Learning to Generate Clinically Coherent Chest X-Ray Reports

Justin Lovelace
Carnegie Mellon University
jlovelac@andrew.cmu.edu

Bobak Mortazavi
Texas A&M University
bobakm@tamu.edu

Abstract

Automated radiology report generation has the potential to reduce the time clinicians spend manually reviewing radiographs and streamline clinical care. However, past work has shown that typical abstractive methods tend to produce fluent, but clinically incorrect radiology reports. In this work, we develop a radiology report generation model utilizing the transformer architecture that produces superior reports as measured by both standard language generation and clinical coherence metrics compared to competitive baselines. We then develop a method to differentiably extract clinical information from generated reports and utilize this differentiability to fine-tune our model to produce more clinically coherent reports.[1]

1 Introduction

Medical imaging (e.g. chest x-ray) is widely used in medicine for diagnostic purposes. However, current clinical practice requires a radiologist with specialized training to manually evaluate x-rays and note their findings in a radiology report. This manual evaluation is time-consuming and providing an automated solution for this task would help streamline the clinical workflow and improve the quality of care.

Image captioning has gained a large amount of attention following the curation of the COCO dataset (Lin et al., 2014) and the initial image captioning work conducted on it (Vinyals et al., 2014; Xu et al., 2015; Lu et al., 2016a). Although image captioning is a widely studied task, radiology report generation offers unique challenges that precludes the direct adaptation of many image captioning models for the task. For example, much of the recent image captioning research has followed the work of Anderson et al. (2018) and utilized pre-trained objection detection models (Ren et al., 2015) to extract image features (Yu et al., 2019; Huang et al., 2019; Yang et al., 2018; Yao et al., 2018). While this works well for general-domain datasets, analagous pre-trained models are not available in the clinical domain.

Radiology reports are also typically longer and more complex than the captions available in standard image captioning datasets such as COCO. Evaluation of medical report generation is also difficult because the language generation metrics typically used to evaluate image captioning systems can not directly evaluate the descriptive accuracy of generated reports which is of critical importance in the medical domain.

There is already a body of past work that has focused on medical report generation, the most related being that of Boag et al. (2019) and Liu et al. (2019) who both developed fully abstractive techniques for report generation. Boag et al. (2019) benchmarked a number of simple baselines on the MIMIC-CXR dataset (Johnson et al., 2019), the largest publicly available dataset of paired chest x-rays and radiology reports. They observed that typical abstractive methods often produce fluent, but clinically incoherent reports that fail to correctly convey essential information (e.g. the presence of a medical condition).

Liu et al. (2019) attempted to directly addressed this problem by using Self-Critical Sequence Training (SCST) (Rennie et al., 2016) to optimize the clinical accuracy of their generated reports. Although their use of SCST did increase the precision of their model, it also greatly decreased recall and ultimately reduced the F1 score of their model[2]. We also focus on improving the clinical coherence of generated reports in this work.

Liu et al. (2019) also utilized recurrent architecture for report generation despite the recent success of the transformer architecture (Vaswani et al.,

[1] https://github.com/justinlovelace/coherent-xray-report-generation

[2] Liu et al. (2019) did not report F1 in their work but it can be calculated from their reported precision and recall.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1235–1243
November 16 - 20, 2020.

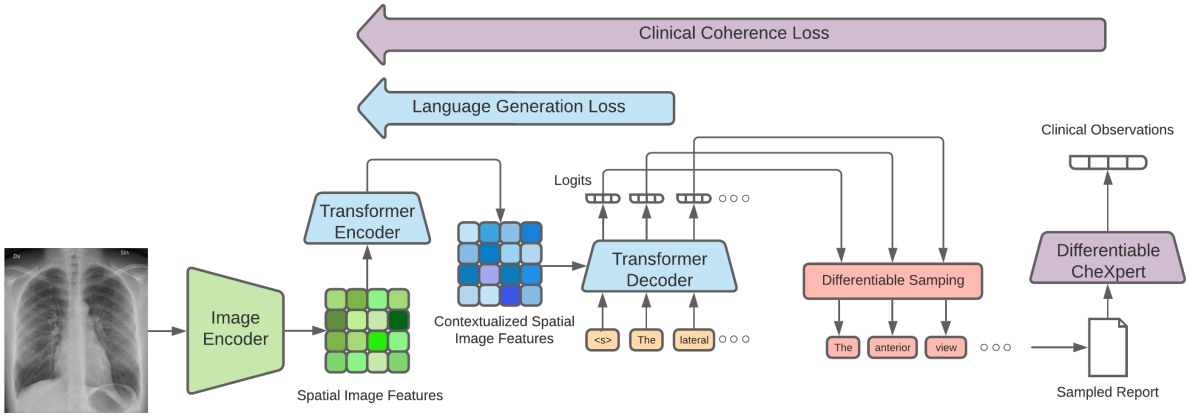

Figure 1: Overview of our proposed framework

2017) with image- and video-captioning tasks outside of the medical domain (Yu et al., 2019; Zhou et al., 2018; Herdade et al., 2019). We also explore whether the transformer architecture is more effective than representative recurrent models for radiology report generation.

Our primary contributions can be summarized as follows: (1) We apply the transformer model to radiology report generation and demonstrate that it is superior to competitive baselines as measured by both language generation metrics and the clinical coherence of the generated reports. (2) We develop a procedure to differentiably extract clinical information from our generated reports and leverage this differentiability to train our report generation model to produce more clinically coherent reports.

2 Dataset

The MIMIC-CXR dataset contains 227,835 imaging reports with 377,110 total images conducted at the Beth Israel Deaconess Medical Center Emergency Department for 65,379 patients (Johnson et al., 2019). The imaging studies are accompanied with free-text radiology reports that record the observations of a practicing radiologist during routine clinical care.

Radiology reports are semi-structured documents composed of a number of possible sections such as *patient history*, *findings*, and *impressions*. We follow the precedent set by previous work and focus on generating the *findings* section because it represents the most direct transcription of the imaging study (Boag et al., 2019; Liu et al., 2019).

We thus constrain our dataset to radiology reports that contain the findings sections and then divide the remaining data into training, validation, and testing sets following a 70%/10%/20% split. We divide the data on the patient ID rather than on specific radiology reports to avoid leaking data from subjects with multiple radiology exams conducted.

3 Methods

We develop an end-to-end report generation framework that consists of two stages. The first stage consists of a report generation model that is trained using a standard language generation objective. For the second stage, we differentiably sample a report from our model and extract the clinical observations from that report. This allows us to introduce an additional learning objective based on the agreement between the observations from the generated and ground truth reports. We utilize this additional objective to fine-tune our model to produce more clinically coherent reports. An overview of this framework is provided in Figure 1. We refer the reader to the supplemental materials for further implementation details.

3.1 Model Architecture

For our report generation architecture we adopt the transformer model introduced by Vaswani et al. (2017) for neural machine translation (NMT). The transformer is an encoder-decoder model where the encoder and decoder both consist of stacked layers of self-attention and position-wise feed-forward neural networks. We refer the reader to Vaswani et al. (2017) for a detailed description of the model.

The primary difference between our setting and that of Vaswani et al. (2017) is that instead of translating a source language to a target language, we must translate an image into a corresponding tex-

tual annotation. Therefore instead of operating on word embeddings, our encoder operates directly on image features.

To extract these features, we apply a pretrained DenseNet-121 model (Huang et al., 2017) to the chest radiographs and extract the final feature matrix before the average pooling layer. We project this feature matrix to the dimensionality d of our model, which provides us with a matrix of spatial features $\mathbf{v} \in \mathbb{R}^{n \times d}$ with n spatial positions. We add a learned positional encoding $\mathbf{P_e} \in \mathbb{R}^{n \times d}$ to the image features to encode spatial information which provides us with the input to our encoder $\mathbf{x_e} = \mathbf{v} + \mathbf{P_e}$.

The decoder is used similarly to its original NMT setting. We pretrain word embeddings for all words that occur at lease 5 times in our training corpus using the continuous-bag-of-words Word2Vec (Mikolov et al., 2013) method. If we let m be the length of a textual report, then the input to our decoder is a sequence of word embeddings $\mathbf{e} \in \mathbb{R}^{m \times d}$ encoded with learned positional embeddings $\mathbf{P_d} \in \mathbb{R}^{m \times d}$ as $\mathbf{x_d} = \mathbf{e} + \mathbf{P_d}$.

We use the standard learned linear transformation followed by the softmax function to convert the feature vectors produced by our decoder to probability distributions for the subsequent word and optimize our model for language generation using the cross-entropy loss function.

3.2 Differentiable CheXpert

To directly train our model to produce clinically accurate reports, we must be able to differentiably extract clinical observations from the generated radiology reports. However, disease labels are typically extracted from radiology reports using a non-differentiable rule-based labeler, CheXpert (Irvin et al., 2019). We develop a differentiable approximation of the CheXpert labeler by training a differentiable model to predict the CheXpert-assigned labels from the reports in our training set.

CheXpert extracts labels for 12 chest x-ray related conditions as well as mentions of support devices. It also has an additional label to represent *no finding*. For each of these 14 label types, it marks the type as either positive, negative, uncertain, or absent. Because positive instances of the conditions are rare, we make the reasonable assumption that an absent label indicates the condition is not present and thus collapse the negative and absent labels to a single label type. Thus we must predict a positive, negative, or uncertain outcome for each of the 14 label types.

We experiment with two model architectures for our differentiable CheXpert, a convolutional neural network (CNN) model and a long short-term memory network (LSTM) model. For our CNN model we apply multiple convolutional filters of varying lengths to the report and utilize a scaled dot-product attention mechanism (Vaswani et al., 2017) to aggregate the feature representations across all spatial positions and convolutional filters. We apply 14 independent attention mechanism for each of the 14 label types extracted by the CheXpert labeler to allow the model to attend to different portions of the narrative for different conditions.

For the LSTM model, we apply a bidirectional LSTM to the report and apply 14 additive attention mechanisms (Bahdanau et al., 2015) to aggregate the output of the LSTM at every position for the 14 label types. For both models, we apply a learned linear transformation and the softmax function to produce a probability distribution over the three possible outcomes for each label type. We train both models using the cross-entropy loss function.

3.3 Differentiable Language Generation

Decoding our model requires sampling discrete tokens from continuous probability distributions which is a non-differentiable operation. To overcome this, we utilize the Gumbel-Softmax trick introduced by Jang et al. (2017); Maddison et al. (2017) to enable differentiable sampling.

This trick utilizes the softmax function as a continuous, differentiable approximation of the argmax operation. If we have k tokens in our vocabulary, then at any given position our model produces probabilities $\{\pi_i\}_{i=1}^k$ over the entire vocabulary. We then sample $\{g_i\}_{i=1}^k$ as independently and identically distributed samples drawn from Gumbel(0,1)[3] and compute the sampled vector $y \in \mathbb{R}^k$ with $y_i = \frac{\exp((\log(\pi_i)+\beta g_i)/\tau)}{\sum_{j=1}^k \exp((\log(\pi_j)+\beta g_j)/\tau)}$ where β controls the magnitude of the noise and τ controls how closely the function approximates the argmax operation [4].

3.4 Fine-Tuning Procedure

We first train a report generation model using the standard natural language generation (NLG) objective $\mathcal{L}_{NLG}$ and then further fine-tune the model

[3]The Gumbel distribution can be sampled by drawing $u \sim$ Uniform(0,1) and computing $g = -\log(-\log(u))$.

[4]The function becomes equivalent to the argmax operation as $\tau \to 0$.

Table 1: Model Performance (Language Generation)

Model	CIDER	METEOR	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1-NN	12.5	13.9	22.8	36.7	21.5	13.8	9.5
SA&T	27.8	14.1	31.0	37.0	24.0	17.0	12.8
AdpAtt	29.9	14.8	31.4	38.4	25.1	17.8	13.4
Transformer	**31.8**	15.7	**31.8**	40.9	26.8	19.1	14.4
Transformer w/ Fine-Tuning	31.6	**15.9**	**31.8**	**41.5**	**27.2**	**19.3**	**14.6**

Table 2: Model Performance (Clinical Coherence)

Model	Micro-Avg			Macro-Avg		
	F1	Prec	Rec	F1	Prec	Rec
1-NN	33.5	34.6	32.4	20.6	21.3	20.0
SA&T	28.2	36.4	23.0	10.1	24.7	11.9
AdpAtt	34.7	41.7	29.8	16.3	**34.1**	16.6
Transformer	39.8	46.1	35.0	21.4	32.7	20.4
Transformer w/ Fine-Tuning	**41.1**	**47.5**	**36.1**	**22.8**	33.3	**21.7**

to be more clinically coherent by introducing an additional clinical coherence objective. We do so by applying the Gumbel-Softmax trick to differentiably sample tokens from our decoder and then apply our differentiable CheXpert to the sampled report. This allows us to introduce a second training objective measuring the agreement between the ground truth CheXpert labels and the labels obtained by applying our differentiable CheXpert to the sampled report. This is implemented using the cross-entropy loss function and we denote this clinical coherence loss $\mathcal{L}_{CC}$. We define the final training objective during this fine-tuning stage as $\mathcal{L} = \lambda\mathcal{L}_{CC} + \mathcal{L}_{NLG}$ where λ is a hyperparameter that determines the balance between the NLG and clinical coherence objectives.

4 Results

We report the CIDER (Vedantam et al., 2014), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002) of our report generation models. While these metrics all measure the language similarity between the the generated and ground truth reports, they can not directly evaluate how effectively the models are producing clinically correct reports. To address this shortcoming, we also report clinical coherence metrics that compare the CheXpert extracted labels for the generated and ground truth reports. For this, we report the macro- and micro-averaged precision, recall, and F1 score for positive annotations. We refer the reader to the supplementary materials to view the results for the individual observations. All models in this work are decoded using beam search with a beam size of 4.

4.1 Baselines

Boag et al. (2019) benchmarked a number of report generation techniques on the MIMIC-CXR dataset and found a simple 1-Nearest Neighbor baseline to be surprisingly effective, particularly with respect to its clinical coherence. As such, we compare against a 1-NN baseline where we retrieve the report from our training set whose DesneNet-induced image features have the highest cosine similarity with the test query image. We also compare against two competitive recurrent image captioning models: Show, Attend, and Tell developed by Xu et al. (2015) and Adaptive Attention developed by Lu et al. (2016b).

4.2 Effect of Model Architecture

We first conduct an experiment to evaluate the effect of using our proposed transformer model by training it using only the standard language generation objective. We report language generation metrics for this experiment in Table 1. We observe that the transformer model offers significant improvements across all metrics compared to the baselines.

We report the metrics for the clinical coherence of our models in Table 2. We observe that our proposed transformer model improves upon the micro-averaged and macro-averaged F1 of our best baselines by 5.1 and 0.8 points respectively.

Table 3: Differentiable CheXpert Performance

Model	Micro-Avg			Macro-Avg		
	F1	Prec	Rec	F1	Prec	Rec
CNN	93.6	92.8	94.5	90.0	89.4	90.8
LSTM	**98.1**	**98.2**	**98.0**	**97.1**	**97.4**	**96.7**

4.3 Differentiable CheXpert

We report the effectiveness of our CNN and LSTM CheXpert models in Table 3. We observe that the LSTM model significantly outperforms the CNN model across all metrics. This is unsurprising because recognizing things like negations can involve identifying long-term relationships in the report that would be difficult for an n-gram model like the CNN to recognize. Because of this finding, we use the LSTM model as our differentiable approximation of CheXpert during the fine-tuning stage.

4.4 Clinical Coherence Fine-Tuning

We fine-tune our transformer model to improve its clinical coherence and report results for this comparison in Tables 1 and 2. We observe that the model still has comparable NLG performance and improves the micro-averaged and macro-averaged F1 of our model by 1.3 and 1.4 points respectively. We conduct McNemar's test (McNemar, 1947; Edwards, 1948) and find that the improvement is statistically significant ($p < 1 \times 10^{-10}$).

5 Limitations and Future Work

Past work on radiology report generation has leveraged the semi-standardized nature of radiology reports to develop extractive or template-based methods (Zhang et al., 2018; Han et al., 2018; Gale et al., 2019). Other work has combined template-based methods with abstractive methods to utilize the advantages of both methods (Li et al., 2018; Biswal et al., 2020). In this work we focused on developing abstractive techniques as was done by past work on the MIMIC-CXR dataset (Liu et al., 2019; Boag et al., 2019). However, in the future we intend to combine the abstractive methods developed in this work with retrieval methods to further improve upon our framework.

Wang et al. (2018) developed a model that jointly extracted conditions from chest x-rays and generated radiology reports. In this work we only focus on report generation, but augmenting our framework with an explicit image classification objective is a potential direction for future work.

While the methods developed in this work lead to significantly improved performance compared to competitive baselines, the clinical coherence of our model is still insufficient for clinical practice. More work must be conducted in the future to continue improving the clinical coherence of automated report generation to enable adoption of such methods.

6 Conclusion

In this work we develop a radiology report generation model utilizing the transformer architecture and demonstrate that it is both more fluent and clinically coherent than competitive baselines. We also develop a procedure to differentiably extract clinical information from generated reports and utilize this differentiability to further fine-tune our model for clinical coherence. Our proposed architecture and fine-tuning procedure improve the micro-averaged and macro-averaged F1 of our best baselines by 6.4 and 2.2 points respectively while achieving superior fluency as measured by all of our computed NLG metrics.

References

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Siddharth Biswal, Cao Xiao, Lucas Glass, Brandon Westover, and Jimeng Sun. 2020. Clinical report auto-completion. In *Proceedings of The Web Conference 2020*, WWW '20, page 541–550, New York, NY, USA. Association for Computing Machinery.

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alsentzer, and Peter Szolovits. 2019. Baselines for chest x-ray report generation. In *Machine Learning for Health Workshop at NeurIPS*.

Allen L. Edwards. 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3):185–187.

W. Gale, L. Oakden-Rayner, G. Carneiro, L. J. Palmer, and A. P. Bradley. 2019. Producing radiologist-quality reports for interpretable deep learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1275–1279.

Zhongyi Han, Benzheng Wei, Stephanie Leung, Jonathan Chung, and Shuo Li. 2018. Towards automatic report generation in spine radiology using weakly supervised framework. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 185–193, Cham. Springer International Publishing.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11137–11147. Curran Associates, Inc.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. 2019. Adaptively aligned image captioning via adaptive attention time. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8942–8951. Curran Associates, Inc.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

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 de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1530–1540. Curran Associates, Inc.

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.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

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 *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269, Ann Arbor, Michigan. PMLR.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016a. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *CoRR*, abs/1612.01887.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016b. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *CoRR*, abs/1612.01887.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. *CoRR*, abs/1612.00563.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9049–9058.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2018. Auto-encoding scene graphs for image captioning. *CoRR*, abs/1812.02378.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. *CoRR*, abs/1809.07041.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *CoRR*, abs/1905.07841.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. *CoRR*, abs/1809.04698.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

A Preprocessing

We extract the findings section from the radiology reports and then utilize spaCy for tokenization. We learn 256 dimension word embeddings using the continuous-bag-of-words Word2Vec method (Řehůřek and Sojka, 2010; Mikolov et al., 2013) for all words that appear at least 5 times in our training set. This leaves us with a vocabulart of 3,913 words. We replace out of vocabulary tokens with a special `<unk>` token that we initialze from a standard Gaussian $(\mu = 0, \sigma^2 = 1)$

We rescale the chest radiographs to a 256×256 image before feeding it to the pretrained DenseNet-121 model. This produces an $8 \times 8 \times 1024$ feature matrix which is the final radiograph representation used as the input to our models.

B Dataset Statistics

After constraining our dataset to reports with a findings section, we are left with 265,259 chest radiographs and 149,459 radiology reports. We report full dataset statistics in Table 4.

C Implementation Details

C.1 Report Generation

We train all of the models used in this work for 64 epochs and anneal the learning rate by a factor of 0.5 every 16 epochs. We train our models using the Adam (Kingma and Ba, 2014) optimizer with a batch size of 32 and tune the initial learning rate independently for each model based on validation performance. We regularize our model using dropout and use gradient clipping to prevent exploding gradients. We evaluate the model with the best BLEU-4 score on the validation set upon the test set. Because conducting beam search upon the validation set after every epoch would greatly increase training time, we generate reports using greedy decoding with teacher forcing and find this to be an effective stopping criterion.

For the fine-tuning procedure, we load the model with the best validation performance and train it for 8 additional epochs with the modified learning objective described in section 3.4. We set $\tau = 1, \beta = 1, \lambda = 0.9$ for this training stage where τ, β are the sampling hyperparameters introduced in section 3.3 and λ is the loss hyperparameter introduced in section 3.4. We utilize the performance of the differentiable CheXpert model upon

Table 4: Dataset Statistics

Category	Total	Training	Validation	Testing
Total Chest X-Rays	265,259	187,071	26,005	52,183
Atelectasis	54,289 / 199,486 / 11,484	38,201 / 140,744 / 8,126	5,368 / 19,556 / 1,081	10,720 / 39,186 / 2,277
Cardiomegaly	64,188 / 195,525 / 5,546	45,440 / 137,687 / 3,944	6,132 / 19,362 / 511	12,616 / 38,476 / 1,091
Consolidation	8,113 / 249,071 / 8,075	5,765 / 175,743 / 5,563	804 / 24,369 / 832	1,544 / 48,959 / 1,680
Edema	19,374 / 232,188 / 13,697	13,813 / 163,585 / 9,673	1,917 / 22,776 / 1,312	3,644 / 45,827 / 2,712
Enlarged Cardiomediastinum	21,851 / 210,373 / 33,035	15,416 / 148,244 / 23,411	2,110 / 20,578 / 3,317	4,325 / 41,551 / 6,307
Fracture	10,645 / 254,011 / 603	7,418 / 179,222 / 431	1,011 / 24,936 / 58	2,216 / 49,853 / 114
Lung Lesion	9,815 / 254,127 / 1,317	6,784 / 179,342 / 945	937 / 24,943 / 125	2,094 / 49,842 / 247
Lung Opacity	79,037 / 182,324 / 3,898	55,861 / 128,420 / 2,790	7,569 / 18,085 / 351	15,607 / 35,819 / 757
No Finding	69,159 / 196,100 / 0	48,640 / 138,431 / 0	6,983 / 19,022 / 0	13,536 / 38,647 / 0
Pleural Effusion	47,449 / 208,363 / 9,447	33,963 / 146,487 / 6,621	4,574 / 20,479 / 952	8,912 / 41,397 / 1,874
Pleural Other	4,925 / 258,891 / 1,443	3,360 / 182,656 / 1,055	521 / 25,357 / 127	1,044 / 50,878 / 261
Pneumonia	10,116 / 234,864 / 20,279	7,098 / 165,578 / 14,395	1,018 / 23,010 / 1,977	2,000 / 46,276 / 3,907
Pneumothorax	8,188 / 254,206 / 2,865	5,905 / 179,185 / 1,981	752 / 24,960 / 293	1,531 / 50,061 / 591
Support Devices	68,858 / 196,071 / 330	48,610 / 138,224 / 237	6,646 / 19,325 / 34	13,602 / 38,522 / 59

We report pos/neg/unc labels for each CheXpert category.

the greedily decoded validation reports to define our stopping criterion for the fine-tuning stage.

Our transformer model has 8 attention heads, 1 encoder layer, and 6 decoder layers. The model dimension is $d = 256$, the dimensionality of the word embeddings, and the feed forward layers have an intermediate dimension of 4096. For our baselines we adapt publicly available implementations for the Show, Attend, and Tell and Adaptive Attention models and utilize the same training schedule as our transformer model. We compute our language generation metrics using the publicly available nlg-eval (Sharma et al., 2017).

C.2 Differentiable CheXpert

For our differentiable CheXpert models, we train the models for a maximum of 64 epochs with a learning rate of 5×10^{-4} using the Adam optimizer and a batch size of 128. We utilize the micro-averaged F1 score for early stopping and terminate training if the validation performance has not improved for 10 epochs. We then evaluate the model with the best validation performance upon the test set.

For our CNN model, we used 4 convolutional filters of lengths 3, 5, 7, and 9. The output dimensionality of the convolutional filters was set to 64. For our LSTM model, we utilize a bidirectional LSTM with a hidden dimension of 128. For each model, we apply 14 independent attention mechanisms corresponding to the 14 label types to allow the model to attend to different sections of the narrative for different conditions. We then project the aggregations induced by the attention mechanism to 3, the number of label types, and then apply the softmax function to produce the final prediction.

The hyperparameters for all of the models used in this work were manually tuned based on validation performance. All training was done using a single NVIDIA GeForce GTX 1080 Ti.

D Supplemental Results

We report detailed results across all CheXpert categories for positive mentions in Table 5 and for uncertain mentions in Table 6[5].

[5]The CheXpert labeler does not produce uncertain mentions for the 'No Findings' category so we report results for the 13 valid categories.

Table 5: Detailed Clinical Coherence of Report Generation Models (Positive Mentions)

Category	1-NN			SA&T			AdpAtt			Transformer			Transformer w Fine-Tuning		
	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec
Atelectasis	29.1	30.0	**28.2**	3.2	40.6	1.7	13.6	43.2	8.1	29.2	**43.2**	22.0	**32.2**	43.0	25.8
Cardiomegaly	35.4	36.9	34.0	28.0	40.9	21.3	36.4	40.4	33.1	40.9	44.1	38.0	**43.3**	**46.9**	**40.2**
Consolidation	4.2	4.5	3.9	0.0	0.0	0.0	0.6	9.4	0.3	**8.1**	**15.8**	**5.4**	7.3	15.7	4.8
Edema	18.0	19.0	17.1	2.6	**44.0**	1.3	17.9	32.6	12.4	25.2	40.7	18.2	**29.8**	37.6	**24.6**
Enlarged Cardiomediastinum	**9.5**	10.1	**8.9**	0.0	6.3	0.0	2.0	8.7	1.2	3.7	10.5	2.3	5.9	**12.3**	3.9
Fracture	**4.8**	**5.3**	**4.4**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lung Lesion	**3.8**	4.4	**3.4**	0.0	0.0	0.0	0.1	0.0	0.0	1.7	**28.6**	0.9	1.4	23.8	0.7
Lung Opacity	**35.4**	37.3	**33.7**	0.5	44.0	0.3	4.4	56.0	2.3	16.7	61.0	9.7	17.1	**64.0**	9.9
No Finding	46.3	**42.7**	50.7	45.7	30.1	94.8	49.1	33.7	90.5	52.2	36.8	**89.8**	**54.1**	39.0	88.2
Pleural Effusion	37.9	39.9	36.1	5.3	60.9	2.8	32.1	69.4	20.8	**48.4**	69.5	**37.2**	48.0	**71.2**	36.2
Pleural Other	**2.3**	**2.7**	**2.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.2	6.7	0.1	0.9	**16.1**	0.5
Pneumonia	4.0	4.1	3.9	0.0	0.0	0.0	2.3	5.0	1.5	**5.1**	6.7	**4.2**	3.9	**7.0**	2.7
Pneumothorax	8.3	9.0	7.6	6.1	7.1	5.4	5.8	5.7	5.9	3.6	**16.0**	2.0	**9.8**	12.9	**7.8**
Support Devices	48.9	52.9	45.5	50.6	71.3	39.2	63.7	73.1	56.5	64.9	**78.1**	55.6	**66.0**	77.0	**57.8**
Macro-Average	20.6	21.3	20.0	10.1	24.7	11.9	16.3	**34.1**	16.6	21.4	32.7	20.4	**22.8**	33.3	**21.7**
Micro-Average	33.5	34.6	32.4	28.2	36.4	23.0	34.7	41.7	29.8	39.8	46.1	35.0	**41.1**	**47.5**	**36.1**

Table 6: Detailed Clinical Coherence of Report Generation Models (Uncertain Mentions)

Category	1-NN			SA&T			AdpAtt			Transformer			Transformer w Fine-Tuning		
	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec
Atelectasis	0.0	0.0	0.0	0.8	12.2	0.4	**6.1**	6.3	**6.0**	2.3	10.7	1.3	1.9	**14.5**	1.0
Cardiomegaly	0.0	0.0	0.0	1.2	4.1	0.7	**3.0**	3.2	**2.8**	1.8	3.9	1.2	2.1	**6.4**	1.3
Consolidation	0.2	6.9	0.1	2.3	6.9	1.4	3.7	4.1	3.3	6.3	8.3	5.1	**6.8**	**9.0**	**5.4**
Edema	3.6	13.0	2.1	4.2	10.5	2.6	9.5	10.0	9.1	**13.3**	16.2	**11.3**	11.5	**16.7**	8.8
Enlarged Cardiomediastinum	10.5	18.8	7.3	14.1	20.5	10.7	14.8	15.4	14.2	17.4	20.5	15.2	**21.1**	**21.0**	**21.1**
Fracture	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lung Lesion	0.0	0.0	0.0	0.0	0.0	0.0	**0.4**	**0.5**	**0.4**	0.0	0.0	0.0	0.0	0.0	0.0
Lung Opacity	0.0	0.0	0.0	1.2	2.4	0.8	1.7	1.7	1.7	1.6	**2.6**	1.2	**2.4**	2.5	**2.4**
Pleural Effusion	0.6	**21.4**	0.3	1.3	8.4	0.7	4.4	4.6	4.1	**6.8**	12.6	**4.6**	5.6	9.7	3.9
Pleural Other	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	0.8	0.6	1.7	0.4	**2.0**	**1.6**	**2.7**
Pneumonia	0.4	6.1	0.2	1.1	15.8	0.6	**8.7**	9.0	**8.5**	7.1	19.7	4.3	8.3	**20.7**	5.2
Pneumothorax	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.8	0.7	0.0	0.0	0.0	0.0	0.0	0.0
Support Devices	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Macro-Average	4.1	4.3	**4.0**	1.2	5.1	0.8	2.0	6.2	1.4	4.4	7.4	3.4	**4.8**	**7.9**	**4.0**
Micro-Average	8.5	8.9	8.2	4.3	**17.3**	2.4	6.1	15.9	3.8	10.3	16.3	7.6	**11.8**	16.6	**9.1**