Progressive Transformer-Based Generation of Radiology Reports

Farhad Nooralahzadeh[1], **Nicolas Perez Gonzalez**[1], **Thomas Frauenfelder**[1],
Koji Fujimoto[2†], **Michael Krauthammer**[1]
[1]University of Zürich and University Hospital of Zürich , [2]Kyoto University
{farhad.nooralahzadeh,nicolas.perez,michael.krauthammer}@uzh.ch
thomas.frauenfelder@usz.ch, †kfb@kuhp.kyoto-u.ac.jp

Abstract

Inspired by Curriculum Learning, we propose a consecutive (*i.e.*, image-to-text-to-text) generation framework where we divide the problem of radiology report generation into two steps. Contrary to generating the full radiology report from the image at once, the model generates global concepts from the image in the first step and then reforms them into finer and coherent texts using a transformer architecture. We follow the transformer-based sequence-to-sequence paradigm at each step. We improve upon the state-of-the-art on two benchmark datasets.

1 Introduction

The analysis of X-rays in medical practice is the most common and important task for radiologists. With years of training, these experts learn to recognize particular features in the image that are later translated to a written report in a clinically appropriate manner. This is a labor intensive and time consuming task, especially difficult for young trainees. With increasing demand on imaging examinations, the burden on radiologists has increased over time, requiring the addition of the technologies to improve their workflow.

Previous research on radiology report generation has mostly focused on image-to-text generation tasks. Jing et al. (2018) introduced a co-attention mechanism to generate full paragraphs. Lovelace and Mortazavi (2020) explored report generation through transformers. More recently, Zhang et al. (2020) used a preconstructed graph embedding module on multiple disease findings to assist the generation of reports. Finally, Chen et al. (2020) proposed to generate radiology reports via memory-driven transformer and showed that their proposed approach outperforms previous models with respect to both language generation metrics and clinical evaluation. These systems have significant potential in many clinical settings, including improvement in workflow in radiology, clinical decision support, and large-scale screening using X-ray images.

In this work, we focus on generating reports from chest X-ray images innovating with a double staged transformer based architecture. Our contributions in this paper can be summarized as follows: (i) We propose to produce radiology reports via a simple but effective progressive text generation model by incorporating high-level concepts into the generation process [1], (ii) We conduct extensive experiments and the results show that our proposed models outperforms the baselines and existing models, *i.e.*, achieving a substantial $+1.23\%$ increase in average over all language generation metrics in IU X-RAY, and the increase of $+3.2\%$ F1 score in MIMIC-CXR, against the best baseline R2GEN, and (iii) We perform a qualitative analysis to further demonstrate the quality and properties of the generated reports.

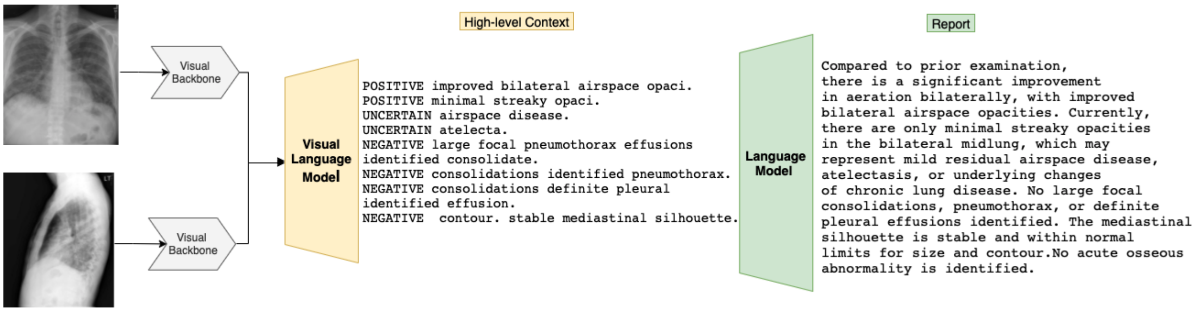
2 Method

An essential challenge in the radiology report generation is modeling the clinical coherence across the entire report. Contrary to generating the full radiology report from the image at once, we propose a consecutive (*i.e.*, image-to-text-to-text) generation framework (inspired by Curriculum Learning (Bengio et al., 2009) and the work of Tan et al. (2020)). As shown in Figure 1, we divide the problem of radiology report generation into two steps. In the first step, the model generates global concepts from the image and then reforms them into finer and coherent text using a transformer architecture. Each step follows the transformer based sequence-to-sequence paradigm.

Model Architecture Instead of generating the full report from an input radiology image, we frame

[1] Our code is available at https://github.com/uzh-dqbm-cmi/ARGON

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2824–2832
November 7–11, 2021. ©2021 Association for Computational Linguistics



POSITIVE improved bilateral airspace opaci.
POSITIVE minimal streaky opaci.
UNCERTAIN airspace disease.
UNCERTAIN atelecta.
NEGATIVE large focal pneumothorax effusions
identified consolidate.
NEGATIVE consolidations identified pneumothorax.
NEGATIVE consolidations definite pleural
identified effusion.
NEGATIVE contour. stable mediastinal silhouette.

Compared to prior examination,
there is a significant improvement
in aeration bilaterally, with improved
bilateral airspace opacities. Currently,
there are only minimal streaky opacities
in the bilateral midlung, which may
represent mild residual airspace disease,
atelectasis, or underlying changes
of chronic lung disease. No large focal
consolidations, pneumothorax, or definite
pleural effusions identified. The mediastinal
silhouette is stable and within normal
limits for size and contour.No acute osseous
abnormality is identified.

Figure 1: Overview of our proposed framework

the generation process such as: $X \rightarrow C \rightarrow Y$, where $X = \{x_1, x_2, ..., x_S\}$, $x_s \in \mathbb{R}^d$. X is a radiology image and x_s is a sequence of patch features extracted from visual extractor and d is the size of the feature vectors. $C = \{c_1, c_2, ..., c_T\}$, $c_t \in \mathbb{V}$, and $Y = \{y_1, y_2, ..., y_{T'}\}$, $y_{t'} \in \mathbb{V}'$, are the generated tokens at intermediate and final steps, respectively. T and T' are the length of generated tokens and $\mathbb{V}$, $\mathbb{V}'$ are the vocabulary of all possible tokens at each step. Our framework can be partitioned into three major components such as: 1) A visual backbone 2) An intermediate encoder-decoder as a visual language model (ViLM) and 3) A final encoder-decoder as a language model (LM).

Visual Backbone Given a set of radiology images (I), the visual backbone extracts the visual features X and results in the source sequence $\{x_1, x_2, ..., x_s\}$ for the subsequent visual language model. The visual backbone can be formulated based on pre-trained Convolutional Neural Networks (CNN), e,g., DenseNet (Huang et al., 2016), VGG (Simonyan and Zisserman, 2015) or ResNet (He et al., 2016). We find DenseNet to be more effective in our generation task and therefore use it as our based visual feature extractor.

Visual Language Model (ViLM) We adapt a state-of-the-art image captioning model, Meshed-Memory Transformer ($\mathcal{M}^2$ Tr.), introduced by (Cornia et al., 2020) for the intermediate step of our architecture. $\mathcal{M}^2$ Tr. is a transformer (Vaswani et al., 2017) based model which presents two adjustments that leveraged the performance of the model: Memory Augmented Encoder and Meshed Decoder. Memory Augmented Encoder extends the set of keys and values in the encoder with additional "slots" to extract a priori information. The priori information is not based on the input; it is encoded in learnable vectors, which are concatenated to keys and values and can be directly updated via SGD. Unlike the original decoder block in transformer, which only performs a cross-attention between the last encoding layer and the decoding layers, the $\mathcal{M}^2$ Tr. presents a meshed connection with all encoding layers. We refer the reader to Cornia et al. (2020) for a detailed description of the Meshed-Memory Transformer.

Given the visual language model structure, the objective of the intermediate generation phase can be formalized as :

$$p_\theta(C \mid I) = \prod_{t=1}^{T} p_\theta(c_t \mid c_{<t}, I)$$

where C at the intermediate step is the high-level context that contains informative and important tokens to serve as skeletons for the following enrichment process. To train the ViLM , we maximize the conditional log-likelihood $\sum_{t=1}^{T} \log p_\theta(C \mid I)$ on the training data to find the optimized θ^*.

Language Model The third component of our architecture is also based on the transformer as a sequence-to-sequence model that follows the conditional probability as:

$$p_{\theta'}(Y|C) = \prod_{t'}^{T'} p_{\theta'}(y_{t'}|y_{<t'} \mid f_{\theta'}(C))$$

where f_θ is an encoder that transforms the input sequence (*e.g.*, high-level context) into another representation that are used by the language model p_θ at decoding step. We employed BART (Lewis et al., 2020) as a pre-trained language model and fine-tune on our target domain. BART includes a BERT-like encoder and GPT2-like decoder. It has an autoregressive decoder and can be directly fine tuned for sequence generation tasks such as paraphrasing and summarization. Similar to the previous module, to train the LM, we maximize the conditional likelihood $\sum_{t'}^{T'} \log p_{\theta'}(Y \mid C)$ using

```
Algorithm 1: Training the Progressive Transformer-Based Generation of Radiology Reports
  Input: Radiology Reports $R$ and Images $I$,
         Pretrained CNNs Model
         DensNet-121, Pretrained LM BART
1 Extract a high-level context $C$ from
    Radiology Reports $R$
2 Fine-tune ViLM and LM independently
  Output: Fine-tuned ViLM and LM for
          report generation from Images $I$ in
          a progressive manner
```

the training set.

Training Algorithm 1 shows the training steps of our proposed architecture. We first extract a high-level context C for each report in training dataset (see Figure 1). To do so, we employed MIRQI tools implemented by Zhang et al. (2020). Each training report is processed with disease word extraction, negation/uncertainty extraction, and attributes extraction based on dependency graph parsing. A similar method proposed in NegBio (Peng et al., 2018) and CheXpert (Irvin et al., 2019) for entity extraction and rule based negation detection is adopted in MIRQI. Then, we construct independent training data for each stage, *i.e.*, fine-tuning of the ViLM and LM. More concretely, given training pairs (I, C), we fine-tune ViLM. On the other hand, the BART is fine-tuned by using training pairs (C, R) in the LM stage. Having fine-tuned the ViLM and LM, the model first generates the intermediate context and subsequently generates the full radiology report by adding finer-grained details at the final stage.

3 Experiments

Datasets We examine our proposed framework on two datasets as follows: i) IU X-RAY (Demner-Fushman et al., 2015), a public radiology dataset that contains 7,470 chest X-ray images and 3,955 radiology reports, each report is associated with one frontal view chest X-ray image and optionally one lateral view image, ii) MIMIC-CXR (Johnson et al., 2019), a large publicly available database of labeled chest radiographs that contains 473,057 chest X-ray images and 206,563 reports. In order to compare our method with previous works, we use the available split on two datasets (*i.e.*, the IU X-RAY and MIMIC-CXR splits available in Chen et al. (2020).)[2]

Evaluation Metrics The evaluation of the models is preformed using general NLG metrics including BLUE (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011) and ROUGE-L (Lin, 2004). However, to address the shortcoming of the conventional NLG metrics in medical abnormality detection (Liu et al., 2019; Lovelace and Mortazavi, 2020; Chen et al., 2020), we also report clinical efficacy (CE) metrics that compare CheXpert extracted labels for the generated and reference reports[3]. To alleviate randomness of the scores, the mean of five different runs are reported.

Baselines We consider the following baselines in our evaluation process: (i) TRANSFORMER: The vanilla transformer is employed in the ViLM component to generate radiology reports in a standard manner, and (ii) $\mathcal{M}^2$ TR.: The Meshed-Memory Transformer is used in the ViLM component to generate text without progressive style.

Moreover, we compare our model with previous studies reported in Chen et al. (2020), *e.g.*, **ST** (Vinyals et al., 2015), **ATT2IN** (Rennie et al., 2017), **ADAATT** (Lu et al., 2017), **TOPDOWN** (Anderson et al., 2018), **COATT** (Jing et al., 2018), **HRGR** (Li et al., 2018), **CMAS-RL** (Jing et al., 2019) and **R2GEN** (Chen et al., 2020) (see Section A in appendix for more detail). For reproducibility, the model configuration and training are described in Section B of the Appendix.

4 Results and Discussion

Effect of progressive generation To show the effectiveness of our model, we conduct experiments with baseline models, including our proposed model (*i.e.*, $\mathcal{M}^2$ TR. PROGRESSIVE) as reported in Table 1. The results shows that $\mathcal{M}^2$ TR. provides better performance than the vanilla transformer which confirms the validity of incorporating memory matrices in the encoder and meshed connectivity between encoding and decoding modules. Our progressive model consistently outperforms the standard and single-stage ViLMs by a large margin on almost all metrics in both benchmark datasets, which clearly highlights the benefits of

[2] https://github.com/cuhksz-nlp/R2Gen

[3] https://github.com/MIT-LCP/mimic-cxr/tree/master/txt

Data	Model	NLG Metrics						CE Metrics		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-Ray	Transformer	0.388	0.246	0.176	0.133	0.163	0.340	-	-	-
	$\mathcal{M}^2$ Tr.	0.475	0.301	0.228	0.180	0.169	0.373	-	-	-
	$\mathcal{M}^2$ Tr. Progressive	**0.486**	**0.317**	**0.232**	**0.173**	**0.192**	**0.390**	-	-	-
MIMIC -CXR	Transformer	0.305	0.188	0.126	0.092	0.128	0.264	0.313	0.224	0.261
	$\mathcal{M}^2$ Tr.	0.361	0.221	0.146	0.101	0.139	0.266	**0.324**	0.241	0.276
	$\mathcal{M}^2$ Tr. Progressive	**0.378**	**0.232**	**0.154**	**0.107**	**0.145**	**0.272**	0.240	**0.428**	**0.308**

Table 1: The performance of baseline and our progressive model on the test sets of IU X-Ray and MIMIC-CXR datasets with respect to NLG and CE metrics. BL-n denotes BLEU score using up to n-grams; MTR and RG-L denote METEOR and ROUGE-L, respectively. The performance of all models is averaged from five runs.

the progressive generation strategy. However the precision of the progressive model is lower than the baselines. We observe that the progressive generation produces long reports mostly by adding the abnormality mentions in negation mode (*e.g.*, *No evidence of pneumonia*, *There is no pneumothorax*), therefore it increases the number of false positives (FPs) in the CE metrics.

In Table 2, we compare our full model (*i.e.*, $\mathcal{M}^2$ Tr. Progressive) with the previous works on the same datasets. In general, memory based transformer methods offer significant improvements across all metrics compared to the recurrent neural networks (RNNs) based architectures. This is illustrated by comparing R2GEN, $\mathcal{M}^2$ Tr. and our full model with the other techniques (see also Table 1). Our model achieves competitive results compare to R2GEN, *i.e.*, $+1.23\%$ average on all NLG metrics in IU X-ray, $+0.83\%$ and $+3.2\%$ average on all NLG metrics and F1 score, respectively, in the MIMIC-CXR dataset. This indicates the benefits of using the $\mathcal{M}^2$ Tr. together with our progressive strategy in the radiology reports generation task. We hypothesise that the use of `MIRQI` in the intermediate context generation provides informative and high-quality plans which results in reasonable descriptions for clinical abnormalities in the last generation stage.

Analysis As a qualitative analysis to explain the effectiveness of our progressive model, we examine some of the generated reports with their references from the MIMIC-CXR test dataset (see Figure 2 in the Appendix). We show the text alignments between the reference text and generated one with the same colors. It can be seen in the top two examples the progressive model is able to provide reports aligned with the reference texts where the baseline model fails to cover them, *e.g.*, *post median sternotomy*, and *mitral valve replacement*, *The mediastinal contours*, *enlargement of the cardiac silhouette*, *bilateral pleural effusions* and *compressive atelectasis* in the top two examples are not generated by $\mathcal{M}^2$ Tr.. Although our model shows improvements in the NLG and CE metrics evaluation, it still fails to generate clinically coherent and error-free reports. For example, in the third example of Figure 2, the *mild pulmonary edema* is incorrect since the *No new parenchymal opacities* in the reference implies negative pulmonary edema. Furthermore, the sentence *left plueral effusion* in the last example is not consistent with the previous text *bilateral pleural effusion.* Additionally, the examples in Figure 2 contain a comparison of study against to the previous study such as *As compared to the previous ...* and *In comparison with the study* ... in the generated reports. This is a little surprising since the model does not have any clue about the previous report of a patient in its design. It can be attributed to the fact that these template sentences are more frequent in the training set. The examples also show that the progressive model generates a more comprehensive report compare to the baseline.It includes occasionally the extra mentions of medical terms compared to the reference text (*e.g.*, *There is no focal consolidation* and *No evidence of pneumonia* in examples 1 and 3, respectively), which result in false-positive mention of observations in the CheXpert labeler of the CE metrics.

5 Conclusion

We propose to produce radiology report via a simple but effective progressive text generation model by incorporating high-level concepts into the generation process. The experimental results show

Data	Model	NLG Metrics						CE Metrics		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-Ray	ST$^{\odot}$	0.216	0.124	0.087	0.066	-	0.306	-	-	-
	Att2in$^{\odot}$	0.224	0.129	0.089	0.068	-	0.308	-	-	-
	AdaAtt$^{\odot}$	0.220	0.127	0.089	0.068	-	0.308	-	-	-
	CoAtt$^{\odot}$	0.455	0.288	0.205	0.154	-	0.369	-	-	-
	HRGR$^{\odot}$	0.438	0.298	0.208	0.151	-	0.322	-	-	-
	CMAS-RL$^{\odot}$	0.464	0.301	0.210	0.154	-	0.362	-	-	-
	R2Gen$^{\odot}$	0.470	0.304	0.219	0.165	0.187	0.371	-	-	-
	$\mathcal{M}^2$ Tr. Progressive	**0.486**	**0.317**	**0.232**	**0.173**	**0.192**	**0.390**	-	-	-
MIMIC -CXR	ST$^{\oplus}$	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
	Att2in$^{\oplus}$	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
	AdaAtt$^{\oplus}$	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
	TopDown$^{\oplus}$	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
	R2Gen $^{\odot}$	0.353	0.218	0.145	0.103	0.142	**0.277**	**0.333**	0.273	0.276
	$\mathcal{M}^2$ Tr. Progressive	**0.378**	**0.232**	**0.154**	**0.107**	**0.145**	0.272	0.240	**0.428**	**0.308**

Table 2: Comparisons of our full model with previous studies on the test sets of IU X-Ray and MIMIC-CXR with respect to language generation (NLG) and clinical efficacy (CE) metrics. $\odot$ refers to the result that is directly cited from the original paper and $\oplus$ represents the replicated results reported on Chen et al. (2020).

that our proposed model outperforms the baselines and a wide range of radiology report generation methods, in terms of language generation and clinical efficacy metrics. Further, the manual analysis demonstrates the ability of the model to produce long and more clinically coherent reports, however there is still room for improvement.

Acknowledgements

This research has received funding from the Kyoto University-University of Zurich Joint Funding Program. The authors would also like to thank the anonymous reviewers for their feedback.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

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, Online. Association for Computational Linguistics.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely connected convolutional networks. *CoRR*, abs/1608.06993.

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.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.

Alistair E. W. 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 large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.

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

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Advances in neural information processing systems*, pages 1530–1540.

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.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. 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.

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.

Yifan Peng, Xiaosong Wang, Le Lu, M. Bagheri, R. Summers, and Z. Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188 – 196.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical Sequence Training for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P Xing, and Zhiting Hu. 2020. Progressive generation of long text. *arXiv preprint arXiv:2006.15720*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, R. Salakhutdinov, R. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12910–12917.

A Previous Models

- **ST** (Vinyals et al., 2015): The model is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image.

- **ATT2IN** (Rennie et al., 2017): The CNN-RNN based model which rather than utilizing a static, spatially pooled representation of the image, it employs the attention model. The attention model dynamically re-weight the input spatial (CNN) features to focus on specific regions of the image at each time step. The model considers a modification of the architecture of the attention model for captioning in Xu et al. (2015), and input the attention-derived image feature only to the cell node of the LSTM.

- **ADAATT** (Lu et al., 2017): It is an adaptive attention encoder-decoder framework which provides a fallback option to the decoder. At each time step, the model decides whether to attend to the image (and if so, to which regions) or to the visual sentinel. The model decides whether to attend to the image and where, in order to extract meaningful information for sequential word generation.

- **TOPDOWN** (Anderson et al., 2018): A combined bottom-up and top-down visual attention mechanism (based on Faster R-CNN). The bottom-up mechanism proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. The model enables attention to be calculated more naturally at the level of objects and other salient regions.

- **COATT** (Jing et al., 2018): A multi-task learning framework which jointly performs the prediction of tags and the generation of paragraphs. The model is based on a hierarchical LSTM model and incorporates a co-attention mechanism to localize regions containing abnormalities and generate narrations for them.

- **HRGR** (Li et al., 2018): A Hybrid Retrieval-Generation Reinforced Agent consists of a CNN to extract visual features which is then transformed into a context vector by an image encoders. Then a sentence decoder (RNNs-based with attention mechanism) recurrently generates a sequence of hidden states which represent sentence topics. A retrieval policy module is employed to decide for each topic state to either automatic generate a sentence, or retrieve a specific template from a template database.

- **CMAS-RL** (Jing et al., 2019): It is a LSTM based framework for generating chest X-ray imaging reports by exploiting the structure information in the reports. It explicitly models the between-section structure by a two-stage framework, and implicitly captured the within-section structure with a Cooperative Multi-Agent System (CMAS) comprising three agents: Planner (PL), Abnormality Writer (AW) and Normality Writer (NW). The entire system was trained with REINFORCE algorithm.

- **R2GEN** (Chen et al., 2020): The model uses ResNet as a visual backbone and generate radiology reports with memory-driven Transformer, where a relational memory is designed to record key information of the generation process and a memory-driven conditional layer normalization is applied to incorporating the memory into the decoder of Transformer. It obtained the state-of-the-art on two radiology report datasets.

B Implementation detail

We adopt the codebase of R2GEN[4] to implement our proposed model. We use DenseNet121 (Huang et al., 2016) pre-trained on CheXpert dataset with 14-class classification setting [5], as the visual backbone to extract visual features with the dimension 1024. For IU X-RAY, the two images are employed to guarantee fair comparison with previous works. In ViLM component, we use the $\mathcal{M}^2$ TR. (Cornia et al., 2020) with 8 attention head, memory size equal to 40, and 3 encoder layers and decoder layers. The model dimension is 512 with the feed forward layers have a dimension of 2048. In LM component, we adapt a pre-trained BART,

[4] https://github.com/cuhksz-nlp/R2Gen

[5] Available in https://nlp.stanford.edu/ysmiura/ifcc/chexpert_auc14.dict.gz

i.e., bart-base[6] for generation of final reports. The model is trained with the Adam optimiser with batch size of 16. The learning rates are set to $5e-5$ and $1e-4$ for the visual extractor and the remaining parameters, respectively. The maximum length in IU X-RAY is set to 60 and in MIMIC-CXR is set to 100. Beam search with beam size of 3 and 5 is used to decode texts during experiments with IU X-RAY and MIMIC-CXR, respectively. The hyper-parameters values are obtained by evaluation of the model with the best BLEU-4 score using the validation set of two benchmark datasets. We train the model using NVIDIA GeForce RTX 2080 Ti for 100 and 30 epochs with early stopping (patience=20) on IU X-RAY and MIMIC-CXR, respectively.

[6] Available in https://huggingface.co/facebook/bart-base

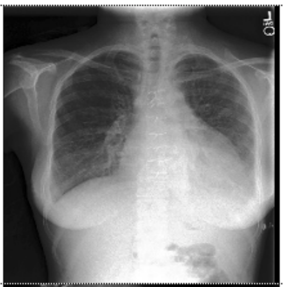
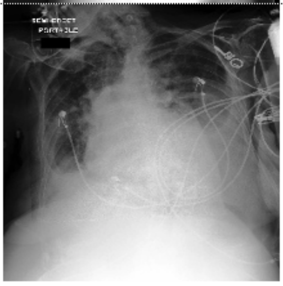
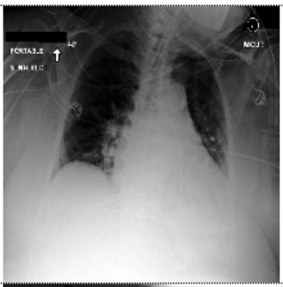
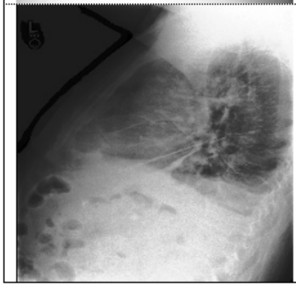
	Reference	$\mathcal{M}^2$ Tr.	$\mathcal{M}^2$ Tr. Progressive
	The patient is status post median sternotomy, CABG, and mitral valve replacement. The heart is mildly enlarged. The mediastinal contours are unchanged with calcification of the aortic knob again noted. Mild pulmonary edema appears progressed compared to the prior exam with small bilateral pleural effusions, also minimally increased compared to the prior exam. Left basilar opacification likely reflects atelectasis. There is no pneumothorax. No acute osseous abnormalities are identified.	Moderate cardiomegaly and mild pulmonary edema are noted. There is a small left pleural effusion. no focal consolidation or pneumothorax is seen.	The patient is status post median sternotomy and mitral valve replacement. Moderate cardiomegaly and mild pulmonary edema are noted. The mediastinal and hilar contours are stable. There is mild pulmonary vascular congestion. There is a small left pleural effusion. There is no focal consolidation or pneumothorax.
	In comparison with the study of ___, there is again enlargement of the cardiac silhouette with extensive bilateral pleural effusions and compressive atelectasis combined with pulmonary vascular congestion.	As compared to the previous the patient has received a nasogastric tube. The course of the tube is the tip of the tube projects over the middle parts of the stomach. No evidence of notably no pneumothorax. The radiograph is unchanged.	In comparison with the study the monitoring and support devices remain in place. Continued enlargement of the cardiac silhouette with bilateral pleural effusions and compressive atelectasis at the bases.
	As compared to the previous radiograph, there is no relevant change. Moderate atelectasis in the retrocardiac lung regions. No new parenchymal opacities. Unchanged position of the endotracheal tube and nasogastric tube.	The mediastinal and hilar contours appear unchanged. The heart size is mildly enlarged. there is no pleural effusion or pneumothorax. The lungs appear clear.	As compared to the previous there is no relevant change. The monitoring and support devices are constant. moderate cardiomegaly with mild pulmonary edema. Retrocardiac atelectasis. No evidence of pneumonia.
	There are small bilateral pleural effusions with fluid extending into the major and minor fissures bilaterally. There is no focal consolidation. Rounded densities projecting over the peripheral right upper lung zone on the AP view may represent pulmonary nodules. There is mild pulmonary vascular congestion interstitial edema. The cardiac silhouette is mild-to-moderately enlarged, but stable. The mediastinal and hilar contours are within normal limits. Partial calcification of the aortic knob is noted.	There is a small bilateral pleural effusions. Bibasilar atelectasis is seen. There is mild pulmonary vascular congestion. The cardiac and mediastinal silhouettes are stable.	There is small bilateral pleural effusions and bibasilar opacities consistent with pulmonary edema. There is mild pulmonary vascular congestion. There is a small left pleural effusion. There is no pneumothorax. There is mild elevation of the left hemidiaphragm.

Figure 2: Illustrations of reports from test dataset as Reference, $\mathcal{M}^2$ Tr. as a baseline model and $\mathcal{M}^2$ Tr. Progressive as a proposed model for selected X-ray chest images. Different colors highlight different medical terms and the detected abnormalities. The text alignments between the reference text and generated one are highlighted with the same colors. Top two images are positive results, the bottom two ones are partial failure cases.