Representative Image Feature Extraction via Contrastive Learning Pretraining for Chest X-ray Report Generation

Yu-Jen Chen[*1] YUJENCHEN@GAPP.NTHU.EDU.TW
Wei-Hsiang Shen[*2] WHSHEN@GAPP.NTHU.EDU.TW
Hao-Wei Chung[1] XDMANWWW@GAPP.NTHU.EDU.TW
Jing-Hao Chiu[1] GWJH101708@GAPP.NTHU.EDU.TW
Da-Cheng Juan[1] DACHENG@GAPP.NTHU.EDU.TW
Tsung-Ying Ho[3] ALBERTYHO@GMAIL.COM
Chi-Tung Cheng[4] ATONG89130@GMAIL.COM
Meng-Lin Li[2] MLLI@EE.NTHU.EDU.TW
Tsung-Yi Ho[1] TYHO@CS.NTHU.EDU.TW

[1] *Department of Computer Science, National Tsing Hua Unviserity, Taiwan*
[2] *Department of Electrical Engineering, National Tsing Hua Unviserity, Taiwan*
[3] *Department of Nuclear Medicine, Chang Gung Memorial Hospital, Linkou, Taiwan*
[4] *Department of Trauma and Emergency Surgery, Chang Gung Memorial Hospital, Linkou, Taiwan*

Editors: Under Review for MIDL 2022

Abstract

Medical report generation is a challenging task since it is time-consuming and requires expertise from experienced radiologists. The goal of medical report generation is to accurately capture and describe the image findings. Previous works pretrain their visual encoding neural networks with large datasets in different domains, which cannot learn general visual representation in the specific medical domain. In this work, we propose a medical report generation framework that uses a contrastive learning approach to pretrain the visual encoder and requires no additional meta information. In addition, we adopt lung segmentation as an augmentation method in the contrastive learning framework. This segmentation guides the network to focus on encoding the visual feature within the lung region. Experimental results show that the proposed framework improves the performance and the quality of the generated medical reports both quantitatively and qualitatively.

Keywords: Radiology, Chest X-ray, Contrastive Learning, Lung Masking, Medical Report Generation

1. Introduction

Medical Images, such as X-rays, have been widely used for supporting clinical diagnosis and diseases treatment. Radiologists are responsible for delivering the interpretation and diagnosis based on their medical expertise and providing the medical reports as reference. However, medical report generation is time-consuming and requires medical expertise from experienced radiologists, which may not be abundant in rural area. Thus, automatic medical report generation is essential in reducing workload, facilitating clinical workflow, and providing additional diagnostic insights.

[*] Contributed equally

© 2022 Y.-J. Chen *et al.*

Most of the current state-of-the-art methods are based on encoder-decoder frameworks (Shin et al., 2016). Many of them (Jing et al., 2017; Li et al., 2018; Chen et al., 2020c) utilize pretrained encoders to extract image features and design decoders to generate the corresponding reports. Although these works design new architectures which result in significant results, their encoders remain to be pretrained with datasets of different domains, where some medical characteristics cannot be fully learned and captured, and lead to limited performance improvement.

In this work, we proposed a medical report generating framework by leveraging the power of Contrastive Learning (CL) and lung segmentation without using extra datasets and additional metadata such as keywords in our model.[1] The encoder is first pretrained with CL using images only and then the entire network is trained end-to-end under a supervised setting (images and reports). Experiments showed that the representation trained by the CL framework results in better performance than other commonly used pretraining settings. In addition, in the CL model, we propose a new augmentation method, lung segmentation, which helps the model attend to the lung region of the given images. The lung segmentation augmentation guides the model to focus on lung regions and generate more detailed and accurate descriptions around the lung regions. In summary, we demonstrate that CL with lung segmentation improves the performance and the quality of the generated medical reports. Moreover, applying the CL pretrained encoder to state-of-the-art methods also boosts their performance.

2. Related Work

2.1. Chest X-ray Report Generation

Most existing report generating approaches (Yuan et al., 2019; Syeda-Mahmood et al., 2020; Xue et al., 2018; Li et al., 2020) follow the traditional encoder-decoder architecture to design their model. Jing et al. (2017) proposed a powerful model combining the co-attention mechanism and the multi-level LSTM to generate a longer paragraph of reports. The co-attention mechanism will combine the visual feature and the paragraph information to achieve better performance. Li et al. (2018) harnessed the reinforcement-learning strategy to update the HRGR-Agent via sentence-level and word-level rewards and maintained a high accuracy in report diagnosis. Finally, Chen et al. (2020c) proposed a memory-based Transformer model to explore the correlation between similar images in the dataset, and such model can report necessary medical terms as well as a meaningful attention map. However, these works focus on designing novel decoders and pretrain their encoders with datasets of domains not related to the application under concern. The medical characteristic cannot be fully learned and captured with such encoder settings. In comparison, our results show that our training framework provides better-encoded features that are applicable to state-of-the-art methods.

1. Source code is available on Github. https://github.com/windstormer/CXR_report_generation_with_CL_pretraining

2.2. Contrastive Learning (CL)

Contrastive learning has shown great promise to learn image representation via reflecting the difference between image features and boosting the performance of various image-related tasks. In fundamental CL frameworks, such as SimCLR (Chen et al., 2020b), MoCo (He et al., 2020), and PIRL (Misra and Maaten, 2020), the feature extractor is trained by minimizing the contrastive loss of the two positive related images and repelling the two negative related images. In the medical domain, several previous works (Chen et al., 2020a; Liu et al., 2021, 2019) also explore the capability that the CL framework achieve performance improvement for medical images.

3. Methodology

Most existing automatic chest X-ray report generation approaches follow an encoder-decoder framework, which consists of an image encoder and a report decoder. The image encoder $E(\cdot) = I \rightarrow V$ encodes the input image I into the visual representation V in latent space. The report decoder $D(\cdot) = V \rightarrow R$ is used to generate the report R from the given image representation V. ResNet is often used as the encoder E and Transformer is often used as the decoder D. The encoder-decoder framework is trained in a supervised manner in which paired chest X-ray images and their corresponding medical reports are required.

In this work, we explore the importance of effective visual representation for chest X-ray report generation. Our image encoder network E is pretrained with contrastive learning techniques that incorporate lung segmentation as one of the augmentation methods.

3.1. Contrastive Learning (CL) Pretraining

In this work, different from most existing approaches whose image encoders are pretrained with large datasets with noisy labels, we adopt the Contrastive Learning (CL) approaches SimCLR (Chen et al., 2020b) and MoCo (He et al., 2020) to pretrain the image encoder.

CL is to minimize the image embedding distance between "positive" image pairs, and maximize the distance between "negative" image pairs. A positive image pair is created by applying two different augmentations to a single image, and a negative image pair is two augmented images that come from two different source images. In this work for X-ray images, the augmentation techniques used are random resized crop and random horizontal flip.

The augmented images are then passed to an encoder neural network $f(\cdot)$, and encoded to their representation vectors $r = f(x) \in R^d$, where d is the dimension of the representation vector, as shown in Fig. 1. Then the representation vectors are projected to a fewer dimension space $z = g(r) \in R^k$ by a project network $g(\cdot)$, which is a neural network of two fully-connected layers.

Finally, for a minibatch of N images, let $i \in I \equiv \{1, 2, ..., 2N\}$ be the index of the augmented sample in the minibatch, the supervised contrastive loss function is formulated as follows:

$$L = -\sum_{i \in I} \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{k \in B(i)} \exp(z_i \cdot z_k/\tau)} \tag{1}$$

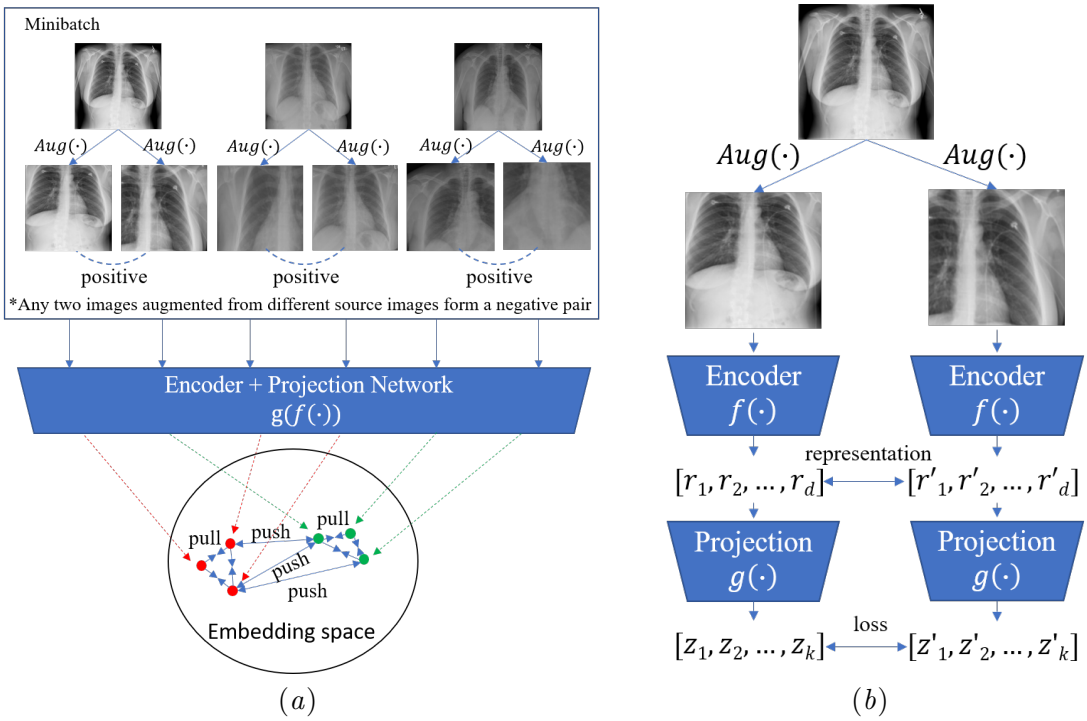


Figure 1: An illustration of the fundamental concept for CL. (a) shows how positive and negative image pairs are created. (b) shows the details of the network, which is a combination of an encoder and a projection network.

where $z_i = g(f(x)) \in R^k$ is the projected vector of the image x, the $\cdot$ operator is the inner product, index $j(i)$ is the positive image of anchor i, $\tau = 0.5$ is the temperature parameter, and $B(i) \equiv I \setminus \{i\}$. The numerator of the loss calculates all the vector similarity of positive pairs, and the denominator is the summation of similarity of all the image pairs (except the positive pairs) inside the minibatch.

3.2. Lung Segmentation Contrastive Learning

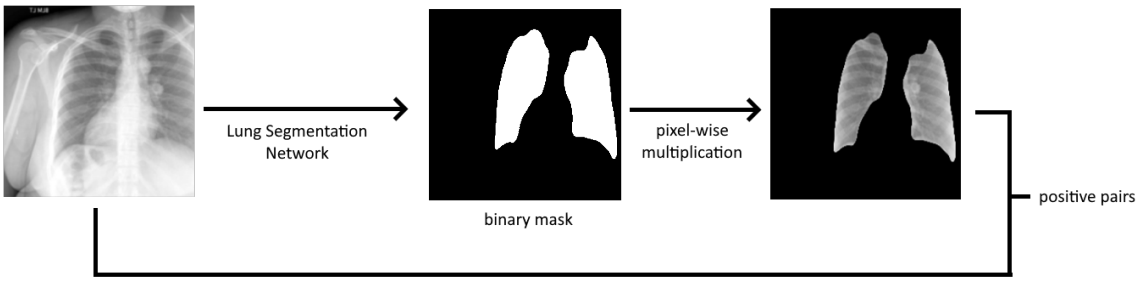


Figure 2: An example of positive pair with the lung masking augmentation added. We can see that the images in a positive pair contain lung image and whole image.

In addition to conventional image augmentation techniques, we explore segmentation methods as augmentation methods to support the model to generate medical findings related to the segmented region. We use a trained lung segmentation model[2] as an augmentation function, which guides the network by focusing on the lung region. The segmentation model is randomly applied with the probability of 0.5 and the binary segmentation mask crops out the lung region in the image, as shown in Fig. 2. Since the contrastive learning setting guides the network to minimize the distance between images with and without lung segmentation crop-off, the network learns to focus less on the non-lung region and generates more detailed features of the lung region. Thus, the report decoder receives lung-specific features and generates reports associated with tissue or abnormality of the lung or rib.

4. Experimental Settings

4.1. Dataset

In this work, we implement report generation on two datasets, Indiana University Chest X-ray Collection (IU X-ray) dataset (Demner-Fushman et al., 2016) which contains 7,470 pairs of frontal and lateral chest X-rays and reports, and the Medical Information Mart for Intensive Care - Chest X-ray (MIMIC-CXR) dataset (Johnson et al., 2019), which contains 377,110 images and 227,835 radiographic reports. We duplicated the data-preprocess of Chen et al. (2020c) and followed their dataset split on both dataset, which is widely adopted by many related works (Jing et al., 2017; Chen et al., 2020c; Rennie et al., 2017; Lu et al., 2017; Li et al., 2018). In detail, IU X-ray dataset is divided by 7:1:2 division of the entire dataset into training/validation/testing sets, as for MIMIC-CXR, we use the official split.

The experiments are evaluated with quantitative NLP metrics that measure the similarity between the predicted and target sentence sequence. The metrics used include BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), and ROGUE-L (Lin, 2004). BLEU measures the similarity between the two sentences by calculating the number of word matches in a weighted manner. These matches are independent of position. METEOR is similar to BLEU but uses the harmonic mean of unigram precision and recall, with recall weighted higher than precision. ROGUE-L takes into account the structural similarity at the sentence level and identifies the longest co-occurring n-grams in the sequence.

4.2. Implementation Details

We adopt the ResNet-50 architecture as the backbone of the encoder in all experiments. At the pretraining phase for the CL approaches (SimCLR and MoCo), all the images in the training set are used, and the model is trained for 500 epochs with the batch size as 64. As for the decoder training (fine-tuning step), the images and the corresponded report in the training set are used, and the entire framework is trained for 200 epochs with the batch size set as 64. Adam optimizer with the learning rate set as 0.001 and decay to 0.00001 followed by the cosine scheduler. We also use the weight decay of 0.000001 to avoid the overfitting problem. The validation set and the test set are used only for evaluation.

2. https://github.com/rani700/xray

5. Results and Discussion

5.1. Comparison across Different Encoder Settings

In this section, we compared pretraining methods for the encoder to show the feasibility of the proposed framework. The compared pretraining methods include using AutoEncoder (AE) pretraining (Baldi, 2012), Imagenet dataset (Imagenet) pretraining, Multi-Label Classification (MLC) pretraining (Bi and Kwok, 2013), without pretraining (scratch), MoCo pretraining, SimCLR pretraining, and SimCLR pretraining with lung segmentation included for the encoder. The labels used in the MLC pretraining are the MeSH (Medical Subject Headings) tag extracted from the reports (the MIMIC-CXR dataset did not provide MeSH tags for the reports). All the experiments in this section are using Transformer as the decoder architecture. Please refer to the appendix for the discussion of two other decoders, LSTM and GRU. The pretrained weight of the encoder is assigned and trained end-to-end with the decoder.

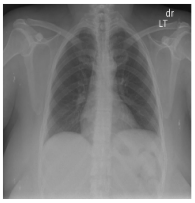
Dataset	Pretraining Method	B-1	B-2	B-3	B-4	M	R-L
IU X-ray	Scratch	0.404	0.249	0.172	0.126	0.161	0.320
	AE	0.417	0.273	0.195	0.148	**0.176**	0.361
	Imagenet	0.420	0.256	0.175	0.127	0.167	0.326
	MLC	0.423	0.270	0.190	0.141	**0.176**	0.359
	MoCo	0.426	0.259	0.175	0.126	0.168	0.325
	SimCLR	**0.456**	**0.289**	**0.202**	**0.150**	**0.176**	**0.362**
	SimCLR (w/ Lung Seg)	0.448	0.282	0.195	0.141	**0.176**	0.340
MIMIC-CXR	Scratch	0.211	0.132	0.090	0.064	0.105	0.250
	AE	0.316	0.186	0.119	0.081	0.119	0.243
	Imagenet	0.270	0.166	0.109	0.076	0.114	0.255
	MoCo	0.334	0.203	0.133	0.091	0.129	0.261
	SimCLR	**0.343**	**0.208**	**0.137**	**0.095**	**0.132**	0.260
	SimCLR (w/ Lung Seg)	0.325	0.198	0.130	0.090	0.129	**0.265**

Table 1: Quantitative comparison among different encoders when using Transformer as decoder. B-n, M, and R-L are short for BLEU-n, METEOR, and ROUGE-L, respectively.

Quantitative Comparison As shown in Table 1, in the IU X-ray dataset, the one without pretraining leads to the worst results, which is expected since the encoder has no prior knowledge of the image. Afterward, due to the expensive pixel-wise image reconstruction, it is difficult for AE to learn the details of images, especially medical images. As for the weight pretrained by Imagenet, though it contains significant information of the nature images, there is a domain gap from chest X-ray images. In MLC pretraining, keywords are extracted from the report, and the model is pretrained with the supervised multi-label classification task. As the encoder has learned little prior knowledge of the reports, we observe that the results are comparable with CL pretraining. Ultimately, CL (SimCLR and MoCo) pretraining outperforms all other pretraining strategies in almost all the metrics. For example, the performance boost at BLEU-1 score compared with other pretraining strategies are up to 9.0%, and 2.25% for SimCLR and MoCo, respectively.

As for the MIMIC-CXR dataset, as shown in Table 1, the improvement trend is similar to that of IU X-ray dataset. CL (SimCLR and MoCo) pretraining outperforms all other

pretraining strategies in all the metrics. The performance boost at BLEU-1 score compared with other pretraining strategies is up to 32.0%, and 28.0% for SimCLR and MoCo, respectively.



AE: the cardiomediastinal silhouette is normal in size and contour no focal consolidation pneumothorax or large pleural effusion t-spine osteophytes

Imagenet: the heart size and mediastinal silhouette are within normal limits for contour the lungs are clear no pneumothorax or pleural effusions the xxxx are intact

CL (Ours): the heart is normal in size the mediastinum is unremarkable there is no pleural effusion pneumothorax or focal airspace disease there is stable irregularity of the posterior left 6th rib which xxxx represents an old fractures

GT: the heart is normal in size the mediastinum is unremarkable the lungs are clear

(a)

AE: the cardiomediastinal silhouette is normal in size and contour no focal consolidation pneumothorax or large pleural effusion t-spine osteophytes

Imagenet: the cardiomediastinal silhouette is normal in size and contour no focal consolidation pneumothorax or large pleural effusion normal xxxx

CL (Ours): the cardiomediastinal silhouette is normal in size and contour hyperexpanded lungs without focal consolidation pneumothorax or large pleural effusion right chest wall surgical clips compatible with prior <unk> negative for acute bone abnormality

GT: normal heart size hyperexpanded lungs without focal consolidation pneumothorax or large pleural effusion left nipple silhouette visualized negative for acute bone abnormality

(b)

Figure 3: Qualitative comparison among different encoders when using Transformer as the decoder. The descriptions marked with red underline represent to match the ground truth. AE and GT stand for the Auto-encoder preatraining and the ground truth, respectively.

Qualitative Comparison Fig. 3 shows the ground truth and the predicted description by different encoders. In case (a), the AE pretrained model generates the description which did not match the ground truth, and the Imagenet pretrained model only describes that the lung is clear, whereas, with CL model, it points out the heart condition and mediastinum condition. In case (b), only CL pretrained model points out "hyperexpanded lungs" and "negative for acute bone abnormality", whereas the Imagenet pretrained model and the AutoEncoder predict some common information. This shows that the CL pretrained encoder generates precise descriptions when using the Transformer decoder.

5.2. Comparison with Previous Studies

We compare our model with existing studies on the test set of both IU X-ray and MIMIC-CXR datasets in Table 2. Results show that the proposed method improves the performance with large margins without any extra information, such as pretraining on other datasets or keyword extraction.

In addition, to show the feasibility of the CL pretrained encoder, we applied our MoCo pretrained encoder to the state-of-the-art model (R2Gen). Table 2 shows the BLEU-1 score improved from 0.455 and 0.354 (R2Gen) to 0.466 and 0.359 (MoCo pretraining + R2Gen) when MoCo pretrained feature is used in IU X-ray and MIMIC-CXR dataset, respectively. In conclusion, results show that CL pretrained encoder gives representative image embeddings that can improve the performance of state-of-the-art methods.

Dataset	Method	B-1	B-2	B-3	B-4	R-L
IU X-ray	Att2in† (Rennie et al., 2017)	0.399	0.249	0.172	0.126	0.321
	AdaAtt† (Lu et al., 2017)	0.436	0.288	0.203	0.150	0.354
	HRGR-Agent† (Li et al., 2018)	0.438	**0.298**	**0.208**	**0.151**	0.322
	CoAtt† (Jing et al., 2017)	0.455	0.288	0.205	0.154	**0.369**
	Transformer (Vaswani et al., 2017)	0.420	0.256	0.175	0.127	0.326
	Ours (SimCLR pretraining + Transformer)	**0.456**	0.289	0.202	0.150	0.362
	R2Gen (Res50) (Chen et al., 2020c)	0.455	0.293	0.210	0.160	0.360
	Ours (MoCo pretraining + R2Gen)	**0.466**	**0.303**	**0.218**	**0.165**	**0.361**
MIMIC-CXR	AdaAt† (Lu et al., 2017)	0.299	0.185	0.124	0.088	0.266
	Att2in† (Rennie et al., 2017)	0.325	0.203	0.136	**0.096**	**0.276**
	Transformer (Vaswani et al., 2017)	0.314	0.192	0.127	0.090	0.265
	Ours (SimCLR pretraining + Transformer)	**0.343**	**0.208**	**0.137**	0.095	0.260
	R2Gen (Res50) (Chen et al., 2020c)	0.354	0.236	0.164	0.118	**0.305**
	Ours (MoCo pretraining + R2Gen)	**0.359**	**0.239**	**0.167**	**0.121**	0.304

Table 2: Performance comparison with previous studies on the IU X-Ray dataset and MIMIC-CXR dataset. B-n and R-L are short for BLEU-n and ROUGE-L, respectively. † indicates the results are quoted from the published paper. Note that we reproduce R2Gen (Chen et al., 2020c) in both datasets with ResNet-50 architecture followed by their recommended setting.

5.3. Comparison between Encoders with and without Lung Segmentation

In this section, we compare the effectiveness of lung segmentation augmentation. We use F1 scores on lung-related MeSH tags to evaluate whether lung segmentation gives more accurate descriptions regarding lung regions, as shown in Table 3. The results show that lung segmentation would improve the accuracy of lung-related tags. This is because the encoder focuses more on lung regions and generates lung-specific features for the decoder. However, since the encoder cannot utilize features in the entire image, the overall performance of the generated descriptions would drop slightly, as shown in Table 1.

Keyword	CL w/ Lung Segmentation	CL w/o Lung Segmentation
pneumothorax	**0.808**	0.752
volume	**0.191**	0.116
effusion	**0.789**	0.772
calcification	**0.120**	0.030

Table 3: F1 score of lung-related keywords appeared in experiments w/ and w/o adding lung segmentation.

6. Conclusion

In this work, we showed that the contrastive learning method provides high-quality image representation for chest X-ray report generation. Unlike most existing approaches, which pretrained the encoder with the dataset of different domains, our method successfully extracted the feature from the target dataset, which there is no domain bias. Furthermore, with lung segmentation included, the network generates more detailed and accurate descriptions regarding the lung regions. Experimental results on two datasets have shown that our framework improved the performance with a large margin from other pretraining strategies.

References

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49. JMLR Workshop and Conference Proceedings, 2012.

Wei Bi and James Kwok. Efficient multi-label classification with many labels. In *International conference on machine learning*, pages 405–413. PMLR, 2013.

Bingzhi Chen, Zheng Zhang, Jianyong Lin, Yi Chen, and Guangming Lu. Two-stream collaborative network for multi-label chest x-ray image classification with lung segmentation. *Pattern Recognition Letters*, 135:221–227, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020b.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020c.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *arXiv preprint arXiv:1805.08298*, 2018.

Xin Li, Rui Cao, and Dongxiao Zhu. Vispi: Automatic visual perception and interpretation of chest x-rays. In *Medical Imaging with Deep Learning*, 2020. URL https://openreview.net/forum?id=43NPhWwbek.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*, 2021.

Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10632–10641, 2019.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

Tanveer Syeda-Mahmood, Ken C. L. Wong, Yaniv Gur, Joy T. Wu, Ashutosh Jadhav, Satyananda Kashyap, Alexandros Karargyris, Anup Pillai, Arjun Sharma, Ali Bin Syed, Orest Boyko, and Mehdi Moradi. Chest x-ray report generation through fine-grained label learning. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 561–571, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59713-9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 457–466, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00928-1.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 721–729, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32226-7.

Acknowledgement

We thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

Appendix A. Comparison among Different Encoders for LSTM and GRU

A.1. LSTM

Quantitative Comparison Table 4 shows that SimCLR pretraining improves the BLEU-1 and BLEU-4 score up to 42% and 31%, respectively, compared to other pretraining strategies on the IU X-ray dataset. Moreover, in the MIMIC-CXR dataset, CL pretraining outperforms other pretraining strategies in all metrics.

Qualitative Comparison Fig. 4 shows the ground truth and the predicted description by different encoders when using LSTM as a decoder. In case (a), the AutoEncoder pretrained model only matches the mediastinum description with GT, and the Imagenet pretrained model describes the lung parenchyma and pleura condition, whereas, with CL model, it points out all of them with the bone condition additionally. In case (b), only the CL pretrained model points out "low lung volumes", whereas the Imagenet pretrained model only predicts that the lung stays normal. This shows that the CL encoder aids the model in generating precise descriptions.

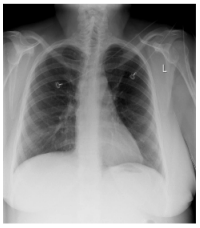
A.2. GRU

Quantitative Comparison Table 5 shows that in the IU X-ray dataset, SimCLR and MoCo pretraining boost the BLEU-1 score up to 7.75% and 7.74%, respectively, compared with other pretraining strategies. As for the MIMIC-CXR dataset, SimCLR pretraining achieves the highest score in all evaluate metrics, about 5.5% higher than without pretraining the encoder (Scratch).

Qualitative Comparison Fig. 5 shows the ground truth and the predicted description by different encoders when using GRU as the decoder. In case (a), the Auto-Encoder pretrained model did not describe the bone condition, and the Imagenet pretrained model

Dataset	Pretraining Method	B-1	B-2	B-3	B-4	M	R-L
IU X-ray	Scratch	0.299	0.185	0.132	0.100	0.134	0.324
	AE	0.389	0.233	0.157	0.111	0.154	0.317
	Imagenet	0.397	0.246	0.170	0.126	0.160	**0.329**
	MLC	0.410	0.250	0.170	0.123	0.164	0.325
	MoCo	0.425	0.260	0.177	0.128	0.169	0.326
	SimCLR	0.426	0.265	0.181	0.131	0.168	0.326
	SimCLR (w/ Lung Seg)	**0.431**	**0.268**	**0.185**	**0.137**	**0.172**	**0.332**
MIMC-CXR	Scratch	0.312	0.193	0.130	0.092	0.128	0.272
	AE	0.282	0.174	0.117	0.083	0.119	0.264
	Imagenet	0.309	0.193	0.131	0.094	0.130	0.276
	MoCo	0.309	0.191	0.128	0.091	0.127	0.270
	SimCLR	**0.330**	**0.206**	**0.139**	**0.099**	**0.134**	**0.278**
	SimCLR (w/ Lung Seg)	0.325	0.203	0.138	0.095	**0.134**	**0.278**

Table 4: Quantitative comparison among different encoders when using LSTM as decoder. B-n, M, and R-L are short for BLEU-n, METEOR, and ROUGE-L, respectively.



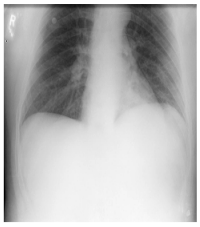
AE: the cardiomediastinal silhouette and vasculature are within normal limits for size and contour the lungs are normally inflated and clear osseous structures are within normal limits for patient age

Imagenet: heart size is normal the lungs are clear there are no focal air space consolidations no pleural effusions or pneumothoraces the hilar and mediastinal contours are normal normal pulmonary vascularity

CL (Ours): the cardiomediastinal silhouette is normal in size and contour aortic atherosclerosis no focal consolidation pneumothorax or large pleural effusion negative for acute bone abnormality

GT: the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size the lungs are clear of focal airspace disease pneumothorax or pleural effusion there are no acute bony findings

(a)



AE: the heart is normal in size the mediastinum is unremarkable the lungs are clear

Imagenet: the cardiomediastinal silhouette is normal in size and contour no focal consolidation pneumothorax or large pleural effusion normal xxxx

CL (Ours): the cardiac silhouette is at the upper limits of normal for size there are low lung volumes with bronchovascular crowding no focal areas of pulmonary consolidation no pneumothorax no pleural effusion minimal degenerative endplate changes of the thoracic spine

GT: the xxxx examination consists of frontal and lateral radiographs of the chest a <unk> of 3 images were obtained the cardiomediastinal contours are within normal limits <unk> for low lung volumes and patient rotation there is xxxx xxxx atelectasis no consolidation pleural effusion or pneumothorax calcified right infrahilar lymph xxxx again seen partially

(b)

Figure 4: Qualitative Comparison among Different Encoders when using LSTM as the decoder. The descriptions marked with red underline represent ground truth matching. AE and GT stand for the Auto-encoder preatraining and the ground truth, respectively.

only describes the lung condition, whereas the CL pretrained model points out both of them. In case (b), only CL pretrained model point out "heart is mildly enlarged" and the lung had no problem, whereas the Imagenet pretrained model only predicts that the lung stays normal, and the AutoEncoder predicts others false information. Qualitatively, the

CL encoder aids the model in generating precise descriptions while using the GRU decoder than other pretraining approaches.

Dataset	Pretraining Method	B-1	B-2	B-3	B-4	M	R-L
IU X-ray	Scratch	0.351	0.221	0.156	0.117	0.153	0.333
	AE	0.389	0.241	0.166	0.121	0.161	0.332
	Imagenet	0.417	0.254	0.173	0.124	0.165	0.328
	MLC	0.412	**0.263**	**0.189**	**0.145**	**0.169**	**0.341**
	MoCo	**0.424**	0.259	0.174	0.124	**0.169**	0.331
	SimCLR	0.423	0.259	0.175	0.126	0.166	0.327
	SimCLR (w/ Lung Seg)	0.415	0.255	0.174	0.127	0.165	0.331
MIMC-CXR	Scratch	0.309	0.193	0.130	0.093	0.128	0.274
	AE	0.299	0.184	0.124	0.088	0.123	0.270
	Imagenet	0.313	0.196	0.132	0.095	0.129	0.275
	MoCo	0.309	0.190	0.127	0.090	0.126	0.269
	SimCLR	0.327	0.204	0.137	0.098	0.134	0.278
	SimCLR (w/ Lung Seg)	**0.335**	**0.209**	**0.140**	**0.100**	**0.136**	**0.279**

Table 5: Quantitative comparison among different encoders when using GRU as decoder. B-n, M, and R-L are short for BLEU-n, METEOR, and ROUGE-L, respectively.

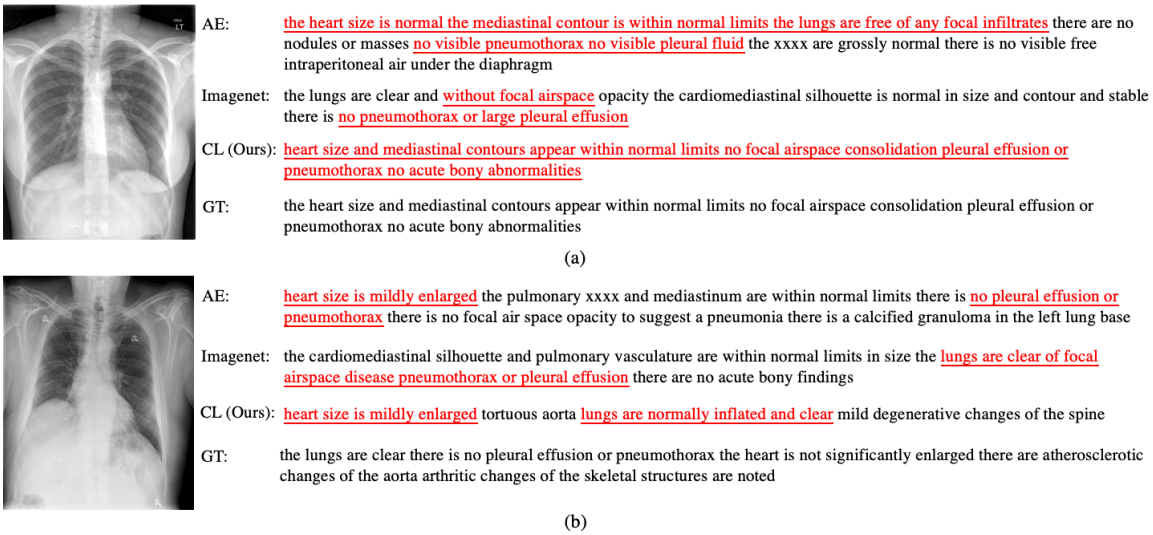


Figure 5: Qualitative comparison among different encoders when using GRU as the decoder. The descriptions marked with red underline represent to match the ground truth. AE and GT stand for the Auto-encoder preatraining and the ground truth, respectively.

Appendix B. Qualitative Comparison between Encoders with and without Lung Segmentation

In this section, we discuss the qualitative comparison between encoders pretrained by using and not using lung segmentation-based augmentation. Fig. 6 shows the ground truth and the predicted descriptions by different encoders. In case (a), the CL model only describes the clearness of the lungs, whereas with lung segmentation, the model can also point out low lung volumes which match the ground truth description. In case (b) and (c), the CL model falsely predicts an expanded (inflated) lungs. In contrast, the lung segmentation model and the ground truth describe a "low lung volume". This shows that lung segmentation aids the model in generating accurate descriptions, especially on the lung area.

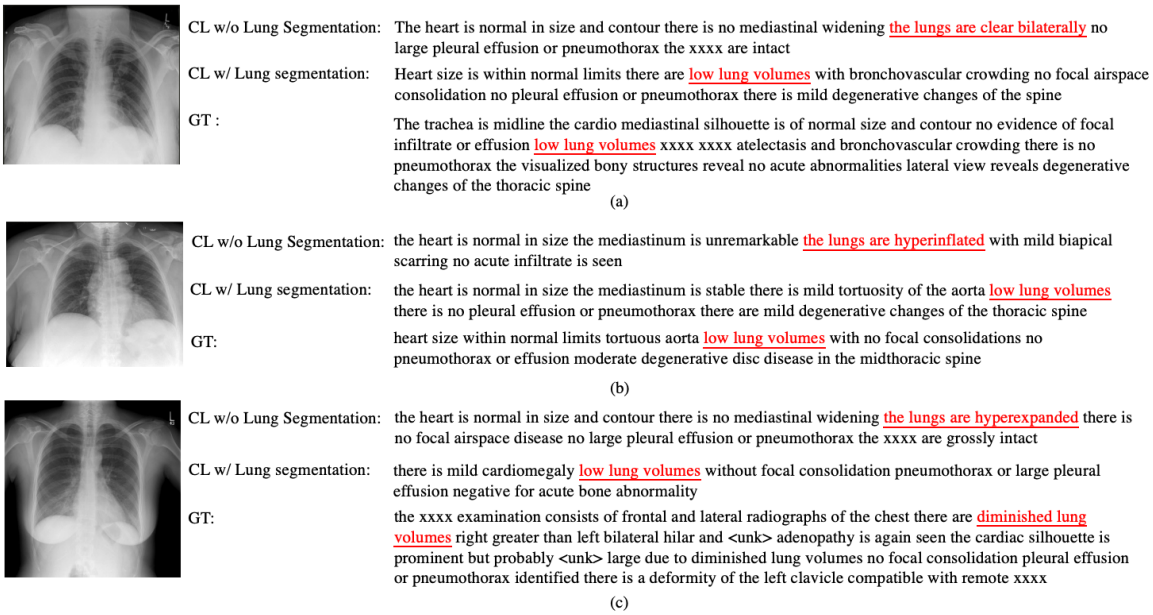


Figure 6: Qualitative comparison for whether lung segmentation is adopted for CL pretraining. With lung segmentation, the model can generate accurate descriptions about the lung, whereas the naive CL sometimes generates false descriptions about the lung volumes. GT stands for the ground truth description.