JPG - Jointly Learn to Align:
Automated Disease Prediction and Radiology Report Generation

Jingyi You♡, Dongyuan Li♡, Manabu Okumura♡, Kenji Suzuki♠
Tokyo Institute of Technology
♡{youjy,lidy,oku}@lr.pi.titech.ac.jp
♠suzuki.k.di@m.titech.ac.jp

Abstract

Automated radiology report generation aims to generate paragraphs that describe fine-grained visual differences among cases, especially those between the normal and the diseased. Existing methods seldom consider the cross-modal alignment between textual and visual features and tend to ignore disease tags as an auxiliary for report generation. To bridge the gap between textual and visual information, in this study, we propose a “Jointly learning framework for automated disease Prediction and radiology report Generation (JPG)” to improve the quality of reports through the interaction between the main task (report generation) and two auxiliary tasks (feature alignment and disease prediction). The feature alignment and disease prediction help the model learn text-correlated visual features and record diseases as keywords so that it can output high-quality reports. Besides, the improved reports in turn provide additional harder samples for feature alignment and disease prediction to learn more precise visual and textual representations and improve prediction accuracy. All components are jointly trained in a manner that helps improve them iteratively and progressively. Experimental results demonstrate the effectiveness of JPG on the most commonly used IU X-RAY dataset, showing its superior performance over multiple state-of-the-art image captioning and medical report generation methods with regard to BLEU, METEOR, and ROUGE metrics.

1 Introduction

Writing radiology reports and predicting disease labels are two essential procedures in clinical practice. However, manually creating them by radiologists is laborious and time-consuming (Jing et al., 2018; Chen et al., 2021b). Therefore, automated radiology report generation and disease prediction, which aim to generate formal-format descriptive texts (Fig. 1 Findings) and clinical conclusive terminologies (Fig. 1 MeSH), have received increasing attention recently (Chen et al., 2020; Miura et al., 2021; Liu et al., 2021b; Nguyen et al., 2021; Liu et al., 2021c; You et al., 2021a). In particular, they not only improve the efficiency of the entire procedure and liberate people from burdensome workloads, but also maintain the high quality of healthcare.

In spite of substantial improvements (Zhang et al., 2020; Wang et al., 2022; Liu et al., 2021a; Shao et al., 2021) have been achieved in the automatic radiology report generation and disease prediction, several challenges remain unsolved. Firstly, following traditional image captioning paradigms (Bhattacharya et al., 2022), current methods mainly adopt a standard encoder-decoder framework with convolutional neural networks (CNNs) encoding radiographs and recurrent neural networks (e.g., LSTM/GRU) or non-recurrent neural networks (e.g., Transformer) decoding reports. As a result, visual and textual information are represented by different encoding methods in their own specific embedding spaces, so that the features are misaligned (e.g., the visual represen-
tation of the regions circled in yellow in Fig. 1 is significantly different from the textual representation of “right upper lobe” in Findings). Therefore, directly applying these visual features to the downstream task will lead to low-quality reports (Chen et al., 2021a,b; Lu et al., 2017).

Furthermore, most existing disease prediction models (Bhattacharya et al., 2022; Sun et al., 2021; Gheflati and Rivaz, 2021; Park et al., 2022) attach a single disease label to each image, where its context (e.g., location, severity, and affected organs) is seldom considered. Automatically mining context-aware disease labels can thus make it easier to understand the disease. Finally, approaches take only visual information as the input of the downstream report generation, which ignores context-aware disease tags as auxiliary textual information. Intuitively, as high-level conclusive features, disease tags can more effectively guide the text generation and alleviate missing keywords.

To overcome the aforementioned problems, we propose to integrate radiology report generation and context-aware disease prediction into an overall framework (JPG), where context-aware disease labels serve as high-level auxiliary information for facilitating the report with the lesion location. Specifically, both visual and textual features are first projected into a shared subspace via a shared base matrix to learn new visual and textual representations. The shared base matrix acts as an intermediate medium, which enables visual and textual information to sufficiently interact and fuse in a manner that relieves misalignment between the features. As for the second issue, we train a CNN-RNN architecture to automatically search for context-aware disease labels. Instead of directly using the output of the CNN, the aligned visual features are applied to initialize the RNN hidden state for context-aware disease label prediction. Consequently, the model can improve the classification accuracy and disease label quality. Finally, we incorporate context-aware disease labels as high-level auxiliary features together with aligned visual features into the decoder, so that the comprehensive disease tags can better guide the report generation.

We highlight the contributions as follows:

- We propose to learn visual and textual representations through a shared subspace to relieve the misalignment across modalities, which can also be easily transplanted to other multi-modal tasks.
- Instead of directly using single labels in the disease prediction task, we propose a strategy to mine context-aware labels to provide a more detailed textual conclusion for lesions in radiographs.

As far as we know, we are the first to use predicted disease contextual labels as high-level auxiliary information for facilitating and guiding the report generation process. Empirical results demonstrate that this scheme proposal outperforms state-of-the-art competitors in terms of the automated radiology report generation.

2 Related Work

2.1 Image Captioning

Image captioning aims to generate sentences that describe images, and it has achieved great success in the cross-modal area (Cornia et al., 2020; Zhou et al., 2020; Shi et al., 2021). Inspired by encoder-decoder architectures used in machine translation, most existing image captioning approaches typically adopt the CNN-RNN framework (Huang et al., 2019; Yan et al., 2021; You et al., 2021b), where a CNN is used to extract visual features from a given image, and a recurrent or non-recurrent network is used to generate the caption. To align visual features with textual features, existing methods adopt a memory network (Chen et al., 2020, 2021b), a relation/consensus graph (Wang et al., 2021a; Bhattacharya et al., 2022), a Transformer network (Ji et al., 2021) or a language model (Sariyildiz et al., 2020; Gupta et al., 2020) to help visual features learn new semantic representations. Among those studies, the most related ones (You et al., 2018; Akbari et al., 2019) directly project visual features to a textual space and consider textual features as basis vectors to learn new representations for visual features. In contrast, in the present study, we design a shared subspace and a base matrix as an intermediate medium to learn new representations for both visual and textual features, which can thereby be better aligned.

2.2 Radiology Report Generation

As one of the applications and extensions of image captioning (Cornia et al., 2020; Zhou et al., 2020; Shi et al., 2021; Huang et al., 2019; Yan et al., 2021) (Appendix 2.1) to the medical domain, radiology report generation aims to annotate radiographs with much more detailed professional reports. According to the strategies for aligning radiological visual and textual features, current methods can be generally classified into three categories: 1) var
ant attention mechanism-based methods seek to integrate and fuse visual and textual features via advanced attention (Jing et al., 2019; Wang et al., 2018; Liu et al., 2019), among which Jing et al. (2018) propose a multi-task hierarchical model with a co-attention mechanism to combine visual and textual features to generate reports. 2) cross-modal memory network-based approaches record the alignment between images and texts through a shared matrix to facilitate the information interaction across modalities (Yin et al., 2019; Chen et al., 2020, 2021b; Wang et al., 2021b). 3) graph convolution network-based models aggregate visual and textual features on pre-trained knowledge graphs or newly constructed multi-modal networks (Zhang et al., 2020; Hu et al., 2019). Jpg offers a new way beyond the above studies to generate radiology reports, since a shared subspace is provided to learn new representations for both visual and textual features in a manner that produces more accurate descriptions for report generation.

2.3 Medical Image Classification

Existing methods have achieved remarkable success at predicting single disease labels for medical images (Bhattacharya et al., 2022; Sun et al., 2021; Gheflati and Rivaz, 2021; You et al., 2022). In particular, informative disease labels have been mined with context information. For example, Shin et al. (2016) predicts disease labels by leveraging a variant of the CNN-RNN framework. Moreover, PP-KED (Liu et al., 2021b) examines abnormal regions and assigns disease topic tags to the abnormalities. Differing from the above-mentioned methods, our Jpg adopts a shared base metric for learning new visual representations and takes it as input for context-aware disease prediction to improve the fluency of disease labels and classification accuracy.

3 Methodology

Figure 2 exhibits an overview of Jpg, which consists of three chief components: (A) shared subspace representation learning, (B) context-aware disease prediction, and (C) radiology report generation. Hereafter, we will give formal notations of variables and task definitions concerning Jpg, and introduce each component subsequently in detail.

3.1 Notations and Task Definition

Given an X-ray image I as input, Jpg is designed to automatically generate a sequence of context-aware disease labels c and a radiology report Y. Specifically, we divide I into p patches, and apply pre-trained CNN-based ResNet (He et al., 2016) as the visual extractor to learn its patch features as X = {x1, x2, . . . , xp}, where xp ∈ Rp with dp representing the dimensionality of patch features. The target output is the corresponding radiology report Y = {y1, y2, . . . , yn}, where yn ∈ Rdy is the word embedding of the n-th generated token, and n denotes the length of the report. Formally, the entire task can be defined as two parts according to Bayes’ theorem as follows:

\[ p(Y, c|X) \propto p(Y|c, X) \cdot p(c|X), \]  (1)

where the radiology report generation process p(Y|c, X) can be formalized as a recursive application of the chain rule as

\[ p(Y|c, X) = \prod_{i=1}^{n} p(y_i|y_{<i}, c, X), \]  (2)

where y_{<i} = {y_1, . . . , y_{i-1}} represents the previously generated tokens so far, and n is the total amount of tokens in target sequence Y.

As described in Eq. 1, jointly learning to align diagnostic disease prediction and radiological report generation can be classified as two subtasks in order. In detail, we first train the model to maximize the probability of producing context-aware disease labels for an X-ray image p(c|X), then maximize the probability of generating a corresponding radiology report p(Y|c, X) conditioned on context-aware disease labels c and visual features X.

3.2 Visual Extractor

As shown in Fig. 2, given a radiology image I organized in 2-dimension format as input, we employ ResNet (He et al., 2016) as a pre-trained visual extractor. Normally, it first decomposes the image into regions of equal size, i.e., patches, and then extracts visual features of each patch from the output of its last convolutional layer. Afterwards, the extracted patch representations x1, x2, . . . , xp are concatenated to constitute the source input for all subsequent modules with the form of visual feature sequence X ∈ Rp×dp as

\[ \{x_1, x_2, . . . , x_p\} = f_v(I). \]  (3)

Note that any type of pre-trained CNNs, e.g., VGG (Simonyan and Zisserman, 2015) or DenseNet (Huang et al., 2017), can be used for the purpose.
3.3 Shared Subspace Representation

Considering that visual and textual features are extracted by different encoding methods (Kim et al., 2020; Huang et al., 2020), directly applying patch features generated by the visual extractor as the input for the downstream text generation task will lead to non-fluent, low-quality reports with missing keywords. To solve this problem, as shown in Fig. 2 (A), both visual and textual features are projected into a shared subspace, and a trainable shared base matrix is designed to learn new representations for them. Therefore, textual and visual features can be fully integrated and interacted to relieve the feature discontinuity across modalities.

Specifically, we define a shared base matrix $B$ with $m$ basis vectors as $B = \{b_1, b_2, \ldots, b_m\}$, where $B \in \mathbb{R}^{m \times d_b}$ with $d_b$ representing the dimensionality of each basis vector. Besides, based on the assumption that the dimension of the shared subspace is $d_s$, visual features $X$, textual features $Y$, and shared base matrix $B$ are projected into the shared subspace respectively as

$$x_i = W_x \cdot x_i \quad & \quad \tilde{X} = X \cdot W_x, \quad (4)$$
$$y_i = W_y \cdot y_i \quad & \quad \tilde{Y} = Y \cdot W_y, \quad (5)$$
$$b_i = W_b \cdot b_i \quad & \quad \tilde{B} = B \cdot W_b, \quad (6)$$

where $W_x \in \mathbb{R}^{d_s \times d_v}$, $W_y \in \mathbb{R}^{d_s \times d_t}$, and $W_b \in \mathbb{R}^{d_s \times d_b}$ are trainable parameters.

To learn new visual and textual representations given base matrix $B$, we calculate the cosine similarity between the previous visual and textual features with $B$ as

$$S_{ij} = \tilde{x}_i^T \cdot \tilde{b}_j \quad & \quad G_{ij} = \tilde{y}_i^T \cdot \tilde{b}_j \quad (7)$$

where $T$ represents matrix transpose, $S_{ij}$ denotes the similarity between the $i$-th visual feature $\tilde{x}_i$ and the $j$-th basis vector representation $\tilde{b}_j$. Similarly, $G_{ij}$ is the similarity between the $i$-th textual feature $\tilde{y}_i$ and $\tilde{b}_j$. To prevent inaccurate representation learning caused by an excessive weight of a certain item, the similarities are further normalized by

$$S_{ij} = \frac{\exp (S_{ij})}{\sum_{k=1}^{m} \exp (S_{ik})} \quad (8)$$
$$G_{ij} = \frac{\exp (G_{ij})}{\sum_{k=1}^{m} \exp (G_{ik})}. \quad (9)$$

Finally, the new visual and textual representations are obtained as

$$r_{x_i} = \sum_{k=1}^{m} S_{ik} \cdot \tilde{b}_k \quad & \quad r_{y_i} = \sum_{k=1}^{m} G_{ik} \cdot \tilde{b}_k \quad (10)$$

where $r_{x_i}$ and $r_{y_i}$ are the $i$-th new visual feature and textual feature, respectively.

The above process guarantees the full integration between textual and visual information; that is, the visual features of a certain patch and its corresponding descriptive textual features maintain...
similar representations in the shared subspace. For example, as shown in Fig. 2, the green solid line and dotted line represent the visual and textual features of left cardiophrenic, respectively, of which ones with similar representations are gathered in the 3D shared subspace as illustrated in Fig. 2 (A).

### 3.4 Context-aware Disease Prediction

Considering that a single disease label cannot fully account for the context of an X-Ray image, including location, severity, and organs affected by a disease, mining context-aware labels for radiographs and using them to train a classification layer for disease prediction are proposed hereafter.

**Mining and pre-training on single labels.** In accordance with Shin et al. (2016), we find 17 simplest unique disease annotation patterns through statistical analysis to label the images and retain 40% of the full dataset. GoogLeNet (Szegedy et al., 2015) is used as the classification layer to train the model on the retained cases. We additionally apply mini-batch normalization (Ioffe and Szegedy, 2015) and random data dropout (Hinton et al., 2012) to alleviate result deviation caused by an unbalanced distribution between normal and pathological cases. Since the majority of disease-related MeSH terms contain up to 5 words, we constrain the GRU decoder to unroll up to 5 timesteps. Specifically, we initialize the first decoder hidden state as the output embedding of the classification layer. The GRU decoder is then trained by minimizing the negative log likelihood between the output sequence and the ground-truth:

$$
L_{\text{Loss}} = - \sum_{t=1}^{N} \{ c_t = s_t | r_{x1}, \ldots, r_{xp} \} \quad (11)
$$

where $c_t$ is the token output on the $t$-th timestep, $s_t$ is the $t$-th reference MeSH term, and $N = 5$.

**Re-training on context-aware labels.** The aforementioned classification layer and GRU decoder are considered as a pre-training procedure to mine the context for previous primary disease labels in the whole dataset. And 57 unique context-aware disease labels on the side of the output of the GRU decoder are obtained. The context-aware labels summarize both the context information and textual semantic information of the image. For example, the coarse-grained label “calcified granuloma” can be attached by more informative and detailed context as “calcified granuloma in right upper lobe” or “small calcified granuloma in left lung base”. This additional labelling procedure improves the quality of clinical practice concerning X-ray diagnosis. As shown in Fig. 2 (B), we re-train the classification layer with 57 context-aware labeled cases, and initialize the GRU hidden state with the output of the classification layer. Eq. 11 is again used as an objective function for the re-training process.

### 3.5 Automated Radiology Report Generation

As shown in Fig. 2 (C), we employ a Transformer-based encoder-decoder architecture for automated radiology report generation. New visual and textual representations are functionalized as the input for the Transformer encoder and decoder, respectively.

Since considering context-aware disease labels as macro-level features also benefits clinical report generation, we concatenate macro-level context-aware labels with micro-level visual features as the input for the Transformer encoder. Specifically, the new representation of micro-level visual features \{r_{x1}, \ldots, r_{xp}\} and macro-level context-aware label features $c$ are first fed into the encoder as

$$\{z_1, \ldots, z_p, z_c\} = f_e(r_{x1}, \ldots, r_{xp}, c), \quad (12)$$

where $f_e(\cdot)$ represents the Transformer encoder. Then, resulting intermediate state \{z_1, \ldots, z_p, z_c\} are fed into the decoder at each decoding step with aligned textual representation of the previously generated sequence \{r_{y1}, \ldots, r_{y_{t-1}}\}. The output at the $i$-th timestep can thus be generated by using

$$y_i = f_d(z_1, \ldots, z_p, z_c, r_{y1}, \ldots, r_{y_{t-1}}), \quad (13)$$

where $f_d(\cdot)$ refers to the Transformer decoder.

### 4 Experiments

#### 4.1 Dataset

We carried out our experiments on the most widely-used and conventional benchmark dataset, namely, Indiana University Chest X-Ray Collection\(^1\) (IU X-RAY) (Demner-Fushman et al., 2016). It contains

<table>
<thead>
<tr>
<th>DATASET</th>
<th>IMAGE</th>
<th>REPORT</th>
<th>PATIENT</th>
<th>AVG. LEN.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAIN</td>
<td>5,226</td>
<td>2,770</td>
<td>2,770</td>
<td>37.56</td>
</tr>
<tr>
<td>VALID</td>
<td>748</td>
<td>395</td>
<td>395</td>
<td>36.78</td>
</tr>
<tr>
<td>TEST</td>
<td>1,496</td>
<td>790</td>
<td>790</td>
<td>33.62</td>
</tr>
</tbody>
</table>

1\(^{https://openi.nlm.nih.gov/}

Table 1: Basic statistics of IU X-RAY with respect to its training, validation, and test sets. “AVG. LEN.” represents the averaged word-based length of reports.
3,955 fully de-identified handwritten radiology reports from the Indiana Network for Patient Care and 7,470 corresponding chest X-ray images from the hospitals’ picture archiving systems. As shown in Fig. 1, each sample is associated with a frontal and/or a lateral chest X-ray image, and each report is comprised of several sections: MeSH\(^3\), Indication, Findings, and Impression, etc. In this work, we use the Findings and MeSH sections as ground-truth reports and disease labels, respectively.

Following the dataset preprocessing procedure of previous studies (Li et al., 2018), we preprocess the reports by tokenizing, converting tokens into lower cases, and removing non-alphabetic tokens. Samples without MeSH or Findings sections in the dataset were excluded. We apply the same split, i.e., 70%/10%/20% for the training/validation/test set, as that stated in Li et al. (2018). The basic statistics of IU X-RAY, in terms of numbers of images, reports, patients, and average length of reports with respect to each split set, are listed in Table 1.

4.2 Baselines

The following excellent baselines are used to examine the effectiveness of the proposed approach on radiology report generation: conventional image captioning methods including Nic (Vinyals et al., 2015), ADAATT (Lu et al., 2017), ATT2IN (Rennie et al., 2017), and VisualGPT (Chen et al., 2021a); and the ones proposed for the medical domain, e.g., COATT (Jing et al., 2018), HRGR (Li et al., 2018), CMA-R-L (Jing et al., 2019), R2GEN (Chen et al., 2020), and CMN (Chen et al., 2021b). In addition, BASE is a vanilla Transformer (Vaswani et al., 2017) used as the backbone encoder-decoder architecture in our full model. We further implement several ablated versions of JPG with the aim of evaluating the different components in it.

4.3 Evaluation Metrics

The performance of the aforementioned baselines, as well as our proposed method, was evaluated by conventional natural language generation (NLG) metrics, including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and ROUGE-L (Lin, 2004), which compare model-generated reports with ground-truth by referring to the overlap of n-grams (BLEU-n), explicit word-to-word matches (METEOR), and longest common subsequence (ROUGE-L). The results based on these metrics were obtained by the standard image captioning evaluation tool\(^4\). We further measured the disease prediction subtask as a multi-label classification problem by the micro-averaged F1 score.

4.4 Implementation Details

Two X-Ray images of a patient were used as the input for both the report generation and disease annotation subtasks to ensure consistency with previous studies (Li et al., 2018; Chen et al., 2021b), where all the CNN input images were rescaled to a size of 256 × 256. We employed ResNet101 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) as the visual extractor to extract patch features with a 7 × 7 × 2048-dimension feature map. The maximum decoding sequence lengths are limited to 60 and 5 tokens for report generation and disease annotation respectively by truncating and zero-padding. 512-dimension word embeddings with random initialization were fine-tuned during training. We randomly initialized the shared subspace as a 512 × 2048 memory matrix, where \(d_s = 512\), and 2048 is the number of shared basis vectors. We adopted GoogLeNet as the classification layer, and a single-layer GRU unrolling up to five timesteps for context-aware disease label prediction. A 3-layer Transformer structure with 8 attention heads and 512-dimension hidden states was used in randomly initialized states as the encoder-decoder backbone.

Our model is trained under a cross entropy loss. As for the optimizer, Adam (Kingma and Ba, 2015) with a learning rate of \(1e-4\) and an initial accumulator value of 0.1 was used. We set the batch size to 16, whereas the target sequences were decoded through beam search with a beam size of 3 at test time to balance the effectiveness and efficiency.

5 Results and Discussion

5.1 Performance of JPG

Table 2 lists the main results on the radiology report generation task. Symbol † indicates statistically significant differences of JPG from BASE using T-test (Yang and Liu, 1999). The results for the conventional image captioning methods are shown at the top, with the ones proposed for the medical domain in the middle, and those for our methods at the bottom. According to Table 2, JPG can gen-

\(^3\)https://www.nlm.nih.gov/mesh/meshhome.html
\(^4\)https://github.com/tylin/coco-caption
Table 2: Comparison of the proposed model with those of previous studies for Findings generation on the test set of IU X-RAY with respect to various NLG metrics, where BLEU-n denotes BLEU scores using up to 4-grams. † marked results significantly surpass BASE using T-test (Yang and Liu, 1999) with $p < 0.05$.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nic (Vinyals et al., 2015)</td>
<td>0.216</td>
<td>0.124</td>
<td>0.087</td>
<td>0.066</td>
<td>-</td>
<td>0.306</td>
</tr>
<tr>
<td>ADAATT (Lu et al., 2017)</td>
<td>0.220</td>
<td>0.127</td>
<td>0.089</td>
<td>0.068</td>
<td>-</td>
<td>0.308</td>
</tr>
<tr>
<td>Attr2In (Rennie et al., 2017)</td>
<td>0.224</td>
<td>0.129</td>
<td>0.089</td>
<td>0.068</td>
<td>-</td>
<td>0.308</td>
</tr>
<tr>
<td>VisualGPT (Chen et al., 2021a)</td>
<td>0.482</td>
<td>0.314</td>
<td>0.221</td>
<td>0.158</td>
<td>0.204</td>
<td>0.375</td>
</tr>
<tr>
<td>CoATT (Jing et al., 2018)</td>
<td>0.455</td>
<td>0.288</td>
<td>0.205</td>
<td>0.154</td>
<td>-</td>
<td>0.369</td>
</tr>
<tr>
<td>HRGR (Li et al., 2018)</td>
<td>0.438</td>
<td>0.298</td>
<td>0.208</td>
<td>0.151</td>
<td>-</td>
<td>0.322</td>
</tr>
<tr>
<td>CMAS-RL (Jing et al., 2019)</td>
<td>0.464</td>
<td>0.301</td>
<td>0.210</td>
<td>0.154</td>
<td>-</td>
<td>0.362</td>
</tr>
<tr>
<td>R2GEN (Chen et al., 2020)</td>
<td>0.470</td>
<td>0.304</td>
<td>0.219</td>
<td>0.165</td>
<td>0.187</td>
<td>0.371</td>
</tr>
<tr>
<td>CMN (Chen et al., 2021b)</td>
<td>0.475</td>
<td>0.309</td>
<td>0.222</td>
<td>0.170</td>
<td>0.191</td>
<td>0.375</td>
</tr>
<tr>
<td>BASE</td>
<td>0.369</td>
<td>0.254</td>
<td>0.179</td>
<td>0.135</td>
<td>0.164</td>
<td>0.342</td>
</tr>
<tr>
<td>JPG-projection</td>
<td>0.458</td>
<td>0.291</td>
<td>0.212</td>
<td>0.159</td>
<td>0.177</td>
<td>0.371</td>
</tr>
<tr>
<td>JPG-auxiliary</td>
<td>0.472</td>
<td>0.308</td>
<td>0.218</td>
<td>0.168</td>
<td>0.188</td>
<td>0.373</td>
</tr>
<tr>
<td>JPG</td>
<td><strong>0.479†</strong></td>
<td><strong>0.319†</strong></td>
<td><strong>0.222†</strong></td>
<td><strong>0.174†</strong></td>
<td><strong>0.193†</strong></td>
<td><strong>0.377†</strong></td>
</tr>
</tbody>
</table>

Figure 3: Classification accuracy of AlexNet, NIN, and GoogLeNet on the test set of IU X-RAY.

We consider three possible reasons for the superior performance of JPG. First, the shared subspace is configured to make up for the gap between different information extracted by word and image embeddings. Compared to simply merging word embeddings of disease tags into patch features as complementary textual information, the additional shared subspace projection makes the aligned visual and textual features much more understandable to each other, so that information is better interacted, and the quality of reports is improved. Second, regarding the improvement of BLEU scores, the introduction of context-aware disease labels provides the report generation process with explicit lesion textual prompts, which prevents our model from generating irrelevant diseases and enables JPG to effectively capture the disease-related keywords. Third, conclusive disease prediction and descriptive report generation are jointly trained and optimized in an overall framework to obtain a globally optimal solution for both subtasks.

As shown in Fig. 3, the three most effective classification networks, AlexNet (Krizhevsky et al., 2012), NIN (Lin et al., 2014), and GoogLeNet (Szegedy et al., 2015) were employed for classification with context-aware disease labels. Compared with adopting patch features directly extracted from the visual extractor, learning new visual representations from a shared subspace can dramatically improve classification accuracy, because new visual representations contain more useful semantic features in regard to the classification task. Therefore, many inspiring context-aware disease labels, such as <opacity lung bilateral interstitial diffuse> and <opacity lung lower_lobe bilateral>, can be obtained.

5.2 Ablation Study

**JPG-projection** To verify the alignment between visual and textual representations within the encoder-decoder architecture, we show the ablation performance in Table 2 by removing the shared
subspace projection and simply using the raw visual extractor and word embedding outputs to both predict disease labels and generate reports, which obviously degrades the model performance with respect to all evaluation metrics. This proves that the shared base matrix plays a critical role in facilitating disease prediction and report generation with sufficient understandable visual representations with semantic meanings, which cannot be replaced by straightforward visual and textual features. Besides, instead of using hard attention to match visual features with textual features, the proposed shared subspace acts as a soft alignment medium to offset the gap between those features; it thus unifies cross-modal features within the same representation space. Furthermore, the shared subspace also provides further fusion patterns for disease tags and chest X-Ray images to communicate with each other and pass both compatible visual and textual information for more accurate reports.

**JPG-auxiliary** Based on the assumption that the remarkable improvement of JPG from baselines is due to jointly training disease prediction and report generation and employing the predicted disease tags as auxiliary information when generating reports, we would like to experimentally evaluate the performance of JPG in terms of a separate learning pattern. In this experiment, the disease prediction and report generation subtasks were treated as two parallel procedures. Specifically, visual and textual features were first projected into a shared subspace to overcome the misalignment of features across modalities. Then, we independently employed a Transformer encoder-decoder structure for report generation without adding context-aware disease labels as auxiliary information on the input side.

According to the last block in Table 2, implementing the subtasks individually degrades the model performance and the quality of generated reports to a certain extent. We consider that in our complimentary interactive learning framework, reports can receive more discriminative lesion locations and semantic features under the guidance and constraint of predicted diseases. And in turn disease prediction accuracy is improved by report generation via visual feature extraction and fusion in a manner that cannot be imitated by separate learning. This result indicates the superiority of JPG over the conventional methods, implying the usage of auxiliary disease tags in the report generation process is promising for identifying salient keywords.

### 5.3 Alignment Visualization and Case Study

To further qualitatively investigate the ability of JPG to overcome the misalignment of features across modalities, Fig. 4 visualizes how the proposed model focuses on the image when generating a certain word or phrase; i.e., it learns from the alignments between visual and textual features. We randomly select an example from the IU-XRAY dataset, and list its original chest X-Ray image with the corresponding ground-truth report for reference. Fig. 4 shows image-text mappings between particular regions (highlighted by colored weights).
of an X-Ray image and words/phrases from its reports generated by BASE and JPG. In detail, we utilize the cross attention weight from the first decoder layer to show the alignment between visual and textual features, since the latter decoder layers couple the textual and visual information, making it difficult to distinguish the most primitive alignment weights. In general, JPG is able to pay attention to relatively accurate patches when generating a word (especially disease terminologies), so it brings about descriptions of higher quality than those produced by BASE.

Fig. 5 exhibits two examples with both front and lateral CXR images and their corresponding reports obtained by ground-truth, BASE, and JPG, where different colors on the texts indicate different clinical terms. These examples indicate that JPG can produce accurate terms and well-aligned descriptions, which abide by a similar content flow as radiologists follow, while BASE sometimes makes factual errors. For example, in both cases, patterns in the ground-truth and generated reports follow the sequence of starting from observations (e.g., “lung volumes” and “cardiomedastinal silhouette”) and concluding with potential diseases (e.g., “pleural effusion” and “pneumothorax”). In addition, JPG-generated reports cover almost all of the necessary clinical terminologies in the ground-truth reports. On the contrary, BASE cannot keep abreast with the description order of the ground-truth, so it generates misaligned and out-of-order sentences. Moreover, several phrases go against fact; e.g., “small pleural effusions” is mistakenly ignored. By longitudinally viewing the reports produced by BASE corresponding to two cases, we can also find that the vanilla Transformer tends to iteratively generate similar sentences.

6 Conclusion

We addressed several fundamental issues concerning clinical disease prediction and radiology report generation in an overall framework, where context-aware disease terminologies act as complementary textual features coupled with visual features of images to guide and facilitate the report generation process. Meanwhile, these explicit clues of lesion location effectively prevent the report generation model from generating factual erroneous texts. The proposed shared subspace provides an interaction platform for different representations extracted by image and word embeddings to overcome the misalignment of information across modalities. Empirical results acquired with the most widely used dataset, including those of ablation studies, demonstrate the effectiveness of the proposed JPG, which achieves the state-of-the-art performance.

Acknowledgements

We would like to gratefully thank the anonymous reviewers for their helpful comments and feedback. Jingyi You and Dongyuan Li acknowledge the support from China Scholarship Council (CSC).

References


