Improving the Factual Correctness of Radiology Report Generation with Semantic Graph-Based Rewards

Anonymous ACL submission

Abstract

Neural image-to-text radiology report generation systems offer the potential to improve radiology reporting by reducing the repetitive process of report drafting and identifying possible medical errors. These systems have achieved promising performance as measured by widely used NLG metrics such as BLEU and CIDEr. However, the current systems face important limitations. First, they present an increased complexity in architecture that offers only marginal improvements on NLG metrics. Secondly, these systems that achieve high performance on these metrics are not always factually complete or consistent due to both inadequate training and evaluation. Recent studies have shown the systems can be substantially improved by using new methods encouraging 1) the generation of domain entities consistent with the reference and 2) describing these entities in inferentially consistent ways. So far, these methods rely on weakly-supervised approaches (rule-based) and named entity recognition systems that are not specific to the chest X-ray domain. To overcome this limitation, we propose a new method, the RadGraph reward, to further improve the factual completeness and correctness of generated radiology reports. More precisely, we leverage the Rad-Graph dataset containing annotated chest X-ray reports with entities and relations between entities. On two open radiology report datasets, our system substantially improves the scores up to 14.2% and 25.3% on metrics evaluating the factual correctness and completeness of reports.

1 Introduction

An important medical application of natural language generation (NLG) is to build assistive systems that take X-ray images of a patient and generate a textual report describing clinical observations in the images (Jing et al., 2018; Li et al., 2018; Chen et al., 2020; Miura et al., 2021). This is
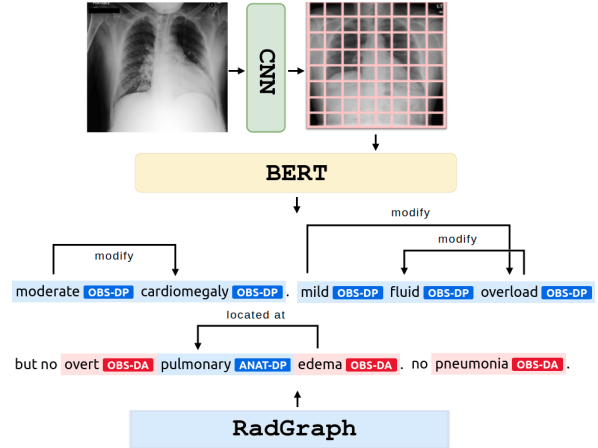

Figure 1: Overview of our radiology report generation pipeline. First, a neural network generates a radiology report given a chest X-ray image. We then leverage RadGraph to create semantic annotations of the output used to design reinforcement learning rewards.

a clinically important task, offering the potential to reduce radiologists' repetitive work and generally improve clinical communication (Kahn Jr et al., 2009).

Recently, a lot of attention has been given to new architectures (Chen et al., 2020, 2021; Alfarghaly et al., 2021) and how the structure of data could be input into the system (Liu et al., 2021a). These systems have achieved promising performance as measured by widely used NLG metrics such as BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015). However, these studies face important limitations. First, they present an increased complexity in architecture that offers only marginal improvements on NLG metrics. Secondly, these systems that achieve high performance on NLG metrics are not always factually complete or consistent due to both inadequate training and evaluation of these systems. Miura et al. (2021) have shown that existing systems are inadequate in factual completeness and consistency, and

1

that an image-to-text radiology report generation (RRG) system can be substantially improved by replacing the widely used NLG metrics with "factually-oriented" methods encouraging 1) the generation of domain entities consistent with the reference and 2) describing these entities in inferentially consistent ways. So far, these new methods rely on weakly-supervised approaches (rule-based) to construct NLI models for radiology reports and biomedical named entity recognition systems (Zhang et al., 2021) that are not specific to chest X-rays.

Despite these "factually-oriented" methods being weakly supervised or being limited to generic biomedical entities, their use showed substantial improvements on a wide range of metrics and board-certified radiologists' evaluations. These findings motivate us to propose a new method to further improve the factual completeness and correctness of generated radiology reports. More precisely, we leverage RadGraph (Jain et al., 2021), a dataset annotated by radiologists containing chest X-ray radiology reports along with annotated entities and relations. These annotations allow us to create two semantic graphs, one for the generated and one for the reference report. We then introduce three simple rewards that score the differences between the two graphs in terms of entities and relations. These rewards can be directly optimized using Reinforcement Learning (RL) to further improve the quality of the generated report by our systems. By doing so, we show on two popular chest X-ray datasets that our models are able to maximize the defined rewards but also outperform the previous works on various NLG and factually-oriented metrics.

In summary our contributions are:

- We propose a simple RRG architecture that 1) is fast to train and suitable for a RL setup and 2) performs equally well as the previous and more complex architectures proposed in the literature.

- We leverage the RadGraph dataset and the associated fine-tuned model to design semantic-based rewards that qualitatively evaluate the factual correctness and completeness of the generated reports.

- We show on two datasets that directly optimizing these rewards outperforms previous approaches that prioritize traditional NLG metrics.

The paper is structured as follows: first, we describe our factually-oriented graph-based rewards (§2). More precisely, we begin by examining the RadGraph dataset (§ 2.1) and how we leveraged the annotations to create our rewards (§ 2.2). Then, we explain the architecture of the model (§ 3) that we used to generate reports and how we trained it using negative log-likelihood (NLL) and RL. The sections that follow afterwards are dedicated to the datasets used for the experiments (§ 4) and the metrics (§ 5) chosen to evaluate the generation of reports. This latter section is divided into two groups: the classic NLG metrics (§ 5.1) and the factually-oriented metrics (§ 5.2). Finally, we present the results (§ 6) and end this paper with a section addressing related works (§ 7).

2 Factually-oriented Graph-Based Reward

In this section, we present a new semantic graph-based reward, called the RadGraph reward, used throughout our experiments. We first start by explaining in Section 2.1 the RadGraph dataset and how we get the annotations that shape our reward. In Section 2.2, we explain how we construct the RadGraph reward and its different variants.
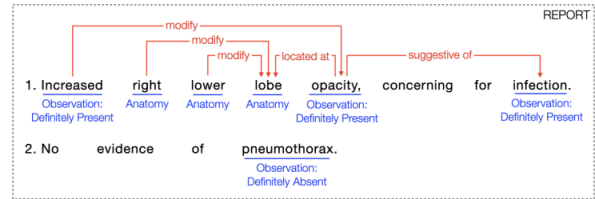
2.1 RadGraph

Figure 2: An example of a report annotated with entities and relations in the RadGraph dataset

RadGraph (Jain et al., 2021) is a dataset of entities and relations in full-text chest X-ray radiology reports based on a novel information extraction schema designed to structure radiology reports. The dataset contains board-certified radiologist annotations of 500 radiology reports from the MIMIC-CXR dataset (Johnson et al., 2019), which correspond in total to 14,579 entities and 10,889 relations. In addition, RadGraph also includes a test dataset of 100 radiology reports, split between two independent sets of board-certified

radiologist annotations on reports from MIMIC-CXR and CheXpert (Smit et al., 2020) datasets (50 reports each).

Entities An entity is defined as a continuous span of text that can include one or more adjacent words. Entities in RadGraph center around two concepts: *Anatomy* and *Observation*. Three uncertainty levels exist for *Observation*, leading to four different entities: *Anatomy* (*ANAT-DP*), *Observation: Definitely Present* (*OBS-DP*), *Observation: Uncertain* (*OBS-U*), and *Observation: Definitely Absent* (*OBS-DA*). *Anatomy* refers to an anatomical body part that occurs in a radiology report, such as a "lung". *Observation* refers to words associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications. As an example, an *Observation* could be "effusion" or more general phrases like "increased".

Relations A relation is defined as a directed edge between two entities. Three levels exist: *Suggestive Of (., .)*, *Located At (., .)*, and *Modify (., .)*. *Suggestive Of (Observation, Observation)* is a relation between two *Observation* entities indicating that the presence of the second *Observation* is inferred from that of the first *Observation*. *Located At (Observation, Anatomy)* is a relation between an *Observation* entity and an *Anatomy* entity indicating that the *Observation* is related to the *Anatomy*. While *Located At* often refers to location, it can also be used to describe other relations between an *Observation* and an *Anatomy*, such as shape or color. *Modify (Observation, Observation)* or *Modify (Anatomy, Anatomy)* is a relation between two *Observation* entities or two *Anatomy* entities indicating that the first entity modifies the scope of, or quantifies the degree of, the second entity.

The authors also released a PubMedBERT (Gu et al., 2021) model fine-tuned on the RadGraph dataset. We leverage this trained model to create the annotations for the datasets used in our experiments. We will refer to this model as RadGraph model in what follows.

2.2 RadGraph reward

Using RadGraph annotation scheme and model, we design F-score style rewards that measure consistency and completeness of generated radiology reports compared to reference reports. Each of our rewards leverages the outputs of
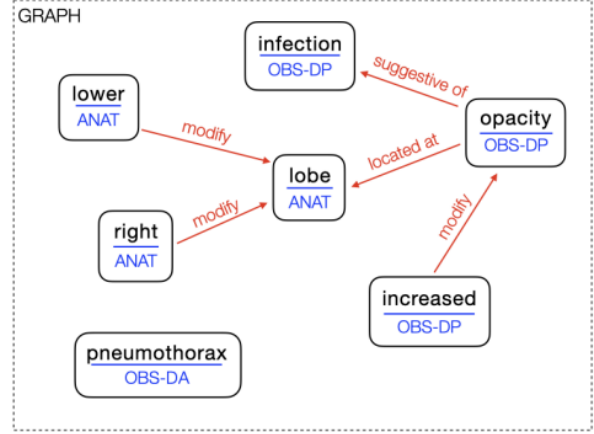

Figure 3: Graph view of the RadGraph annotations for the report in Figure 2.

the released fine-tuned PubMedBERT model on RadGraph, namely the entities and the relations, on both a generated report and its reference.

The RadGraph annotations of a report can be represented as a graph $\mathcal{G}(V, E)$ with the set of nodes $V = \{v_1, v_2, \ldots, v_{|V|}\}$ containing the entities and the set of edges $E = \{e_1, e_2, \ldots, e_{|E|}\}$ the relations between pairs of entities. The graph is directed, meaning that the edge $e = (v_1, v_2) \neq (v_2, v_1)$. An example is depicted in Figure 3. Each node or edge of the graph also has a label, which we denote as v_{i_L} for an entity i (for example "OBS-DP" or "ANAT") and e_{ij_L} for a relation $e = (v_i, v_j)$ (such as "modified" or "located at"). We now proceed to describe three of our rewards.

RG$_\text{E}$ This reward focuses only on the nodes V. For the generated report y, we create a new set of node-label pairs $\bar{V}_y = \{(v_i, v_{i_L})\}_{i \in [1..|V|]}$ comprising all entities and their corresponding labels. We proceed to construct the same set for the reference report $\hat{y}$ and denote this set $\bar{V}_{\hat{y}}$.

RG$_\text{ER}$ This reward focuses on the nodes V and whether or not a node has a relation in E. For the generated report y, we create a new set of triplets $\bar{V}_y = \{(v_i, v_{i_L}, \epsilon_i)\}_{i \in [1..|V|]}$. The value of ϵ_i is 1 if v_i has a relation in E else 0. We proceed to construct the same set for the reference report $\hat{y}$ and denote this set $\bar{V}_{\hat{y}}$.

RG$_\overline{\text{ER}}$ This reward focuses on the nodes V and their relations in E. For the gener-

3

ated report y, we create a new set of tuples $\bar{V}_y = \{(v_i, v_{i_L}, (v_i, v_j), e_{ij_L}) \mid i \in [1..|V|], j \in [1..|V|], j \neq i, (v_i, v_j) \in E\}$. In addition, for all the nodes v_i with no relations, we include a tuple (v_i, v_{i_L}) in $\bar{V}_y$. We proceed to construct the same set for the reference report $\hat{y}$ and denote this set $\bar{V}_{\hat{y}}$.

Finally, RG_E, RG_{ER} and $RG_{\overline{ER}}$ are defined as the harmonic mean of precision and recall between their respective sets $\bar{V}_{\hat{y}}$ and $\bar{V}_y$.

As an illustration, we provide in Appendix B the set $\bar{V}$ of the graph $\mathcal{G}$ in Figure 3.

3 Model

Architecture   To encode an X-ray image, we extract convolutional visual features $\boldsymbol{D}$ of size 49×1024 using a Densenet-121 (Huang et al., 2017). To generate language, we use a one-layered BERT (Vaswani et al., 2017; Devlin et al., 2019) with cross-attention over the visual features. More formally, the cross-attention of the transformer layer is written:

$$\text{Cross-Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d}})V \quad (1)$$

where Q is the BERT hidden state of size d and K and V are the visual features $\boldsymbol{D}$. The full detail of the model can be found at Appendix D.

Training   If we denote $\boldsymbol{\theta}$ as the model parameters, then $\boldsymbol{\theta}$ is learned by maximizing the likelihood of the observed report $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n)$ or in other words by minimizing the negative log-likelihood. The objective function is given by:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{t=1}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_t | \boldsymbol{y} < t, \boldsymbol{D}) \quad (2)$$

After the NLL training, we start a RL training that optimizes one of our factually-oriented rewards. The loss function in equation 2 is now given by:

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{Y} \sim p_{\boldsymbol{\theta}}} \boldsymbol{r}(\boldsymbol{Y}) \quad (3)$$

where $\boldsymbol{r}(\boldsymbol{Y})$ is the reward given to the generated report. We use the SCST algorithm (Rennie et al., 2017) to approximate the expected gradient of our non-differentiable reward function. The expression becomes:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \approx -(\boldsymbol{r}(\boldsymbol{Y}) - \boldsymbol{r}(\bar{\boldsymbol{Y}})) \nabla_{\boldsymbol{\theta}} \log_{p_{\boldsymbol{\theta}}}(\boldsymbol{Y}) \quad (4)$$

Here $\boldsymbol{r}(\bar{\boldsymbol{Y}})$ acts as a baseline (Sutton et al., 1998) to reduce the variance of $\boldsymbol{r}(\boldsymbol{Y})$. In our case, $\boldsymbol{r}(\bar{\boldsymbol{Y}})$ is the expected reward by sampling from the model during training.

Hyper-parameters   Our model consists of 1 Transformer block of size 768 with a feed-forward layer of size 3072. As optimizer, we pick Adam (Kingma and Ba, 2015) with a learning rate of $3e^{-4}$ and mini-batch size of 128. We decode with a beam-search of size 3.

4 Datasets

To carry out our experiments, we use two chest X-ray datasets: MIMIC-CXR (Johnson et al., 2019) and Open-i Chest X-ray (Indiana University, Demner-Fushman et al. (2012)). For both datasets, we use the official splits but we discard the reports that do not contain a *Findings* section. MIMIC-CXR thus consists of $152,173$ training samples, $1,196$ for validation and $2,347$ for testing; similarly, the Indiana dataset, originally containing $5,935$ training images, 740 for validation and 740 for testing, consists of only $3,335$ reports in total if we do not count the multiplicity of images in each study and if we discard the reports without *Findings*. Since this renders the dataset too small to both train and test a RRG model, Open-i dataset is only used for testing purposes.

For each sample of both datasets, we generate what we consider the ground-truth radiology diagnostic by running the CheXbert labeler (Smit et al., 2020) on the ground-truth *Findings* section. This creates for each ground-truth report the associated diagnostic label, which describe for 14 possible observations the degree of presence (e.g. consolidation or edema for which the report can be positive, negative, uncertain or unspecified). This label is used to compute the F1 CheXbert score (see Section 5.2).

5 Metrics

In this section, we proceed to report all the metrics used to evaluate the *Findings* generated by our model against the human-redacted *Findings*. We divided our metrics into two categories, the NLG metrics as widely reported in the NLG literature, and the factually-oriented metrics, specific to the

4

evaluation of factual correctness and completeness of radiology reports.

5.1 NLG-oriented

We report the BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) metrics to evaluate the generations. To be consistent with previous work (Chen et al., 2021), we also report the ROUGE-L metric (Lin, 2004).

5.2 Factually-oriented

The following presented metrics evaluate the factual correctness and completeness of the generated *Findings* in different ways. We proceed to describe them and their differences.

fact$_{ENT}$ (Miura et al., 2021)   A named entity recognizer (NER) is applied to the generated report $\hat{y}$ and the corresponding reference y, giving respectively two sets of extracted entities $\mathbb{E}_{\hat{y}}$ and $\mathbb{E}_y$. fact$_{ENT}$ is defined as the harmonic mean of precision and recall between the two sets $\mathbb{E}_{\hat{y}}$ and $\mathbb{E}_y$. The clinical model of Stanza (Qi et al., 2020) is used as NER.

fact$_{ENTNLI}$ (Miura et al., 2021)   This score is an extension of **fact$_{ENT}$** with Natural Language Inference (NLI). Here, an entity e of $\mathbb{E}_{\hat{y}}$ is not automatically considered correct if present in $\mathbb{E}_y$. To be considered valid, the sentence $s_{\hat{y}}$ containing entity e must not present a contradiction with its counterpart sentence s_y in the reference report. The counterpart sentence s_y in the reference report is the sentence with the highest BERTScore (Zhang et al., 2019) against $s_{\hat{y}}$. The NLI model outputs whether sentence s_y is a contradiction of $s_{\hat{y}}$. We use the NLI model weights of Miura et al. (2021), which relies on a BERT-architecture.

F$_1$CheXbert (Zhang et al., 2020)   This score uses CheXbert (Smit et al., 2020), a Transformer-based model trained to output abnormalities (fourteen classes) of chest X-rays given a radiology report as input.   F$_1$CheXbert is the F1-score between the prediction of CheXbert over the generated report $\hat{y}$ and the corresponding reference y.   To be consistent with previous works, the score is calculated over 5 observations: atelectasis, cardiomegaly, consolidation, edema and pleural effusion.

RG rewards   We use the rewards explained in Section 2.2 as evaluation scores.

5.3 RadGraph vs fact$_{ENT}$ and fact$_{ENTNLI}$

Theoretically, RG$_E$ encapsulates both fact$_{ENT}$ and fact$_{ENTNLI}$ concepts.   Indeed, RG$_E$ focuses on having the right entities and also the right entity labels. Given that the label of an entity contains the notion of an anatomy or observation, the former being always present by definition and the latter having a degree of presence ("present", "absent" or "uncertain"), RG$_E$ can penalize a report presenting a contradiction with the reference, in the same fashion fact$_{ENTNLI}$ does.

Moreover, RG$_E$ presents two more advantages. First, it does not rely on an external model, such as the BERTScore, to map a hypothesis sentence with its counterpart in the reference report to run the NLI model. Secondly, the entities and entity labels evaluated by RG$_E$ are computed by a NER model specifically trained on chest X-rays, while fact$_{ENT}$ relies on a general-purpose biomedical NER system.

6 Results

In this section, we discuss the results displayed in Table 1. We divide the section into Quantitative analysis (Section 6.1), Qualitative analysis (Section 6.2) and Limitations (Section 9).

6.1 Quantitative analysis

Using NLL   The models reported in the NLL section are trained using only the negative log likelihood loss.   We can see that our simple approach, referred to as "ours (NLL)", performs similarly on the NLG metrics (better in BLEU but worse in ROUGEL compared to Chen et al. (2021)), but outperforms previous works on the factually-oriented metrics. On MIMIC-CXR, our baseline is up 2.3% on fact$_{ENT}$ and up 8.0% on fact$_{ENTNLI}$ compared to Miura et al. (2021).

Using RL   These results are the main contribution of our paper. We notice that optimizing any of the presented rewards in Section 2.2 improved the RadGraph scores, validating the design of our rewards.   Our best performing model is RG$_{ER}$ accross both datasets. On MIMIC-CXR, it shows an improvement of respectively 61.3%,

5

Model	BL4	ROUGEL	F_1cXb	fact$_{ENT}$	fact$_{ENTNLI}$	RG$_E$	RG$_{ER}$	RG$_{\overline{ER}}$
MIMIC-CXR								
Using NLL								
Liu et al. (2019)	7.6	—	29.2	—	—	—	—	—
Chen et al. (2020)	8.6	27.7	34.6	—	—	—	—	—
Chen et al. (2021)	10.6	**27.8**	40.5	—	—	—	—	—
Miura et al. (2021)	**11.5**	—	44.7	27.3	24.4	—	—	—
ours (NLL)	10.5	25.3	**44.8**	**28.0**	**26.8**	23.0	20.2	15.3
Using RL								
Miura et al. (fact$_{ENT}$)	11.1	—	56.7	39.5	34.8	—	—	—
Miura et al. (fact$_{ENTNLI}$)	11.4	—	56.7	38.5	37.9	—	—	—
ours (RG$_E$)	11.4	26.3	59.4	41.2	42.7	36.8	32.5	21.5
ours (RG$_{ER}$)	11.4	**26.5**	**62.2**	**42.5**	**43.3**	**37.1**	**34.7**	23.5
ours (RG$_{\overline{ER}}$)	**11.6**	25.9	51.4	41.8	40.9	35.4	31.6	**23.8**
Open-i								
Using NLL								
Chen et al. (2021)	12.0	**29.8**	—	—	—	—	—	—
Miura et al. (2021)	**12.1**	28.8	32.2	**40.6**	**42.9**	—	—	—
ours (NLL)	11.4	—	**33.1**	**40.6**	42.6	29.2	26.4	18.1
Using RL								
Miura et al. (fact$_{ENT}$)	12.0	—	48.3	44.4	46.8	—	—	—
Miura et al. (fact$_{ENTNLI}$)	13.1	—	47.8	43.6	47.1	—	—	—
ours (RG$_E$)	13.1	32.5	46.8	44.3	51.2	**44.1**	38.8	29.9
ours (RG$_{ER}$)	**13.9**	**32.7**	**49.1**	**46.0**	**58.9**	43.6	**41.2**	31.9
ours (RG$_{\overline{ER}}$)	12.1	30.6	45.3	43.3	53.2	41.9	39.1	**32.2**

Table 1: Comparison of our models against the state of the art (we took the best four models performing on MIMIC-CXR and the two best on Open-i). Results reported for Open-i are from models trained on MIMIC-CXR and tested on the entirety of the Open-i dataset. BL4 and F_1cXb refers to the BLEU4 and F_1CheXbert metrics. In **bold** are highlighted the best scores per category (i.e. NLL/RL and MIMIC-CXR/Open-i)

71.7% and 53.5% on the RG$_E$, RG$_{ER}$ and RG$_{\overline{ER}}$ metrics compared to our NLL baseline. This model also reports the best scores on the NLG metrics as well as improvements of +7.59% on fact$_{ENT}$ and +14.2% on fact$_{ENTNLI}$ over Miura et al. (2021). This means reports are generated with more factually-correct entities and less contradictions. On the F_1CheXbert score, our model also reports an improvement of +9.7% compared to the fact$_{ENTNLI}$ model. It is also worth noting that if we were to include all abnormalities in the computation of the F_1CheXbert, RG$_{ER}$ score would be **56.0%**, meaning that our model also performs well on less represented classes.

On the out-of-domain test-set of open-i, our model trained with RG$_{ER}$ reward reports a similar trend, with a noticeable improvement of 16.7% on the fact$_{ENTNLI}$ metric (55.0 vs 47.1).

RG$_{ER}$ vs RG$_{\overline{ER}}$ Surprisingly, RG$_{ER}$ outperforms RG$_{\overline{ER}}$ on both datasets and on every metrics. A hypothesis is that the RG$_{\overline{ER}}$ reward is too restrictive on the relations. Indeed, for an entity and its relation, a point of precision is only given if the label of the relation and the target entity of the relation are both found in the reference report. On MIMIC-CXR test-set, only 20.4% of the relations generated by our model are correct (5071 out of 24818 generated relations). It means that for 79.6% of the generated relations, our model received a negative signal even if some of these relations were partly correct. By contrast, when optimizing RG$_{ER}$ where a point of precision is given when the relation for an entity exists in the generated and reference report, regardless of the label
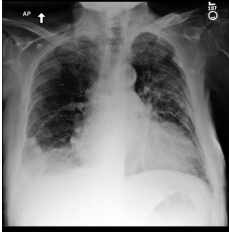
Table 2 with examples:

Image	Ours (NLL)	Ours (RG$_{ER}$)	Reference
	in comparison with the study of ___ , the monitoring OBS-DP and support OBS-DP devices OBS-DP remain in place OBS-DP place . continued enlargement OBS-DP of the cardiac ANAT-DP silhouette ANAT-DP with bilateral ANAT-DP pleural ANAT-DP effusions OBS-DP and compressive OBS-DP atelectasis OBS-DP at the bases ANAT-DP . in the appropriate clinical setting , supervening OBS-U pneumonia OBS-U would have to be considered .	there is no relevant OBS-DA change OBS-DA . there is no evidence of pneumothorax OBS-DA . the lung ANAT-DP volumes ANAT-DP remain low OBS-DP . the monitoring OBS-DP and support OBS-DP devices OBS-DP are in constant OBS-DP position OBS-DP .	as compared to the previous radiograph , there is no relevant change . the monitoring OBS-DP and support OBS-DP devices OBS-DP are constant OBS-DP . no evidence of pneumothorax OBS-DA . no other acute OBS-DA interval OBS-DA changes OBS-DA .
RG$_{ER}$ reward:	26.0%	52.6%	
ChexBert labels	Cardiomegaly, Pneumonia, Atelectasis Support Devices	Support Devices, No Finding	Support Devices, No Finding
	compared to the prior study there is no significant interval change OBS-DA .	as compared to the previous radiograph , the lung ANAT-DP volumes ANAT-DP are low OBS-DP . there is a small OBS-DP right ANAT-DP pleural ANAT-DP effusion OBS-DP with areas OBS-DP of atelectasis OBS-DP in the right ANAT-DP lung ANAT-DP bases ANAT-DP . there is no focal OBS-DA consolidation OBS-DA , pleural ANAT-DP effusion OBS-DA or pneumothorax OBS-DA . the heart ANAT-DP size ANAT-DP is mildly OBS-DP enlarged OBS-DP . the mediastinal ANAT-DP and hilar ANAT-DP contours ANAT-DP are unchanged OBS-DP .	as compared to ___ , interval worsening OBS-DP moderate OBS-DP pulmonary ANAT-DP edema OBS-DP . right ANAT-DP moderate OBS-DP pleural ANAT-DP effusion OBS-DP has also slightly OBS-DP increased OBS-DP . small OBS-DP left ANAT-DP effusion OBS-DP persists . left ANAT-DP lower ANAT-DP lobe ANAT-DP parenchymal ANAT-DP opacity OBS-DP in the superior ANAT-DP segment ANAT-DP is now obscured OBS-DP by increasing OBS-DP pulmonary ANAT-DP edema OBS-DP . moderate OBS-DP cardiomegaly OBS-DP . no pneumothorax OBS-DA .
RG$_{ER}$ reward:	0.0%	22.2 %	
ChexBert labels	No findings	Enlarged Cardiomediastinum, Cardiomegaly	Enlarged Cardiomediastinum, Cardiomegaly Lung Opacity, Pleural Effusion

Table 2: Cherry picked examples that compare two of our models' outputs: Ours (NLL) and Ours (RG$_{ER}$).

and the target entity, 38.7% of the relations are correct (9974 out of 25769 generated relations). This is 18.4% more than RG$_{\overline{ER}}$. We assume that this relaxed constraint encourages the model to generate relations and therefore more factually correct and complete reports.

Impression section We also evaluate our models on their potential to generate the *Impression* section of a report, based on the corresponding chest X-ray image, instead of the *Findings* section (see Appendix E). *Impression* highlights the key observations and conclusions of the radiology study. Automating this task is also critical because the *Impression* section is the most important part of a radiology report, and can be time-consuming and error-prone to produce. In addition, generating *Impression* is related to Radiology Report Summarization (Zhang et al., 2020) where a system has to summarize the *Findings* section of a report into *Impression*, making this choice even more relevant. MIMIC-CXR now consists of 185, 816 and 1, 521 samples for the training and validation sets. The MIMIC-CXR and Open-i test sets now have respectively 2, 224 and 3, 820 samples.

6.2 Qualitative analysis

First, we performed a human evaluation to further confirm whether the generated radiology reports are more factually complete and consistent. Two board-certified radiologists were asked to perform up to a hundred studies, where they had to choose between two findings given the chest X-ray. The two findings are from fact$_{ENTNLI}$ and our model RG$_{ER}$. On average, the radiologists favored our model. We give more details of the study in Appendix C.

The Figure 4 shows the number of entities and relations generated by our best model RG$_{ER}$, aggregated per label, on the MIMIC-CXR test-set. The takeaways are that 1) our model generates 20% more *Anatomy* entities compared to the ground-truth reports 2) for the 4 most frequent labels, namely *OBS-DP*, *modify*, *located_at*, *OBS-DA*, the model generates between -12% and +24% entities and relations 3) our model barely generated any occurrences of the two most under-represented labels: *OBS-U* and *suggestive_of*.

It is also interesting to note that the median word-length of the RG$_{ER}$ findings is 7% lower than the

ground-truth findings and 25% lower for the NLL findings (on MIMIC-CXR). We can see in Table 2 two cherry picked examples showing reports generated by our two models. We see in both instances that the length of the reports from our RG$_{ER}$ model is closer to the references lengths compared to the NLL model.

7 Related work

First, we describe the studies that presented architectural novelties. Usually, they focus on improving the widely used NLG metrics such as BLEU and ROUGE. We then proceed to go over the previous works that used Reinforcement Learning (RL) to optimize NLG or factually-oriented metrics. Finally, we quickly mention a few projects that do not fall into the first two categories.

Architectural novelties Chen et al. (2020) proposed to 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. In Chen et al. (2021), authors investigated cross-modal memory networks to enhance the encoder-decoder framework for radiology report generation, where a shared memory is designed to record the alignment between images and texts so as to facilitate the interaction and generation across modalities. Liu et al. (2021a) used Posterior-and-Prior knowledge to imitate the working patterns of radiologists, who first examine the abnormal regions and assign the disease topic tags to the abnormal regions, and then rely on the years of prior medical knowledge and prior working experience accumulations to write reports. More specifically, the prior knowledge consists of a medical knowledge graph built using unsupervised topic modeling. Topics are defined as nodes and are grouped by the organ or body part that they relate to. They connect their nodes with bidirectional edges, resulting in closely connected related topics. Another notable work used a Knowledge Graph Auto-Encoder (Liu et al., 2021b) taking as input a knowledge graph constructed in an unsupervised manner.

We differ from these two last works in two ways: 1) our semantic graph is generated by a model trained on high-quality human-annotated data and 2) these studies used the graph as input to their model while we use our graph as an evaluation metric that can be directly optimized using Reinforcement Learning.

Reinforcement Learning (RL) (Liu et al., 2019) improved Radiology Report Generation by optimizing the correctness of the output of CheXpert. The metric they optimized is equivalent to the CheXbert metric presented in our paper. (Miura et al., 2021) proposed new metrics to evaluate the factual correctness and consistency of the generated report. The metrics are included in our work and referred to as fact$_{ENT}$ and fact$_{ENTNLI}$ in Section 5.2.

Our paper is an original contribution to these two preceding works: we present a new evaluation metric that answers previous weaknesses. First, our graph-based annotations are generated by a model trained on a dataset with entities and relations labeled by radiologists. Second, our annotations are specific to the chest X-ray domain. Finally, our reward captures new semantic nuances such as new entity labels (anatomy, observation, absent, present) and new relationship levels between words.

Other works Two previous works generated radiology reports using retrieval methods. Endo et al. (2021) used a retrieval-based radiology report generation approach using a pre-trained contrastive language-image model. At test time, they retrieved the most likely report in the training dataset given the representation of the encoded X-ray. Li et al. (2018) employed a hierarchical decision-making procedure. For each sentence, a high-level retrieval policy module chooses to either retrieve a template sentence from an off-the-shelf template database, or invoke a low-level generation module to generate a new sentence. The decisions are updated via reinforcement learning, guided by sentence-level and word-level rewards. Finally, Yang et al. (2021) decided to input the RadGraph annotations in addition to the X-ray image and showed moderate to no improvement compared to previous works.

	MIMIC-CXR												
	ANAT-DP		OBS-DP		modify		located_at		OBS-DA		OBS-U		suggestive_of
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
ours (NLL)	29.7	23.6	26.6	15.7	13.6	9.2	14.2	11.1	29.8	48.4	13.5	3.0	9.8	2.1
ours (RG_{ER})	38.0	45.5	33.6	29.8	16.5	14.6	18.3	22.8	49.5	47.3	0.0	0.0	11.1	0.1

Table 3: Precision and recall per each entity and relation label on the MIMIC test-set, for the RG_{ER} model. For each label and study, we assess whether the true words from the ground-truth report correspond to the generated words from the generated report - or pairs of source and target words in the case of relations.

8 Conclusion

In this paper, we leveraged the RadGraph dataset containing annotated chest X-ray reports with entities and relations between entities to design a new reward that qualitatively evaluate the factual correctness and completeness of the generated reports. We showed on two datasets that directly optimizing these rewards outperforms previous approaches that prioritize traditional NLG metrics or leverage unsupervised out-of-domain systems as factual-oriented metrics. Our best model reports up to +14.2% improvements on these factual metrics on the MIMIC-CXR and +25.3% on Open-i.

9 Limitations

In this section, we highlight four limitations of our work.

First, rewards based solely on entities (Miura et al., 2021) or entities and relations cannot be optimized without counter-effects happening. Indeed, optimizing $fact_{ENT}$ or RG_{ER} will encourage the model to discard the grammar and generate reports such as "left lower right base uper opacities pleural cardiopulmonary cardiopulmonary atelectasis" to maximize the precision of entities generated. For this reason, we follow the settings of previous work and optimize one of our RG metrics alongside the BERTScore (Zhang et al., 2019) and the NLL loss (with weights 0.495, 0.495 and 0.01 respectively)

Secondly, our model capability at correctly connecting observations and corresponding anatomies remains limited. Table 3 depicts the precision and recall per entity and relation labels: in general, we observe that precision and recall have similar values among each label, but vary significantly from one label to the other. Excluding the two under-represented labels, *OBS-U* and *suggestive_of*, entities have a macro-averaged re-
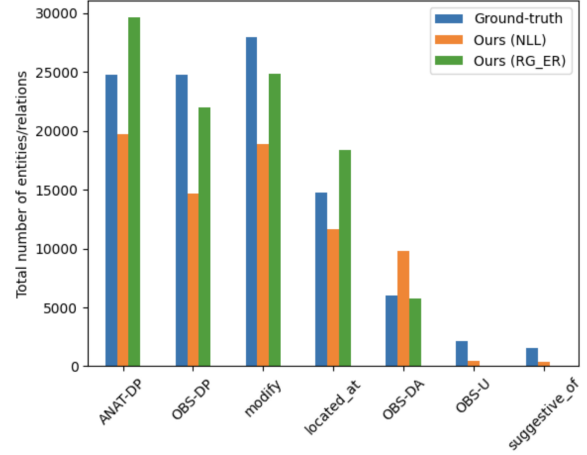

Figure 4: Number of entities and relations generated by our best model RG_{ER}, aggregated per label, on the MIMIC-CXR test-set. We also show the ground-truth distribution by running the RadGraph model on the reference test-set.

call of 40.9% compared to only 18.7% for relations.

More critically, the *OBS-U* entities are not correctly learned by our RG_{ER} model, as underlined in Table 3, the recall for this label being 0. We measured that 30% of the words labeled as *OBS-U* in the reference are incorrectly generated as *OBS-DP* or *OBS-DA* by our mode. The rest of the "missed" *OBS* and *ANAT* entities are due to the entity being not present in the generated report. Concerning the errors on the relation labels, we noticed that 15% of the relations have the incorrect relation label, while the rest of the errors are due to the relation being just absent.

Finally, we note that even though our model fits on a single GPU of 12 GB, training the model using RL is computationally expensive. A RL epoch is between 7 to 10 hours on MIMIC-CXR (depending on the randomness of the sampling) against 50 minutes for NLL training.

References

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

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.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv e-prints*, page arXiv:1901.07042.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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*, volume 31. Curran Associates, Inc.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.

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

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. PMLR.

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

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.

10

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

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.

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

Richard S Sutton, Andrew G Barto, et al. 1998. Introduction to reinforcement learning.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shuxin Yang, Xian Wu, Shen Ge, Shaohua Kevin Zhou, and Li Xiao. 2021. Knowledge matters: Radiology report generation with general and specific knowledge. *arXiv preprint arXiv:2112.15009*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

A Code release

To help with further research, we make our code publicly available. More specifically, we release the code of all the factually-oriented metrics presented in Section 5.2 in one package. Our code also includes SCST (Rennie et al., 2017) on top of the widely-used NLP library HuggingFace (Wolf et al., 2020). We hope this effort will improve reproducibility of the factually-oriented metrics and allow for a fairer comparison of the performance of future radiology report generation systems. Our code is available at https://anonymous.4open.science/r/rrg_emnlp-D382/. This repository is anonymized and respect the double-blind process.

B Set $\bar{V}$ of Figure 3

$V = \{$lower, infection, right, lobe, opacity, pneumothorax, increased$\}$
$E = \{$(right, lobe), (lower, lobe), (opacity, infection), (opacity, lobe), (increased, opacity)$\}$

$\bar{V}$ of RG$_E$ = {(lower, anat), (infection, obs-dp), (right, anat), (lobe, anat), (opacity, obs-dp), (increased, obs-dp), (pneumothorax, obs-da)}

$\bar{V}$ of RG$_{ER}$ = {(lower, anat, 1), (infection, obs-dp, 0), (right, anat, 1), (lobe, anat, 0), (opacity, obs-dp, 1), (increased, obs-dp, 1), (pneumothorax, obs-da, 0)}

$\bar{V}$ of RG$_{\bar{ER}}$ = {(lower, anat, lobe, modify), (infection, obs-dp), (right, anat, lobe, modify), (lobe, anat), (opacity, obs-dp, infection, suggestive of), (opacity, obs-dp, lobe, located_at), (increased, obs-dp, opacity, modify), (pneumothorax, obs-da)}

C Qualitative Study

To evaluate qualitatively how our RG$_{ER}$ model compares to previous models such as fact$_{ENTNLI}$ model, we built a study that asked radiologists to assess for each test image, based on their experience and the clinical expectations, which corresponding generated report is the best.

The clinical studies chosen for this experience are from the MIMIC-CXR test set and contain the following labels:

```
{
    "Lung Opacity": 39,
    "Pleural Effusion": 36,
    "Support Devices": 36,
    "Atelectasis": 26,
    "Cardiomegaly": 19,
    "Lung Lesion": 19,
    "Pleural Other": 16,
    "Enlarged Cardiomediastinum": 15,
    "Pneumonia": 15,
    "Fracture": 14,
    "Edema": 13,
    "Consolidation": 13,
    "Pneumothorax": 10,
    "No Finding": 5
}
```
Listing 1: Labels of the 100 clinical studies selected for the qualitative comparison of RRG models. There is more than 100 labels since one report can contain multiple labels. We made sure to have at least 10 instances for each abnormality.

The radiologists received the following instructions: "Please evaluate the radiology reports given the chest x-ray using the following criteria: 1. factual correctness (is it correct ?); 2. factual completeness (how complete is the report ?); 3. factual consistency (is there any contradiction within the report ?)".

Radiologists were asked to choose, for each pair of generated reports, one being from fact$_{ENTNLI}$ model and the other from RG$_{ER}$ model, a score on a scale ranging from -2 (report 1 is much better) to 2 (report 2 is much better). Then, we aggregated the scores for each pair of reports and for each labeler, and computed the average score of reports coming from RG$_{ER}$ model compared to reports coming from fact$_{ENTNLI}$ model.

Compared to fact$_{ENTNLI}$, our RG$_{ER}$ model is:

Reader 1: 0.485 ± 0.662 Reader 2: 0.891 ± 0.461
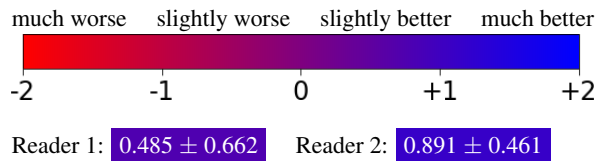
According to both radiologists, on average reports from our RG$_{ER}$ model are prefered to fact$_{ENTNLI}$. We notice that scores can be improved and the reports of our model are not systematically better. Following our ideas in Section 9, we would like to improve upon our current RRG model in the future.

D Model

Encoder:

932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981

```
{
    encoder: CNN
    backbone: densenet121
    output_layer: features (of size 49x1024)
    dropout_out: No dropout
    output_size: 1024
}
```

The decoder is based on HuggingFace (Wolf et al., 2020):

```
{
    decoder: BertGenerationDecoder
    add_cross_attention: true
    attention_probs_dropout_prob: 0.1
    hidden_act: gelu
    hidden_dropout_prob: 0.1
    hidden_size: 768
    initializer_range: 0.02
    intermediate_size: 3072
    is_decoder: true
    layer_norm_eps: 1e-05
    max_position_embeddings: 514
    num_attention_heads: 12
    num_hidden_layers: 1
    position_embedding_type: absolute
    type_vocab_size: 1
    vocab_size: 9877 (for mimic-cxr)

}
```

The model has 25.8M learnable parameters and fits on a single GPU of 12GB.

The training hyper parameters are as such:

```
{

    batch_size: 128
    optimizer: RAdam
    optim_params:
        lr: 0.0003
        weight_decay: 0.
    lr_decay: ReduceLROnPlateau
    lr_decay_params:
        factor: 0.8
        patience: 1
        min_lr: 0.000001
        threshold_mode: abs
    early_stop: 10
    early_stop_metric: ROUGEL
}
```

The plateau is monitored on ROUGEL metric during eval.

E Results on the impression section

	MIMIC-CXR		
Model	F_1cXb	fact$_{ENT}$	fact$_{ENTNLI}$
ours (RG$_{ER}$)	52.1	42.6	40.5
	RG$_E$	RG$_{ER}$	RG$_{\overline{ER}}$
ours (RG$_{ER}$)	39.6	33.2	23.1

Table 4: Results of our RG$_{ER}$ model trained on generating the *Impression* section of MIMIC-CXR instead of *Findings*.

13