Chest X-Ray Radiology Reports Generation with Pre-trained CLIP Model

Jungmin Park[1] **Dami Jung**[2] **Chanhyeok Choi**[3]
Department of Computer Science Engineering
Ulsan National Institute of Science and Technology
{jm971004[1], dam2j[2], chan4184[3]}@unist.ac.kr

Abstract

The automated generation of clinically accurate radiology reports can provide highly interpretable information for doctors, and it has the potential to improve patient care and reduce radiologist workload. In this work, we will focus on training a model using a contrastive language-image pre-training model for better performance. We aims to train the model to generate better report and more efficient in training speed than previous models, and propose an idea for incorporating text simplification to generate easy-to-understand sentences from the generated reports for the public.

1 Introduction

Automated radiology report generation can potentially improve radiology reporting and alleviate the workload of radiologists. Research on a model that prepares a radiation report by training chest X-ray images is in progress, such as using weakly supervised contrastive loss for medical report generation (Yan et al., 2021 [1]) or adapting CLIP-based models to the the chest x-ray classification and generating images from text (Wiehe et al., 2022 [2]). However, report generated from previous models still not guarantee good predictions for rare pathology and not very readable for general public. In this project, we focus on developing our own model using nontrivial methods and improve a prior method in order for automated report generation deployment to advance patient care. Our model should generate a radiology report given a chest X-ray as shown in Fig 1. In addition, we're proposing an idea as future work for text simplification on generated report to re-generate simple and easy-to-understand contents for laypeople.

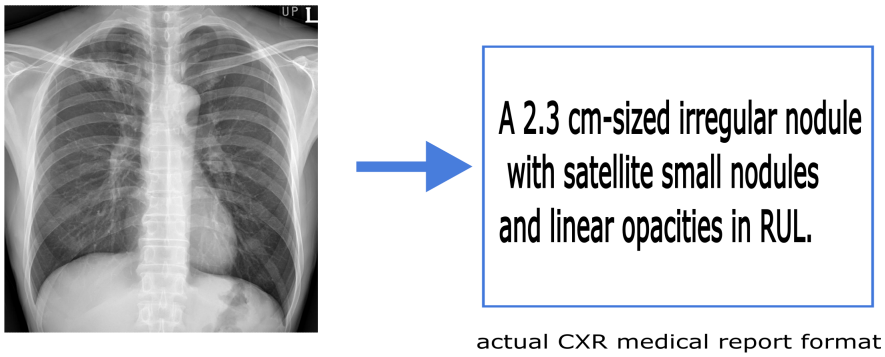


Figure 1: Example of Report Generation

Preprint. Under review.

2 Related Work

Our key research paper in this work was Chest X-Ray Report Generation from Chest-X Ray Images, proposed in Stanford CS224N Custom Project (Haritaoglu et al., 2022 [3]). Many methods have been proposed to solve the task of medical report generation. There are three the most common approaches for automatic chest x-ray report generation:

- Template-based approach, which classifies whether certain conditions are present or not, and then retrieves the corresponding template sentence for each condition.
- Retrieval-based approach, which reuses text from previous similar examples.
- Generation-based approach, which uses an image-encoder + text-decoder architecture that encodes the image and then decodes that into a sequence of words.

This paper([1]) adopted generation-based approaches which is represented by the model with encoder - decoder architecture. In detail, it followed the pretrained UNet image encoder, and used Cross Entropy Loss at the output of the text decoder. Also, it experimented with Contrastive Loss by using the labels from CheXbert Pathology Labeler to contrast the embeddings at the transformer decoder output.

There are five well-known datasets that can be used for Chest X-Ray report generation training and evaluation: IU X-RAY, PEIR GROSS, ICLEF CAPTION, MIMIC-CXR, and CheXpert. In practice, the Stanford team only used two datasets, IU X-RAY and CheXpert, due to various conditions. Then, they evaluated their model using conventional natural language generation (NLG) metrics and clinical efficacy (CE) metrics on both the internal dataset.

Previous method has shown promising performance in metrics that measure descriptive accuracy, but it can fail to produce complete, consistent, and clinically accurate reports. In particular, the model cannot be expected to make predictions for rare pathologies if they do not appear in the reference corpus. The paper([1]) also mentions experiment with lateral images from IU X-ray dataset as future work to get higher accuracy on generated medical reports. Therefore, we propose a new approach that addresses the shortcomings of the existing architecture. We will basically use a generation-based approach, but experiment a more advanced pre-trained UNet image encoder.

And, earlier works in radiology report generation have focused heavily on performing image captioning with a transformer architecture. However, one main shortcoming of the image-captioning models is the presence of medically inconsistent information frequently found in the generated reports. To address this issue, CXR-RePaiR (Endo et al., 2021[6]) adopts an image-text retrieval method that retrieves a report whose CLIP (Radford et al., 2021[14]) text embedding scores the highest cosine similarity with the chest X-ray's CLIP image embedding. To further improve performance, we reconfigure the Contrastive Loss using the CLIP model pretrained with MIMIC-CXR dataset instead of the CheXpert Pathology labeler.

3 Approach

3.1 Model architecture

Our work follows the [3] with making differences in the image encoder, text decoder architecture and loss function. We use EfficientUNet pre-trained on ImageNet dataset as an image encoder. The proposed model has outperformed the other state-of-the-art models for lung segmentation in chest x-rays.

As shown in Figure 2, we are planning to experiment with concatenating labels from chest X-Ray images with the embeddings from EfficientUNet. We also experiment with Cross Entropy Loss by using target reports embeddings from IU X-ray to contrast the embeddings at the transformer decoder output. The predicted report should be computed with CLIP embeddings between image x and report r. The CLIP model passes r (or s) through a text encoder E_T to produce a text embedding T. Similarly, it passes x through an image encoder E_I to produce an image embedding I. These embeddings are utilized to compute a cosine similarity of image-text direction for source and target, which is described in more detail in section 3.1.2. We use a pre-trained CLIP model on radiology report-image pairs.

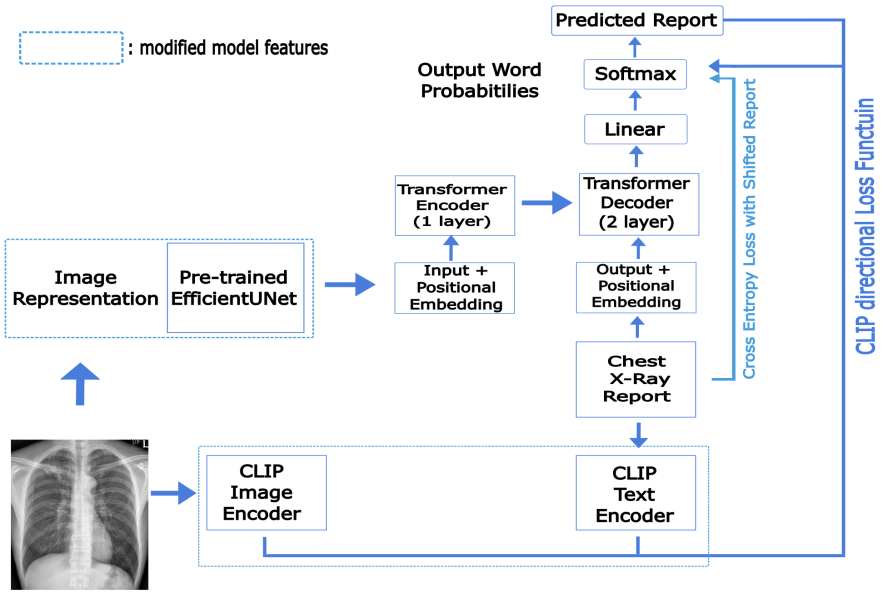


Figure 2: Generation Based Image Encoder/ Text Decode

3.1.1 Replacement with pretrained EfficientUNet encoder and Quantitative evaluation with MIMIC-CXR dataset

Our baseline used UNet pretrained on brain MRI images as an image encoder. UNet is the most popular segmentation architecture among the deep learning community. Despite its high performance, there is scope for improvement. So, in our model architecture, modified pre-trained EfficientNet-B4 will be used to experiment an effect as encoder (Agrawal et al., 2022 [7]).

3.1.2 Improved performance with the new directional Loss function with CLIP model

In Endo et al., 2021 [6], the baseline approach of a contrastive language image model (CLIP) trained on radiology report-image pairs showed good performance across overall evaluations. However, it is only applied to retrieval-based generation method. We propose to apply CLIP directional loss function with pre-trained CLIP model to transformer decoder for clinically accurate radiology reports. CLIP directional loss is proposed in StyleGAN-NADA[13] to steer generative models towards a user-defined text caption since CLIP provides a score of how close an image is to a caption. An illustration of the CLIP-space directions is provided in Fig. 3. And training scheme by the directions is shown in Fig. 4. We expect our model to learn that a directional vector of ΔD_1 aligns to one of ΔD_2. The directional loss is given by:

$$\Delta D_1 = E_T(r) - E_I(x) \tag{1}$$

$$\Delta D_2 = E_T(M(x)) - E_I(x) \tag{2}$$

$$L_{directional} = 1 - \frac{\Delta D_1 \cdot \Delta D_2}{|\Delta D_1| \cdot |\Delta D_2|} \tag{3}$$

Here, E_T and E_I are pretrained CLIP's text and image encoder, and $M(x)$ is report texts predicted by the model where M is our model. r and x are ground truth report texts and target chest X ray image, respectively.

Then, a total loss L_{total} can be composed as following:

$$L_{total} = L_{CE} + L_{directional} \tag{4}$$

where L_{CE} is a Cross Entropy Loss with shifted report deriven from our baseline and $L_{directional}$ is our CLIP directional loss.

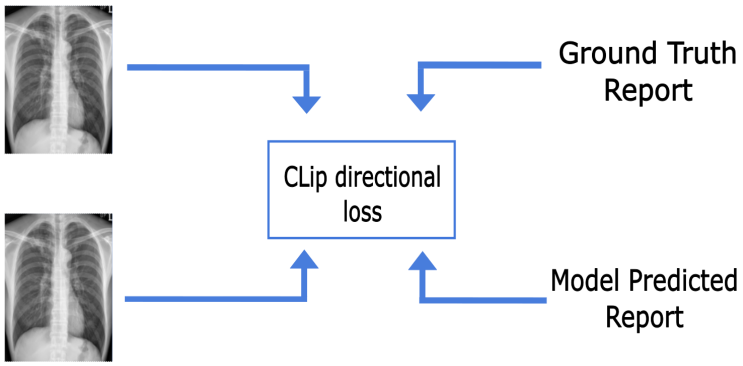


Figure 3: CLIP directional Loss

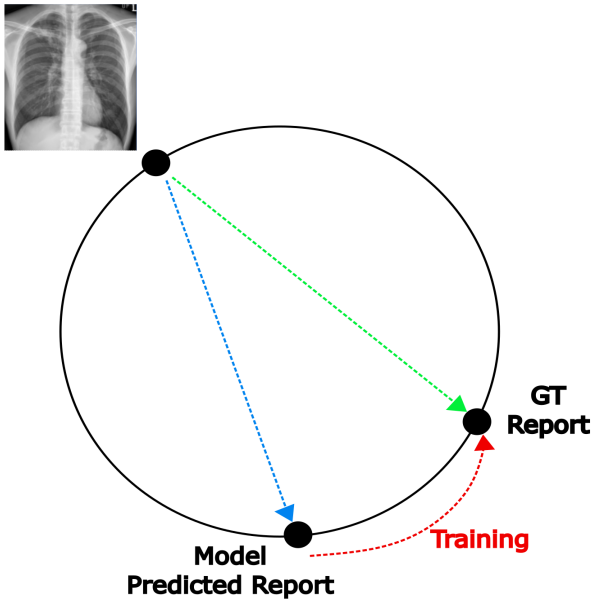


Figure 4: CLIP latent space for ΔD_1 and ΔD_2

4 Experiments

4.1 Data

In this work, we are planning to use two dataset, MIMIC-CXR(Johnson et al., 2019[4],[5]) and IUXray.

- MIMIC-CXR: MIMIC-CXR is a huge dataset containing 377,110 chest X-ray images and 227,835 free-text format reports. It is outrageous data, more than our computer's hard disk or Google Drive could handle. So, we decided to use a CLIP model, pretrained with MIMIC-CXR dataset. [1]
- IU X-ray: IU X-ray is an open access chest X ray dataset from Indiana University. We used the dataset that includes frontal and lateral images alongside with their reports. After filtering out lateral images, images with empty reports and duplicate reports, we got a dataset size of 2259 samples for training, and 640 samples for validation. We created a vocabulary from the findings section of the reports and had a vocabulary size of 2321 with cased words.

[1] https://physionet.org/content/mimic-cxr/2.0.0/

4.2 Evaluation method

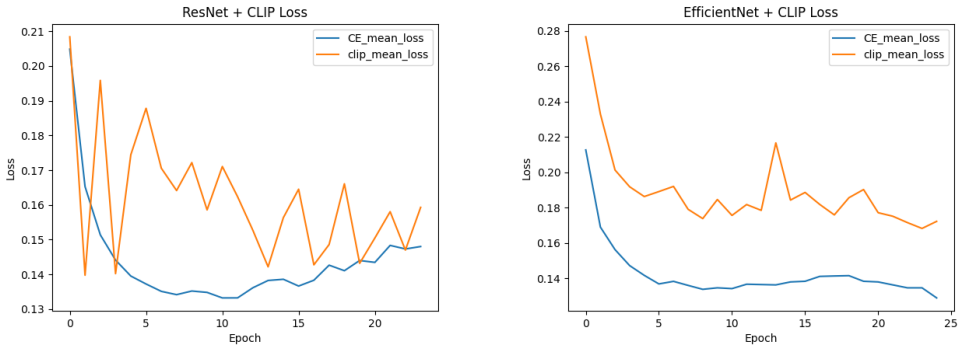
We focus on various metrics to evaluate the clinical efficacy (CE) of the generated reports. Then, we evaluate performance by three automatic metrics: BLEU (Papineni et al., 2002 [9]), ROUGE-L (Lin, 2004 [10]), and METEOR (Denkowski and Lavie, 2011 [11]). These metrics are computed as a natural language generation (NLG) metric.

4.3 Experimental details

During training we used batch size of 16 and learning rate of 2e-3 with a cosine learning rate decay. We ran training until after the test loss has gone above the 5 of the minimum test loss. Test loss generally reached to its lowest around epoch 4/5 and continued to rise above 5 until epoch 8/9. Each epoch took approximately 8 minutes with data running parallel on GPU A100 in Google Colab Pro Plus. Running each epoch took relatively long time, which is because the model generated a batch-sized report for every epoch, and batch-sized images and report texts are also encoded by CLIP.

4.4 Results

Fig 5 shows 2 losses(CE mean loss and CLIP mean loss) in ResNet and EfficientUNet with CLIP Loss model. In (a), the case of CE mean loss decreases very quickly from the early stage of learning, then takes an optimal learning point near epoch 10, and after that point, the loss tends to increase slightly. In the case of CLIP mean loss, the loss changes unstable before epoch 5, and then becoming stable with decreasing trend of loss. In (b), CLIP mean loss is decreased stable than (a) and CE mean loss continued to decreased without optimal point.



(a) (b)

Figure 5: Loss of ResNet and EfficientUNet with CLIP loss

5 Analysis

We compared four x-ray report generating models(two encoder, Efficient UNet and ResNet with/without CLIP Loss)in same environment.

In terms of efficacy, the reports predicted by the two models with EfficientUNet for X-ray photographs showed significant differences(see Table 1). Without CLIP loss, report efficacy of only EfficientUNet model was poor. In 500 epochs medical report-like sentences could not generated. In contrast, with CLIP loss, key symptoms and medical terms of given X-ray, 'cardiomedial', 'cardiomegaly', 'no effusion', no edema' appeared in epochs of 27. This shows model with CLIP loss has much higher efficacy than model without CLIP loss. In two models with ResNet has similar results(see Table 2). 'normal heart size' was appeared in CLIP applied model as actual report while without CLIP model does not.

In terms of learning speed, 500 epochs of without CLIP loss and 27 epochs with CLIP loss both took around 4 hours. The learning time for each epoch takes longer for models with CLIP loss, but when

actual report	Apical lordotic frontal view. Considering differences in technical factors XXXX stable cardiomediastinal silhouette with mild cardiomegaly. No focal alveolar consolidation, no definite pleural effusion seen. Dense left lower lung nodule suggests a previous granulomatous process. No typical findings of pulmonary edema.
predicted report by **without CLIP loss**	and of of of of..of..of...
predicted report by **with CLIP loss**	borderline cardiomediastinal cardiomegaly without.... edema bilateral bilateral with without effusion

Table 1: Result of Report Generation by EfficientUNet

actual report	The heart is normal in size and contour. There is no mediastinal widening. No focal airspace disease. No large pleural effusion or pneumothorax. The XXXX are intact.
Res w/o CLIP	The and are are are . . are . in . . . are
Res w CLIP	silhouette size and size cardiomediastinal normal . . . No No . No .

Table 2: Result of Report Generation by ResNet

comparing the results of learning over the total learning time, model with CLIP loss shows faster learning.

In terms of BLEU-n score, the two models shows performance differences(see Table 3). Since a medical report is usually made up of several words, it is important to show a greater performance difference in a higher n-value in BLEU-n. EfficientUNet models learned with CLIP loss as a baseline of BLEU-1 showed 16.5 percent higher scores, and this difference gradually increased to eight times higher scores in BLEU-4. Similar patterns has shown in ResNet model except BLEU-1 case. Our CLIP Loss increased the scores, which showed a probability of improving a model. ResNet models got slightly higher scores than EfficientUNet models.

BLEU-n	1	2	3	4
Eff w/o CLIP	0.1948	0.0074	0.0005	0.00005
Eff w CLIP	0.2270	0.0110	0.0021	0.0004
Res w/o CLIP	0.2276	0.0133	0.0009	0.00005
Res w CLIP	0.2180	0.0226	0.0030	0.0005

Table 3: BLEU-n score

In terms of Rouge-L score, the model trained with CLIP loss got increased scores except Rouge-L:R case(see Table 4). P(precision) is related to correct ratio and R(recall) is related to symptom finding ratio of model prediction. X-ray is early stage of medical diagnosis so we focus more on recall scores. From this perspective, EfficientUNet with CLIP loss was best and ResNet without CLIP loss was the second best. And same as F(f-measure, harmonic mean).

In conjunction with the results in BLEU-n, it can be confirmed that the application of CLIP loss can expect many performance improvements in the EfficientUNet encoder model, but not always in the ResNet encoder model.

Rouge-L	P	R	F
Eff w/o CLIP	0.2299	0.1912	0.2088
Eff w CLIP	0.2546	0.2165	0.2340
Res w/o CLIP	0.2496	0.2119	0.2292
Res w CLIP	0.2624	0.2084	0.2323

Table 4: Rouge-L score

In terms of METEOR score, ResNet models were better than EfficientUNet models(see Table 5).

A common feature that can be found when looking at the three scores is that the application of CLIP loss in models using EfficientUNet rather than ResNet shows a greater improvement in scores.

METEOR	score
Eff w/o CLIP	0.1123
Eff w CLIP	0.1333
Res w/o CLIP	0.1319
Res w CLIP	0.1435

Table 5: METEOR score

6 Conclusion

6.1 Summary of results and significance of the work

From this work, we examine that model with CLIP loss showed approved performance both in actual report generation and in evaluation methods we used like BLEU score, ROUGE-L and METEOR. In addition, unlike our expectation, loss with ResNet encoder showed better performance than pretrained EfficientNet-B4 encoder. We can also expect more accurate and acceptable report when applying CLIP loss after finding better baseline model architecture. Although ResNet looks better than EfficientUNet in terms of scores from various evaluation indicators, the results of the prediction reports seen in Tables 1 and 2 confirmed that EfficientUNet may produce better report results. Therefore, follow-up research is needed on a model that enables higher scores and more accurate medical report writing while using evaluation indicators other than the evaluation indicators used in this study or applying other methods other than CLIP loss.

6.2 Ideas for future work

To improve the model performance with less loss and better BLEU score when generating medical reports, we could try the following things for the future work:

- Use MIMIC-CXR dataset from MIT: We finished getting licence to access MIMIC-CXR dataset, but the datasize was too large to use as 4.6 TB even though we bought extra storage on google and upgraded plan on Colab. However, as mentioned in the proposal, the MIMIC-CXR dataset has the largest amount of data among the five commonly used x-ray datasets, so if computing resources are available to handle the MIMIC-CXR dataset, it is expected to yield better results on training a model with MIMIC-CXR.
- Sufficient computing resources: For this project, we used Colab environment provided by Google, with upgraded plan. However, it was still unavoidable from limitation of GPU usage from google, so we struggled on training the model with sufficient epochs we planned. By extra resources, we will be able to evaluate our model better with sufficient training and find optimized epochs to get best performance from our model.
- Modify model architecture: We can try and experiment with different model architecture as baseline model referencing other papers.

Further more, we would like to propose idea for text simplification on generated report to re-generate simple and easy-to-understand contents for laypeople.

There are many text simplification model for medical texts, and we would like to use NapSS(Narrative Prompting and Sentence-matching Summarization) (Lu et al., 2023 [15]) for baseline model to generate simple text. NaPSS is a model which summarize the paragraph level medical text first and then simplifies it. It contains their own human evaluation assessment, involving general readers and medical specialists. NapSS showed better performance than the seq2seq model on English medical corpus in improving lexical similarity.

We propose the idea of generating simplified text by passing the output of our model as the input to NapSS. To improve accuracy and generate more understandable text, we plan to replace the current word embedding of the existing NapSS model with BioWordVec and Med-BERT, then compare the results, which BioWordVec is an biomedical word and sentence embeddings using PubMed and the clinical notes from MIMIC-III Clinical Database (Lu, 2019 [16]) [2] and Med-BERT is a contextualized embedding model pretrained on a structured electronic health records dataset of 28,490,650 patients,

[2]https://github.com/ncbi-nlp/BioSentVec

originally adapts the BERT framework (Rasmy, 2021 [17]). [3] Also we could check the result after adding more reward function like consine similarity with word vectors, to find the best set of loss functions. Since our target is chest x-ray radiology report, we could expect higher accuracy from this model by training extra chest x-ray reports from IU X-ray for example.

6.3 Code

The code used for this project is provided in following github:

- https://github.com/ChanHyeok-Choi/cxr-gen-report-CLIP

7 Contribution of each team member

- Jungmin Park: Produce a ppt & Presenter, Draw figures, Write a report & Analyzer
- Dami Jung: Produce a ppt & Presenter, Investigate references, Write & Organize a report
- Chanhyeok Choi: Develop model architecture & loss function, Evaluate a model on metrics, Write a report

References

[1] Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., & McAuley, J. (2021) Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation. *cs.CL*

[2] Wiehe, A.O., Schneider, F., Blank, S., Wang, X., Zorn, H., & Biemann, C. (2022) Language over Labels: Contrastive Language Supervision Exceeds Purely Label-Supervised Classification Performance on Chest X-Rays. *Proceedings of the AACL-IJCNLP 2022 Student Research Workshop*

[3] Haritaoglu, E.d., Timashov, A., & Tan, M. (2022) Chest X-Ray Report Generation from Chest-X Ray Images. *Stanford CS224N Custom Project*

[4] Johnson, A. , Pollard, T. , Mark, R. , Berkowitz, S., & Horng, S. (2019). MIMIC-CXR Database (version 2.0.0). PhysioNet. https://doi.org/10.13026/C2JT1Q.

[5] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6, 317 (2019). https://doi.org/10.1038/s41597-019-0322-0

[6] Endo, M., Krishnan, R., Krishna, V., Ng, A. Y. & Rajpurkar, P. (2021). Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model. *Proceedings of Machine Learning for Health, PMLR* (158):209-219

[7] Agrawal, E., & Choudhary, P. (2022). EfficientUNet: Modified encoder-decoder architecture for thelung segmentation in chest x-ray images. *Expert System* https://doi.org/10.1111/exsy.13012

[8] Van, H., Kauchak, D., & Leroy, G. (2020). AutoMeTS: The Autocomplete for Medical Text Simplification. *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1424-1434. International Committee on Computational Linguistics.

[9] Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[10] Lin, C. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*, pages 74–81

[11] Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91.

[12] Torchxrayvision. In https://github.com/mlmed/torchxrayvision.

[13] Gal, R., Patashnik, O., Maron, H., Chechik, G., & CohenOr, D. (2021). Stylegan-nada: Clip-guided domain adaptation of image generators.

[14] Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I.. (2021). Learning transferable visual models from natural language supervision. *CoRR, abs/2103.00020.*

[3] https://github.com/ZhiGroup/Med-BERT

[15] Lu, J., Li, J., Wallace, B.C., He, Y., and Pergola, G. (2023). NapSS: Paragraph-level Medical Text Simplification via Narrative Prompting and Sentence-matching Summarization

[16] Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH

[17] Rasmy, L., Xiang, Y., Xie, Z., Tao C., and Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction.