Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, Chun-Nan Hsu
University of California, San Diego
{ayan,zehe,xil135,jiangdu,e8chang}@ucsd.edu
{agentili,jmcauley,chunnan}@ucsd.edu

Abstract

Radiology report generation aims at generating descriptive text from radiology images automatically, which may present an opportunity to improve radiology reporting and interpretation. A typical setting consists of training encoder-decoder models on image-report pairs with a cross entropy loss, which struggles to generate informative sentences for clinical diagnoses since normal findings dominate the datasets. To tackle this challenge and encourage more clinically-accurate text outputs, we propose a novel weakly supervised contrastive loss for medical report generation. Experimental results demonstrate that our method benefits from contrasting target reports with incorrect but semantically-close ones. It outperforms previous work on both clinical correctness and text generation metrics for two public benchmarks.

1 Introduction

Automated radiology report generation aims at generating informative text from radiologic image studies. It could potentially improve radiology reporting and alleviate the workload of radiologists. Recently, following the success of deep learning in conditional text generation tasks such as image captioning (Vinyals et al., 2015; Xu et al., 2015), many methods have been proposed for this task (Jing et al., 2017; Li et al., 2018; Liu et al., 2019; Jing et al., 2020; Ni et al., 2020; Chen et al., 2020b).

Unlike conventional image captioning benchmarks (e.g. MS-COCO (Lin et al., 2014)) where referenced captions are usually short, radiology reports are much longer with multiple sentences, which pose higher requirements for information selection, relation extraction, and content ordering. To generate informative text from a radiology image study, a caption model is required to understand the content, identify abnormal positions in an image and organize the wording to describe findings in images. However, the standard approach of training an encoder-decoder model with teacher forcing and cross-entropy loss often leads to text generation outputs with high frequency tokens or sentences appearing too often (Ranzato et al., 2015; Holtzman et al., 2019). This problem could be worse for chest X-ray report generation, since the task has a relatively narrow text distribution with domain-specific terminology and descriptions for medical images, and models often struggle to generate long and diverse reports (Harzig et al., 2019; Boag et al., 2020).

To tackle these challenging issues, we propose to introduce contrastive learning into chest X-ray report generation. However, simply using random non-target sequences as negative examples in a contrastive framework is suboptimal (Lee et al., 2020), as random samples are usually easy to distinguish from the correct ones. Hence, we further introduce a weakly supervised contrastive loss that assigns more weights to reports that are semantically close to the target. By exposing the model to these “hard” negative examples during training, it could learn robust representations which capture the essence of a medical image and generate high-quality reports with improved performance on clinical correctness for unseen images.

Overall, our contributions are three-fold:

• We propose a novel objective for training a chest X-ray report generation model with a contrastive term, which contrasts target reports with incorrect ones during training.

• We develop a weakly supervised method to identify “hard” negative samples and assign them with higher weights in our contrastive loss to further encourage diversity.

• We conduct comprehensive experiments to show the effectiveness of our method, which outperforms existing methods on both clinical correctness and text generation metrics.
2 Related Work

Medical Report Generation Medical report generation, which aims to automatically generate descriptions for clinical radiographs (e.g. chest X-rays), has drawn attention in both the machine learning and medical communities. Many methods have been proposed to solve this task. For example, Jing et al. (2017) proposed a co-attention hierarchical CNN-RNN model. Li et al. (2018, 2019) used hybrid retrieval-generation models to complement generation. Most recently, Ni et al. (2020) proposed to learn visual semantic embeddings for cross-modal retrieval in a contrastive setting, and showed improvements in identifying abnormal findings. However, this is hard to scale or generalize since we need to build template abnormal sentences for new datasets. Chen et al. (2020b) leveraged a memory-augmented transformer model to improve the ability to generate long and coherent text, but didn’t address the issue of generating dominant normal findings specifically. Our work proposes to incorporate contrastive learning into training a generation-based model, which benefits from the contrastive loss to encourage diversity and is also easy to scale compared to retrieval-based methods.

Contrastive Learning Contrastive learning (Gutmann and Hyvärinen, 2010; Oord et al., 2018) has been widely used in many fields of machine learning. The goal is to learn a representation by contrasting positive and negative pairs. Recent work showed that contrastive learning can boost the performance of self-supervised and semi-supervised learning in computer vision tasks (He et al., 2020; Chen et al., 2020a; Khosla et al., 2020). In natural language processing, contrastive learning has been investigated for several tasks, including language modeling (Huang et al., 2018), unsupervised word alignment (Liu and Sun, 2015) and machine translation (Yang et al., 2019; Lee et al., 2020). In this work, we are interested in applying contrastive learning to chest X-ray report generation in a multi-modality setting. Different from previous work in applying contrastive learning to image captioning (Dai and Lin, 2017), we leverage a recent contrastive formulation inspired by (Chen et al., 2020a) which transforms representations into latent spaces, and propose a novel learning objective for the medical report generation task specifically.

3 Method

Generating Reports with Transformer We leverage a memory-driven transformer proposed in (Chen et al., 2020b) as our backbone model, which uses a memory module to record key information when generating long texts.

Given a chest X-ray image \( I \), its visual features \( X \) are extracted by pre-trained convolutional neural networks (e.g. ResNet (He et al., 2016)). Then we use the standard encoder in transformer to obtain hidden visual features \( H_X \). The decoding process at each time step \( t \) can be formalized as

\[
\hat{y}_t = \text{Decoder}(H_X, y_1, \ldots, y_{t-1}).
\]
We use a cross-entropy (CE) loss to maximize the conditional log likelihood \( \log p_\theta (Y | X) \) for a given \( N \) observations \((X^{(i)}, Y^{(i)})_{i=1}^N\) as follows:

\[
L_{CE} = \sum_{i=1}^{N} \log p_\theta (Y^{(i)} | X^{(i)}).
\]  

(2)

**Labeling Reports with Finetuned BERT**  As shown in Figure 1, we first extract the embeddings of each report from CheXBERT (Smit et al., 2020), a BERT model pretrained with biomedical text and finetuned for chest X-ray report labeling. We use the [CLS] embedding of BERT to represent report-level features. We then apply K-Means to cluster the reports into \( K \) groups. After clustering, each report \( Y \) is assigned with a corresponding cluster label \( l \), where reports in the same cluster are considered to be semantically close to each other.

**Weakly supervised Contrastive Learning**  To regularize the training process, we propose a weakly supervised contrastive loss (WCL). We first project the hidden representations of the image and the target sequence into a latent space:

\[
z_x = \phi_x (\tilde{H}_X), \quad z_y = \phi_y (\tilde{H}_Y),
\]  

(3)

where \( \tilde{H}_X \) and \( \tilde{H}_Y \) are the average pooling of the hidden states \( H_X \) and \( H_Y \) from the transformer, \( \phi_x \) and \( \phi_y \) are two fully connected layers with ReLU activation (Nair and Hinton, 2010). We then maximize the similarity between the pair of source image and target sequence, while minimizing the similarity between the negative pairs as follows:

\[
L_{WCL} = \sum_{i=1}^{N} \log \frac{\exp(s_{i,i})}{\sum_{l \neq i} \exp(s_{i,l}) + \alpha \sum_{l = l} \exp(s_{i,l})},
\]  

(4)

where \( s_{i,j} = \text{sim}(z_x^{(i)}, z_y^{(j)})/\tau, \text{sim} \) is the cosine similarity between two vectors, \( \tau \) is the temperature parameter, \( \alpha \) is a hyperparameter that weighs the importance of negative samples that are semantically close to the target sequence, i.e., with the same cluster label \( l_i = l_j \) in Eq. (4). Empirically, we find that these samples are “hard” negative samples and the model would perform better by assigning more weights to distinguish these samples.

Overall, the model is optimized with a mixture of cross-entropy loss and weakly supervised contrastive loss:

\[
L_{\text{loss}} = (1 - \lambda) L_{CE} + \lambda L_{WCL}.
\]  

(5)

where \( \lambda \) is a hyperparameter that weighs the two losses.

### 4 Experiments

#### 4.1 Experimental Setup

**Datasets**  We conduct experiments on two datasets: (1) MIMIC-ABN, which was proposed in (Ni et al., 2020) and contains a subset of images of MIMIC-CXR with abnormal sentences only, with 26,946/3,801/7,804 reports for train/val/test sets. (2) MIMIC-CXR (Johnson et al., 2019), the largest radiology dataset to date that consists of 222,758/1,808/3,269 reports for train/val/test sets.

**Evaluation Metrics**  Performance is first evaluated on three automatic metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Denkowski and Lavie, 2011).

We then use the CheXpert labeler to evaluate the clinical accuracy of the abnormal findings reported by each model, which is a state-of-the-art rule-based chest X-ray report labeling system (Irvin et al., 2019). Given sentences of abnormal findings, CheXpert will give a positive and negative label for 14 diseases. We then calculate the Precision, Recall and Accuracy for each disease based on the labels obtained from each model’s output and from the ground-truth reports.

**Implementation Details**  We use Adam as the optimizer with initial learning rates of 5e-5 for the visual extractor and 1e-4 for other parameters. The learning rate is decayed by a factor of 0.8 per epoch. We set the number of labels \( K \) to 13 and 14 for MIMIC-ABN and MIMIC-CXR, the weighting parameters \( \lambda \) and \( \alpha \) to 0.2 and 2 respectively. We conduct a grid-based hyperparameter search for weighting factor \( \lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\} \) and temperature \( \tau \in \{0.1, 1, 10\} \) by evaluating the models on the validation sets of the two datasets.

Words with frequency less than 3 and 10 are discarded for MIMIC-ABN and MIMIC-CXR respectively. The maximum sequence lengths are set to 64 and 100 for MIMIC-ABN and MIMIC-CXR. The projection heads consists of two convolutional layers with ReLU activation, and the latent dimensions are set to 256 for both visual and text projection layers. The hidden sizes of both the encoder and decoder are 512 with 8 heads and 3 layers. The batch size for training the two datasets is 128. We report the results using models with the highest BLEU-4 on the validation set. We use a beam size of 3 for generation to balance the generation effectiveness and efficiency.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>NLG metrics</th>
<th>CE metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU-1</td>
<td>BLEU-4</td>
</tr>
<tr>
<td>MIMIC-ABN</td>
<td>ST</td>
<td>14.9</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>HCR</td>
<td>8.4</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>CVSE</td>
<td>19.2</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>MDT</td>
<td>24.6</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>MDT+WCL</td>
<td>25.6</td>
<td>6.7</td>
</tr>
<tr>
<td>MIMIC-CXR</td>
<td>ST</td>
<td>29.9</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>TopDown</td>
<td>31.7</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>MDT</td>
<td>35.3</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>MDT+WCL</td>
<td>37.3</td>
<td>10.7</td>
</tr>
</tbody>
</table>

Table 1: The performance of all baselines and our full model on the test sets of MIMIC-ABN and MIMIC-CXR datasets with respect to natural language generation (NLG) and clinical efficacy (CE) metrics. Results are reported in percentage (%). ST is CNN+LSTM with attention (Xu et al., 2015). HCR (Jing et al., 2017) is a hierarchical CNN-RNN model. CVSE (Ni et al., 2020) is a cross-modal retrieval model. TopDown (Anderson et al., 2018) is a widely-used image captioning model. MDT is a memory-driven transformer proposed in (Chen et al., 2020b).

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation</th>
<th>NLG metrics</th>
<th>CE metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU-1</td>
<td>BLEU-4</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td>25.0</td>
<td>6.6</td>
</tr>
<tr>
<td>adversarial</td>
<td></td>
<td>25.1</td>
<td>6.6</td>
</tr>
<tr>
<td>excluding</td>
<td></td>
<td>24.8</td>
<td>6.6</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>25.4</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Table 2: Ablation study on the MIMIC-ABN dataset. baseline is the vanilla contrastive loss ($\alpha = 1$) introduced in (Chen et al., 2020a); excluding excludes negative samples in the same cluster ($\alpha = 0$); adversarial is proposed in (Lee et al., 2020) where they construct extra negative samples via adversarial attack. Results are averaged over three runs.

4.2 Performance comparison

We compare our approach to other methods on two datasets. As shown in Table 1, first, our method (MDT+WCL) outperforms previous retrieval (CVSE) and generation based models (MDT) on most text generation metrics. Second, our contrastive loss significantly improves clinical efficacy metrics, demonstrating its capability to accurately report abnormal findings. Finally, the relative difference between MDT and MDT+WCL is higher on MIMIC-CXR, which contains a larger training set for learning robust representations.

4.3 Analysis

Different Contrastive Losses We compare our contrastive formulation with other variants in Table 2. Our approach achieves the best performance over all baselines on different metrics. Both adversarial and excluding have similar performance compared to the vanilla contrastive framework (Chen et al., 2020a), not showing improvements for medical report generation. On the other hand, we identify the reports in the same cluster as “hard” negative samples and assign more weights on them during training, guiding the model to better distinguish reference reports from inaccurate ones.

Length Distributions To further evaluate the generation quality of our method in addition to NLG and CE metrics, we compare the length of generated reports to the ground truth. To do this, we categorize all reports generated on the MIMIC-CXR test set into 20 groups (within the range of [0, 100] with an interval of 5). As shown in Figure 2, the baseline model MDT has a sharper distribution, while adding our weakly supervised contrastive loss leads to a length distribution which is smoother and closer to the ground truth, demonstrating its effectiveness to generate more diverse and accurate reports, and generalize on unseen images.
WCL: linear left basilar opacity is likely atelectasis versus scarring. M with chest pain.

GT: the lungs are hyperinflated but clear of consolidation or effusion. Hyperinflation without acute cardiopulmonary process.
WCL: the lungs are hyperinflated but clear of consolidation or effusion. Hyperinflation without acute cardiopulmonary process.

GT: there is bibasilar atelectasis. A lineq cardiac monitoring device projects over the subcutaneous tissue of the left lower chest. I with shortness of breath. Evaluate for pneumonia.
WCL: the lungs are hyperinflated with flattening of the diaphragms suggesting chronic obstructive pulmonary disease. There is mild bibasilar atelectasis.

Table 3: Generated samples from MIMIC-ABN dataset.

GT: there is moderate amount of right-sided subcutaneous emphysema which is similar in appearance compared to prior. Right-sided chest tube is again visualized. There is no increase in the pneumothorax. Bilateral parenchymal opacities are again visualized and not significantly changed. The tracheostomy tube is in standard location. Right subclavian line tip is in the mid svc.
WCL: tracheostomy tube tip is in unchanged position. Right-sided port-a-cath tip terminates in the low svc. Left-sided port-a-cath tip terminates in the proximal right atrium unchanged. Heart size is normal. Mediastinal and hilar contours are similar. Innumerable bilateral pulmonary nodules are re-demonstrated better assessed on the previous ct. Small right pleural effusion appears slightly increased compared to the prior exam. Small left pleural effusion is similar. No new focal consolidation or pneumothorax is present. There are no acute osseous abnormalities.

GT: the lungs are mildly hyperinflated as evidenced by flattening of the diaphragms on the lateral view. Diffuse interstitial markings compatible with known chronic interstitial lung disease are unchanged. There is no pleural effusion or evidence of pulmonary edema. There is no focal airspace consolidation or pneumothorax is present. There are no acute osseous abnormalities.
WCL: lung volumes are low. Heart size is mildly enlarged. The aorta is tortuous and diffusely calcified. Crowding of bronchovascular structures is present without overt pulmonary edema. Patchy opacities in the lung bases likely reflect areas of atelectasis. No focal consolidation pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

Table 4: Generated samples from MIMIC-CXR dataset.

Case Study We present examples of generated reports and their corresponding ground truth from MIMIC-ABN and MIMIC-CXR. As shown in Table 3 and Table 4, our method (WCL) is able to generate similar contents which are aligned with the ground truth (GT) written by radiologists. For example, abnormal findings in specific positions (e.g., “low lung volumes” and “enlarged heart size”) are reported, and potential diseases are also noted (e.g., “atelectasis”).

5 Conclusion

In this paper, we present a weakly supervised contrastive learning framework for generating chest X-ray reports. Our contrastive loss could lead to better results on both clinical correctness and text generation metrics than previous methods. We also show that exposing the model to semantically-close negative samples improves generation performance. In the future, we will extend our method to other medical image datasets other than chest X-rays.

Acknowledgments

This work was partially supported by the Office of the Assistant Secretary of Defense for Health Affairs through the AIMM Research Program endorsed by the Department of Defense under Award No. W81XWH-20-1-0693 and NSF Award #1750063. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense and the National Science Foundation.

References


William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter


