, and 1,624 image-report pairs, respectively. The MIMIC-CXR hidden test set was newly introduced in the RadSum23 challenge served by vilmedic (Delbrouck et al., 2022b), and it comprises 1,000 out-of-domain image-report pairs collected from CheXpert images (Irvin et al., 2019)[4].

4.2 Results

Table 1 shows the official results of the leaderboard on the hidden test set. Our model ranked first place among other systems on the leaderboard and achieved the state-of-the-art performance in

[3]The mentioned statistics are derived from reports containing both the *findings* and *impression* sections simultaneously.

[4]Although it comes from CheXpert dataset, we name it as MIMIC-CXR hidden test set, following the challenge description

Track	Rank	Team Name	BLEU	ROUGE	Bertscore	CheXbert F1	Radgraph F1
Hidden Test	1	**ku-dmis-msra (ours)**	**18.62**	34.57	**55.90**	**72.36**	**43.20**
	2	utsa-nlp	16.33	**34.97**	55.54	69.41	42.86
	3	knowlab	14.41	33.63	54.72	67.20	39.98
	4	shs-te-dti-mai	14.59	32.43	53.99	6 8.99	38.40
	5	aimi	5.15	31.84	47.83	64.18	32.05
Public Test	1	utsa-nlp	**25.87**	**47.86**	64.74	**77.93**	**51.84**
	2	**ku-dmis-msra (ours)**	25.58	47.75	**64.80**	76.29	50.96
	3	knowlab	22.97	46.15	63.43	75.14	48.04
	4	e-health csiro	17.97	44.14	61.47	71.67	44.95
	5	iuteam1	10.10	40.44	56.44	58.01	39.48

Table 1: Official results of the leaderboards on MIMIC-CXR hidden test set and MIMIC-CXR test set. The Models are ranked based on F1-Radgraph score. The best score is displayed in bold typeface and the score of the second best model is underlined.

Model	ROUGE	CheXbert F1	Radgraph F1
CheXOFA (Ensem.)	47.75	76.29	50.96
w/o. Fact Calib.	47.04	75.98	50.14
CheXOFA (Single)	46.03	75.67	48.16
Text-only	46.30	73.59	47.78
w/o. pre-training	43.79	73.67	45.73

Table 2: Ablation study on the public test set. We provide the performance of ensembled (Ensem.) and single results. Every component such as factual calibration (Fact Calib.), encoding multimodal inputs (Text-only), and pre-training contributes to the improvement.

four out of five evaluation metrics. Especially, our model significantly outperformed the second-best model by 2.29 BLEU4 and 2.95 F1-cheXbert.

Table 1 also presents the official results on the public test set of MIMIC-CXR. Our model achieved competitive performance with the best model in four out of five metrics and the best performance based on Bertscore. Overall results indicate that our method could generalize to diverse datasets, achieving outstanding performances on both hidden and visible test sets. Conclusively, our method has remarkable capabilities to summarize radiology reports by capturing essential medical observations.

4.3 Ablation Study

We performed an ablation study to analyze how each method contributes to the overall performance. Table 2 shows the evaluation scores when removing each method from our best model on the test set of MIMIC-CXR. Factual calibration improves the factual correctness scores, 0.3 of the F1-CheXbert score and 0.8 of F1-Radgraph score. Using a single CheXOFA model shows a performance drop compared to the ensemble model by approximately 1.8 in F1-Radgraph. Nevertheless, it achieves competitive performances with other participating systems. Allowing the model to focus on the report (Text-only) achieves similar performance in ROUGE-L and F1-Radgraph scores relevant to the lexical overlap. However, F1-cheXbert score significantly degrades, which indicates that models benefit from using multimodal information. Fine-tuning a vanilla OFA model performs poorly in most scores, which shows the importance of the pre-training task.

5 Conclusion

We proposed a pre-trained VLM, CheXOFA, for the chest X-ray domain. We showed that pre-traing with the report generation task improves the downstream task, report summarization. Taking advantage of its multimodal nature, we improved the performance by jointly encoding visual and linguistic features. Furthermore, we explored subtle techniques such as ensemble and factual calibration to improve the model performance. Our experimental results demonstrated that proposed methods benefit the summarization performance. Our model ranked first on the hidden test set in RadSum23 shared task. We showed promising results about the domain-specific VLM in the chest X-ray tasks. We hope that our method can shed light on automating radiology report generation.

Acknowledgement

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program) (RS-2022-00155958) and the ICT Creative Consilience program(IITP-2023-2020-0-01819) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation). It was also supported by the National Research Foundation of Korea, South Korea (NRF-2023R1A2C3004176) and a grant of the Korea Health Technology RD Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health Welfare, Republic of Korea (grant number: HR20C0021).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022*, pages 1–21, Cham. Springer Nature Switzerland.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 679–689. Springer.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. BDKG at MEDIQA 2021: System report for the radiology report summarization task. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 103–111, Online. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34, Dublin, Ireland. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Cassie Zhang, and Daniel Rubin. 2021. Qiai at mediqa 2021: Multimodal radiology report summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 285–290.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jinpeng Hu, Zhihong Chen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2022a. Improving radiology summarization with radiograph and anatomy prompts. *arXiv preprint arXiv:2210.08303*.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4980–4990.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022b. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, D. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curt P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *ArXiv*, abs/2106.14463.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng.

2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schütze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021a. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020b. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. 2020c. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020d. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of* *the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.