

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

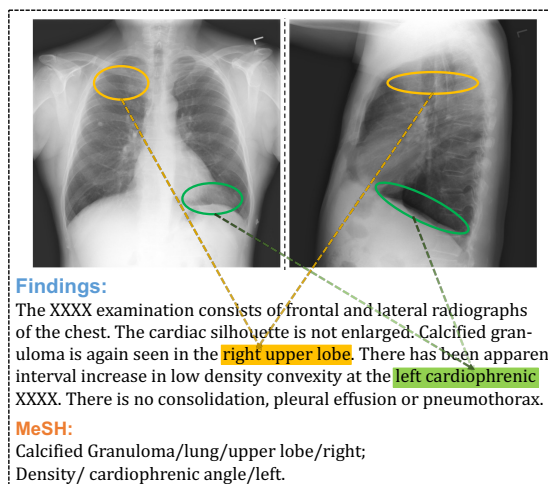


Figure 1: Chest X-ray images and an accompanying report, including *Findings* and *MeSH* labels, from the IU X-RAY dataset. We marked the aligned visual and textual features in different colors for better illustration.

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, current 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) *vari-*

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 \mathbf{I} as input, JPG is designed to automatically generate a sequence of context-

aware disease labels \mathbf{c} and a radiology report \mathbf{Y} . Specifically, we divide \mathbf{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 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$, where $\mathbf{x}_p \in \mathbb{R}^{d_x}$ with d_x representing the dimensionality of patch features. The target output is the corresponding radiology report $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$, where $\mathbf{y}_n \in \mathbb{R}^{d_y}$ 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:

$$p(\mathbf{Y}, \mathbf{c} | \mathbf{X}) \propto p(\mathbf{Y} | \mathbf{c}, \mathbf{X}) \cdot p(\mathbf{c} | \mathbf{X}), \quad (1)$$

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

$$p(\mathbf{Y} | \mathbf{c}, \mathbf{X}) = \prod_{i=1}^n p(\mathbf{y}_i | \mathbf{y}_{<i}, \mathbf{c}, \mathbf{X}), \quad (2)$$

where $\mathbf{y}_{<i} = \{\mathbf{y}_1, \dots, \mathbf{y}_{i-1}\}$ represents the previously generated tokens so far, and n is the total amount of tokens in target sequence \mathbf{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 $p(\mathbf{c} | \mathbf{X})$, then maximize the probability of generating a corresponding radiology report $p(\mathbf{Y} | \mathbf{c}, \mathbf{X})$ conditioned on context-aware disease labels \mathbf{c} and visual features \mathbf{X} .

3.2 Visual Extractor

As shown in Fig. 2, given a radiology image \mathbf{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 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ are concatenated to constitute the source input for all subsequent modules with the form of visual feature sequence $\mathbf{X} \in \mathbb{R}^{p \times d_x}$ as

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\} = f_v(\mathbf{I}). \quad (3)$$

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.

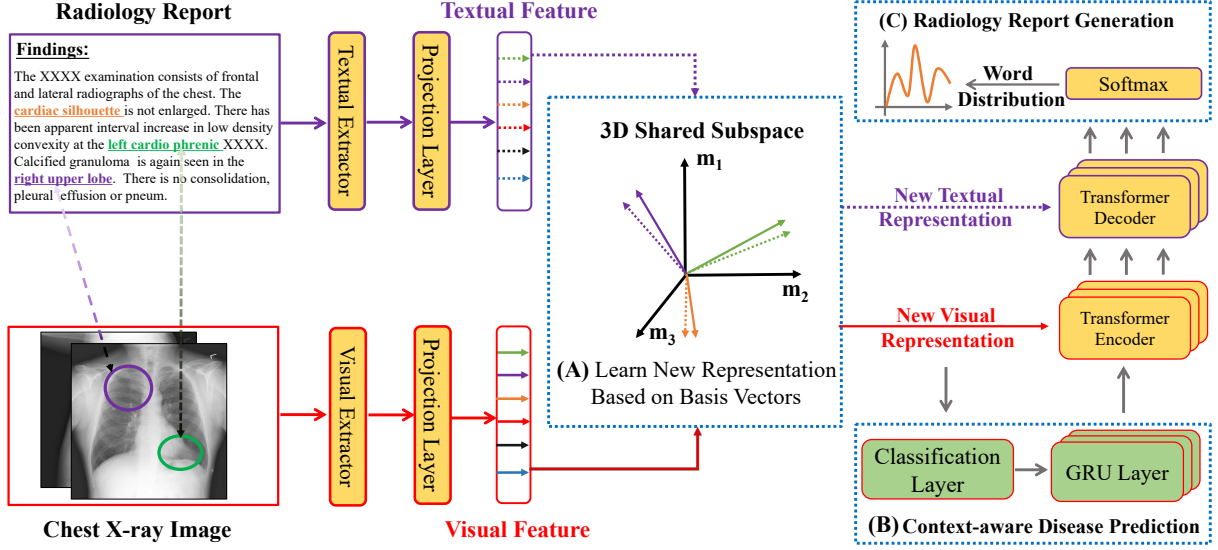


Figure 2: Model overview. JPG first captures textual and visual features through word embeddings and a visual extractor. Those features are then projected onto a shared subspace ((A) as a 3-dimension example) to learn new representations based on shared basis vectors. Finally, a RNN decoder (B) and a Transformer-based encoder-decoder architecture (C) are employed to generate context-aware disease labels and radiology reports, respectively.

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 \mathbf{B} with m basis vectors as $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_m\}$, where $\mathbf{B} \in \mathbb{R}^{m \times d_b}$ 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 \mathbf{X} , textual features \mathbf{Y} , and shared base matrix \mathbf{B} are projected into the shared subspace respectively as

$$\tilde{\mathbf{x}}_i = \mathbf{W}_x \cdot \mathbf{x}_i \quad \& \quad \tilde{\mathbf{X}} = \mathbf{X} \cdot \mathbf{W}_x, \quad (4)$$

$$\tilde{\mathbf{y}}_i = \mathbf{W}_y \cdot \mathbf{y}_i \quad \& \quad \tilde{\mathbf{Y}} = \mathbf{Y} \cdot \mathbf{W}_y, \quad (5)$$

$$\tilde{\mathbf{b}}_i = \mathbf{W}_b \cdot \mathbf{b}_i \quad \& \quad \tilde{\mathbf{B}} = \mathbf{B} \cdot \mathbf{W}_b, \quad (6)$$

where $\mathbf{W}_x \in \mathbb{R}^{d_x \times d_s}$, $\mathbf{W}_y \in \mathbb{R}^{d_y \times d_s}$, and $\mathbf{W}_b \in \mathbb{R}^{d_b \times d_s}$ are trainable parameters.

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

$$\mathbf{S}_{ij} = \tilde{\mathbf{x}}_i^T \cdot \tilde{\mathbf{b}}_j \quad \& \quad \mathbf{G}_{ij} = \tilde{\mathbf{y}}_i^T \cdot \tilde{\mathbf{b}}_j \quad (7)$$

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

$$\mathbf{S}_{ij} = \frac{\exp(\mathbf{S}_{ij})}{\sum_{k=1}^m \exp(\mathbf{S}_{ik})} \quad (8)$$

$$\mathbf{G}_{ij} = \frac{\exp(\mathbf{G}_{ij})}{\sum_{k=1}^m \exp(\mathbf{G}_{ik})}. \quad (9)$$

Finally, the new visual and textual representations are obtained as

$$\mathbf{r}_{x_i} = \sum_{k=1}^m \mathbf{S}_{ik} \cdot \tilde{\mathbf{b}}_k \quad \& \quad \mathbf{r}_{y_i} = \sum_{k=1}^m \mathbf{G}_{ik} \cdot \tilde{\mathbf{b}}_k \quad (10)$$

where \mathbf{r}_{x_i} and \mathbf{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

DATASET	IMAGE	REPORT	PATIENT	AVG. LEN.
TRAIN	5,226	2,770	2,770	37.56
VALID	748	395	395	36.78
TEST	1,496	790	790	33.62

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.

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:

$$\mathcal{L}_{Loss} = - \sum_{t=1}^N \{c_t = s_t | \mathbf{r}_{x_1}, \dots, \mathbf{r}_{x_p}\} \quad (11)$$

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 $\{\mathbf{r}_{x_1}, \dots, \mathbf{r}_{x_p}\}$ and macro-level context-aware label features \mathbf{c} are first fed into the encoder as

$$\{\mathbf{z}_1, \dots, \mathbf{z}_p, \mathbf{z}_c\} = f_e(\mathbf{r}_{x_1}, \dots, \mathbf{r}_{x_p}, \mathbf{c}), \quad (12)$$

where $f_e(\cdot)$ represents the Transformer encoder. Then, resulting intermediate state $\{\mathbf{z}_1, \dots, \mathbf{z}_p, \mathbf{z}_c\}$ are fed into the decoder at each decoding step with aligned textual representation of the previously generated sequence $\{\mathbf{r}_{y_1}, \dots, \mathbf{r}_{y_{i-1}}\}$. The output at the i -th timestep can thus be generated by using

$$\mathbf{y}_i = f_d(\mathbf{z}_1, \dots, \mathbf{z}_p, \mathbf{z}_c, \mathbf{r}_{y_1}, \dots, \mathbf{r}_{y_{i-1}}), \quad (13)$$

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 Collection¹ (IU X-RAY) (Demner-Fushman et al., 2016). It contains

¹<https://openi.nlm.nih.gov/>

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*², *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), CMAS-RL (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 tool³. 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 $7 \times 7 \times 2048$ -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-

²<https://www.nlm.nih.gov/mesh/meshhome.html>

³<https://github.com/tylin/coco-caption>

METHOD	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
NIC (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	-	0.306
ADAATT (Lu et al., 2017)	0.220	0.127	0.089	0.068	-	0.308
ATT2IN (Rennie et al., 2017)	0.224	0.129	0.089	0.068	-	0.308
VisualGPT (Chen et al., 2021a)	0.482	0.314	0.221	0.158	0.204	0.375
COATT (Jing et al., 2018)	0.455	0.288	0.205	0.154	-	0.369
HRGR (Li et al., 2018)	0.438	0.298	0.208	0.151	-	0.322
CMAS-RL (Jing et al., 2019)	0.464	0.301	0.210	0.154	-	0.362
R2GEN (Chen et al., 2020)	0.470	0.304	0.219	0.165	0.187	0.371
CMN (Chen et al., 2021b)	0.475	0.309	0.222	0.170	0.191	0.375
BASE	0.369	0.254	0.179	0.135	0.164	0.342
JPG-projection	0.458	0.291	0.212	0.159	0.177	0.371
JPG-auxiliary	0.472	0.308	0.218	0.168	0.188	0.373
JPG	0.479[†]	0.319[†]	0.222[†]	0.174[†]	0.193[†]	0.377[†]

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 n-grams. [†] marked results significantly surpass BASE using T-test (Yang and Liu, 1999) with $p < 0.05$.

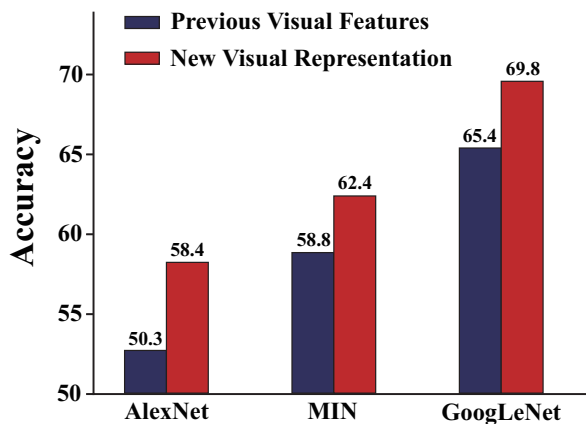


Figure 3: Classification accuracy of AlexNet, NIN, and GoogLeNet on the test set of IU X-RAY.

erate more accurate and fluent radiology reports compared with the baselines.

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

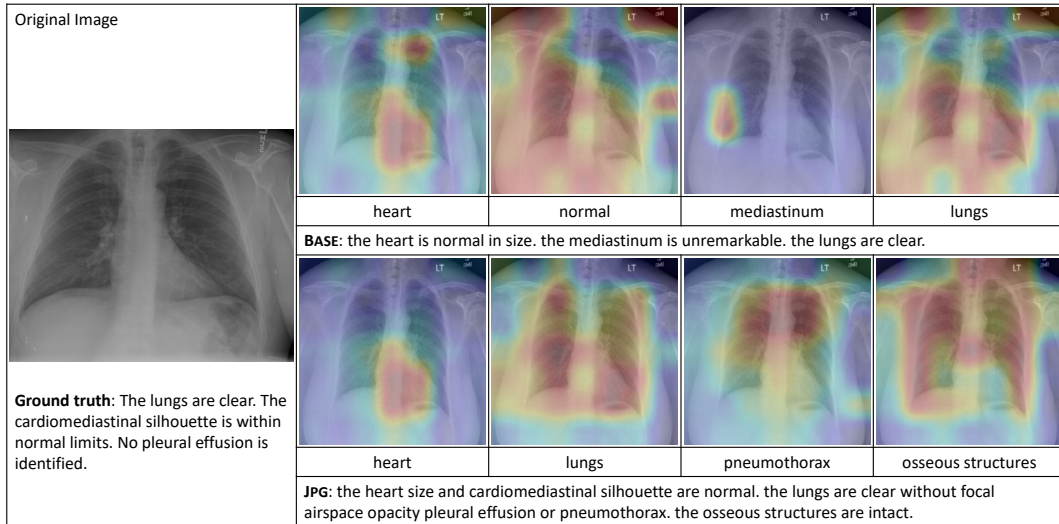


Figure 4: Visualization of image-text mappings between particular regions (indicated by colored weights) of a chest X-ray image and tokens from its reports generated by BASE and JPG, respectively.

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)


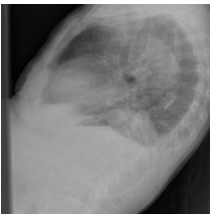
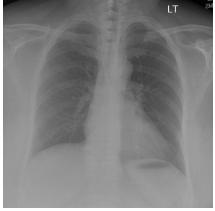
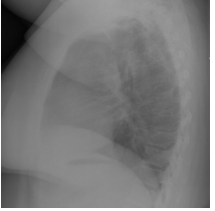
Frontal	Lateral	Ground-truth	BASE	JPG
		Low lung volumes. Stable ectasia of the thoracic aorta. Stable right upper mediastinal. Bilateral small pleural effusions and bibasilar airspace opacities. The heart size and mediastinal silhouette are within normal limits for contour. No pneumothorax. Stable wedging of the anterior thoracic vertebral bodies.	Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality. Low lung volumes bilaterally. The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion.	There are low lung volumes with bibasilar opacities representing subsegmental atelectasis. The cardio the cardiac silhouette is of the xxx of normal in size and contour. There is no pneumothorax or large pleural effusion.
		The lungs are clear. The cardiome-diastinal silhouette is within normal limits. No pleural effusion is identified.	The lungs are clear bilaterally. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.	The heart size and cardiome-diastinal silhouette are normal. The lungs are clear without focal airspace opacity pleural effusion or pneumothorax. The osseous structures are intact.

Figure 5: Example reports of BASE and JPG.

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 “cardiome-diastinal 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

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

- 2019, Long Beach, CA, USA, June 16-20, 2019, pages 12476–12486. Computer Vision Foundation / IEEE.
- Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna. 2022. [Radiotransformer: A cascaded global-focal transformer for visual attention-guided disease classification](#). *CoRR*, abs/2202.11781.
- Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2021a. [Visualgpt: Data-efficient image captioning by balancing visual input and linguistic knowledge from pretraining](#). *CoRR*, abs/2102.10407.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021b. [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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5904–5914. 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 2020, Online, November 16-20, 2020*, pages 1439–1449. Association for Computational Linguistics.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. [Meshed-memory transformer for image captioning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. Computer Vision Foundation / IEEE.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. [Preparing a collection of radiology examinations for distribution and retrieval](#). *J. Am. Medical Informatics Assoc.*, 23(2):304–310.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.
- Michael J. Denkowski and Alon Lavie. 2011. [Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems](#). In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 85–91. Association for Computational Linguistics.
- Behnaz Gheflati and Hassan Rivaz. 2021. [Vision transformer for classification of breast ultrasound images](#). *CoRR*, abs/2110.14731.
- Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. [Contrastive learning for weakly supervised phrase grounding](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 752–768. Springer.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.
- Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. [Improving neural networks by preventing co-adaptation of feature detectors](#). *CoRR*, abs/1207.0580.
- Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. [Language-conditioned graph networks for relational reasoning](#). In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 10293–10302. IEEE.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. [Densely connected convolutional networks](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society.
- Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. [Attention on attention for image captioning](#). In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4633–4642. IEEE.
- Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander G. Hauptmann. 2020. [Unsupervised multimodal neural machine translation with pseudo visual pivoting](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8226–8237. Association for Computational Linguistics.
- Sergey Ioffe and Christian Szegedy. 2015. [Batch normalization: Accelerating deep network training by reducing internal covariate shift](#). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org.
- Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. [Improving image captioning by leveraging intra- and inter-layer global representation in transformer network](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational*

- Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1655–1663. AAAI Press.
- Baoyu Jing, Zeya Wang, and Eric P. Xing. 2019. [Show, describe and conclude: On exploiting the structure information of chest x-ray reports](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6570–6580. Association for Computational Linguistics.
- Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. [On the automatic generation of medical imaging reports](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2577–2586. Association for Computational Linguistics.
- Eun-Sol Kim, Woo-Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. 2020. [Hypergraph attention networks for multimodal learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14569–14578. Computer Vision Foundation / IEEE.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. [Imagenet classification with deep convolutional neural networks](#). In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114.
- 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 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1537–1547.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Min Lin, Qiang Chen, and Shuicheng Yan. 2014. [Network in network](#). In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Fenglin Liu, Shen Ge, and Xian Wu. 2021a. [Competence-based multimodal curriculum learning for medical 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3001–3012. Association for Computational Linguistics.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. [Exploring and distilling posterior and prior knowledge for radiology report generation](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13753–13762. Computer Vision Foundation / IEEE.
- Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. 2021c. [Auto-encoding knowledge graph for unsupervised medical report generation](#). *CoRR*, abs/2111.04318.
- Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. [Clinically accurate chest x-ray report generation](#). In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2019, 9-10 August 2019, Ann Arbor, Michigan, USA*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. [Knowing when to look: Adaptive attention via a visual sentinel for image captioning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250. IEEE Computer Society.
- Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5288–5304. Association for Computational Linguistics.
- Hoang T. N. Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. [Automated generation of accurate & fluent medical x-ray reports](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3552–3569. Association for Computational Linguistics.
- 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

- Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, Chang Min Park, and Jong Chul Ye. 2022. [AI can evolve without labels: self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation](#). *CoRR*, abs/2202.06431.
- Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. [Self-critical sequence training for image captioning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.
- Mert Bülent Sariyildiz, Julien Perez, and Diane Larlus. 2020. [Learning visual representations with caption annotations](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII*, volume 12353 of *Lecture Notes in Computer Science*, pages 153–170. Springer.
- Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. 2021. [Transmil: Transformer based correlated multiple instance learning for whole slide image classification](#). *CoRR*, abs/2106.00908.
- Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. [Enhancing descriptive image captioning with natural language inference](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 269–277. Association for Computational Linguistics.
- Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. 2016. [Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2497–2506. IEEE Computer Society.
- Karen Simonyan and Andrew Zisserman. 2015. [Very deep convolutional networks for large-scale image recognition](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. 2021. [Lesion-aware transformers for diabetic retinopathy grading](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10938–10947. Computer Vision Foundation / IEEE.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. [Going deeper with convolutions](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. [Show and tell: A neural image caption generator](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.
- Song Wang, Liyan Tang, Mingquan Lin, George Shih, Ying Ding, and Yifan Peng. 2022. [Prior knowledge enhances radiology report generation](#). *CoRR*, abs/2201.03761.
- Xiaomei Wang, Lin Ma, Yanwei Fu, and Xiangyang Xue. 2021a. [Neural symbolic representation learning for image captioning](#). In *ICMR '21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021*, pages 312–321. ACM.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. [Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9049–9058. Computer Vision Foundation / IEEE Computer Society.
- Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. 2021b. [A self-boosting framework for automated radiographic report generation](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2433–2442. Computer Vision Foundation / IEEE.
- Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. 2021. [Control image captioning spatially and temporally](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2014–2025. Association for Computational Linguistics.
- Yiming Yang and Xin Liu. 1999. [A re-examination of text categorization methods](#). In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 42–49. ACM.

- Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. [Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network](#). In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 728–737. IEEE.
- Jingyi You, Chenlong Hu, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2021a. [Robust dynamic clustering for temporal networks](#). In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2424–2433. ACM.
- Jingyi You, Chenlong Hu, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2021b. [Abstractive document summarization with word embedding reconstruction](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, pages 1586–1596. INCOMA Ltd.
- Jingyi You, Dongyuan Li, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2022. [Joint learning-based heterogeneous graph attention network for timeline summarization](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4091–4104. Association for Computational Linguistics.
- Quanzeng You, Zhengyou Zhang, and Jiebo Luo. 2018. [End-to-end convolutional semantic embeddings](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5735–5744. Computer Vision Foundation / IEEE Computer Society.
- Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. 2020. [When radiology report generation meets knowledge graph](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12910–12917. AAAI Press.
- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. [Unified vision-language pre-training for image captioning and VQA](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.