

METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens

Zhanyu Wang
University of Sydney
zhanyu.wang@sydney.edu.au

Lingqiao Liu
University of Adelaide
lingqiao.liu@adelaide.edu.au

Lei Wang
University of Wollongong
leiw@uow.edu.au

Luping Zhou
University of Sydney
luping.zhou@sydney.edu.au

Abstract

In clinical scenarios, multi-specialist consultation could significantly benefit the diagnosis, especially for intricate cases. This inspires us to explore a “multi-expert joint diagnosis” mechanism to upgrade the existing “single expert” framework commonly seen in the current literature. To this end, we propose METransformer, a method to realize this idea with a transformer-based backbone. The key design of our method is the introduction of multiple learnable “expert” tokens into both the transformer encoder and decoder. In the encoder, each expert token interacts with both vision tokens and other expert tokens to learn to attend different image regions for image representation. These expert tokens are encouraged to capture complementary information by an orthogonal loss that minimizes their overlap. In the decoder, each attended expert token guides the cross-attention between input words and visual tokens, thus influencing the generated report. A metrics-based expert voting strategy is further developed to generate the final report. By the multi-experts concept, our model enjoys the merits of an ensemble-based approach but through a manner that is computationally more efficient and supports more sophisticated interactions among experts. Experimental results demonstrate the promising performance of our proposed model on two widely used benchmarks. Last but not least, the framework-level innovation makes our work ready to incorporate advances on existing “single-expert” models to further improve its performance.

1. Introduction

Interpreting radiology images (e.g., chest X-ray) and writing diagnostic reports are essential operations in clinical practice and normally require a considerable manual work-

load. Therefore, radiology report generation, which aims to automatically generate a free-text description based on a radiograph, is highly desired to ease the burden of radiologists while maintaining the quality of health care. Recently, substantial progress has been made towards research on automated radiology report generation models [4, 5, 15–17, 21, 22, 33–35, 41, 43, 44, 47]. Most existing studies adopt a conventional encoder-decoder architecture following the image captioning paradigm [6, 25, 32, 37, 48] and resort to optimizing network structure or introducing external or prior information to aid report generation. These methods, in this paper, are collectively referred to as “single-expert” based diagnostic captioning methods.

However, diagnostic report generation is a very challenging task as disease anomalies usually only occupy a small portion of the whole image and could appear at arbitrary locations. Due to the fine-grained nature of radiology images, it is hard to focus on the correct image regions throughout the report generation procedure despite different attentions developed in recent works [16, 21]. Meanwhile, it is noticed that in clinic scenarios, multi-specialist consultation is especially beneficial for those intricate diagnostic cases that challenge a single specialist for a comprehensive and accurate diagnosis. The above observations led us to think, could we design a model to simulate the multi-specialist consultation scenario? Based on this motivation, we propose a new diagnostic captioning framework, METransformer, to mimic the “multi-expert joint diagnosis” process. Built upon a transformer backbone, METransformer introduces multiple “expert tokens”, representing multiple experts, into both the transformer encoder and decoder. Each expert token learns to attend distinct image regions and interacts with other expert tokens to capture reliable and complementary visual information and produce a diagnosis report in parallel. The optimal report is selected through an expert voting strategy to produce the final report. Our design is based on

the assumption that it would be easier for multiple experts than a single one to capture visual patterns of clinical importance, which is verified by our experimental results.

Specifically, we feed both the expert tokens (learnable embeddings) and the visual tokens (image patches embeddings) into the Expert Transformer encoder which is comprised of a vision transformer (ViT) encoder and a bilinear transformer encoder. In ViT encoder, each expert token interacts not only with the visual tokens but also with the other expert tokens. Further, in the bilinear transformer encoder, to enable each “expert” to capture fine-grained image information, we compute higher-order attention between expert tokens and visual tokens, which has proved to be effective in fine-grained classification tasks [20]. It is noteworthy that the expert tokens in the encoder are encouraged to learn complementary representations by an orthogonal loss so that they attend differently to different image regions. With these carefully learned expert token embeddings, in the decoder, we take them as a guide to regulate the learning of word embeddings and visual token embedding in the report generation process. This results in M different word and visual token embeddings, thus producing M candidate reports, where M is the number of experts. We further propose a metric-based expert voting strategy to generate the final report from the M candidates.

By utilizing multiple experts, our model, to some extent, is analogous to ensemble-based approaches, while each expert token corresponds to an individual model. While it enjoys the merits of ensemble-based approaches, our model is designed in a manner that is computationally more efficient and supports more sophisticated interactions among experts. Therefore, it can scale up with only a trivial increase of model parameters and achieves better performance, as demonstrated in our experimental study.

Our main contributions are summarized as follows.

First, we propose a new diagnostic captioning framework, METransformer, which is conceptually “multi-expert joint diagnosis” for radiology report generation, by introducing learnable expert tokens and encouraging them to learn complementary representations using both linear and non-linear attentions.

Second, our model enjoys the benefits of an ensemble approach. Thanks to the carefully designed network structure and the end-to-end training manner, our model can achieve better results than common ensemble approaches while greatly reducing training parameters and improving training efficiency.

Third, our approach shows promising performance on two widely used benchmarks IU-Xray and MIMIC-CXR over multiple state-of-the-art methods. The clinic relevance of the generated reports is also analyzed.

2. Related Work

Image Captioning. Natural image captioning task aims at automatically generating a single sentence description for the given image. A broad collection of methods was proposed in the last few years [1, 6, 24, 25, 28, 32, 37, 42, 45, 48] and have achieved great success in advancing the state-of-the-art. Most of them adopt the conventional encoder-decoder architecture, with convolutional neural networks (CNNs) as the encoder and recurrent (e.g., LSTM/GRU) or non-recurrent networks (e.g., Transformer) as the decoder with a carefully designed attention module [24, 25, 37] to produce image description. However, compare with image captioning, radiographic images have fine-grained characteristics and radiology report generation aims to generate a long paragraph rather than a single sentence, which brings more challenge to the model’s attention ability.

Medical report generation. Most of the research efforts in medical report generation can be roughly categorized as being along two main directions. The first direction lies in promoting the model’s structure, such as introducing a better attention mechanism or improving the structure of the report decoder. For example, some works [12, 35, 38, 39, 43, 46] utilize a hierarchically structured LSTM network to better handle the long narrative nature of medical reports. Jing et al [12] proposed a multi-task hierarchical model with co-attention by automatically predicting keywords to assist in generating long paragraphs. Xue et al [38, 39] presented a different network structure involving a generative sentence model and a generative paragraph model that uses the generated sentence to produce the next sentence. In addition, Wang et al [35] introduced an image-report matching network to bridge the domain gap between image and text for reducing the difficulty of report generation. To further improve performance, some works [4, 5, 33] employ a transformer instead of LSTM as the report decoder, which has achieved good results. Work [5] proposes to generate radiographic reports with a memory-driven Transformer designed to record key information of the generation process. The second research direction studies how to leverage medical domain knowledge to guide report generation. Most recently, many works [16, 21, 22, 40, 41, 47] attempt to integrate knowledge graphs into medical report generation pipeline to improve the quality of the generated reports. Another group of works [34, 44] utilizes disease tags to facilitate report generation. Yang et al [41] present a framework based on both general and specific knowledge, where the general knowledge is obtained from a pre-constructed knowledge graph, while the specific knowledge is derived from retrieving similar reports. The work [34] proposed a medical concepts generation network to generate semantic information and integrate it into the report generation process. It is worth mentioning that our approach is orthogonal to the methods mentioned above, for example, the advanced

memory-enhanced decoder used in [5] can also be applied to our framework for further performance improvement.

3. Method

As shown in Figure.1, our encoder comprises a vision transformer (ViT) [9] encoder and an expert bilinear transformer encoder. The ViT encoder takes both the expert and the visual tokens as the input and computes the linear attention between every two tokens. The encoded tokens are further sent into the expert bilinear transformer encoder where high-order interactions are computed between the expert tokens and the visual tokens. The enhanced expert tokens are then used to regulate the embeddings of visual and word tokens by an adjust block and sent into the expert decoder to produce expert-specific reports in parallel, and then an expert voting strategy is used to generate the final report.

3.1. Multi-expert ViT Encoder

Our ViT encoder adopts vision transformer [9]. In addition to the common input of visual patches/tokens embedding and position embedding, we further introduce expert tokens embedding and segment embedding.

Visual patches Embedding. Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, followed [9], we reshape \mathbf{x} into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \dot{C})}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. We consider \mathbf{x}_p as visual tokens and N is the input sequence length.

Expert tokens Embedding. In addition to N visual tokens, we post-pend M learnable embeddings $\mathbf{x}_e \in \mathbb{R}^{M \times D}$ which have the same dimension as the visual token embeddings and are called expert tokens. We introduce an Orthogonal Loss as in Eqn. 6 to enforce orthogonality among the expert token embeddings, encouraging different expert tokens to attend different image regions.

Segment Embedding. We introduce two types of the segment, “[0]” and “[1]”, to separate the input tokens from different sources [29], i.e., “[0]” for visual tokens and “[1]” for expert tokens. A learned segment embedding \mathbf{E}_{seg} is added to each input token to indicate its segment type.

Position Embedding. A standard 1D learnable position embedding \mathbf{E}_{pos} is added to each input token to indicate its order in the input sequence.

Model Structure. The expert ViT encoder adopts the same structure as the standard Transformer [30], which consists of alternating layers of multi-headed self-attention (MSA) and Multi-Layer Perceptron blocks (MLP). Layer-norm(LN) [2] is applied before every block and residual connections are applied after every block. Mathematically:

$$\begin{aligned} \mathbf{z}_0 &= [\mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}; \mathbf{x}_e^1; \mathbf{x}_e^2; \dots; \mathbf{x}_e^M] \\ &\quad + \mathbf{E}_{pos} + \mathbf{E}_{seg}, \\ \hat{\mathbf{z}}_l &= \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \\ \mathbf{z}_l &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}_l)) + \hat{\mathbf{z}}_l, \end{aligned} \quad (1)$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 C) \times D}$ is a learnable matrix parameter to map visual tokens to a fixed dimension D . The subscript $l = 1 \dots L$ and L is the total number of the transformer layers. $\mathbf{E}_{pos} \in \mathbb{R}^{(N+M) \times D}$ is the position embeddings and $\mathbf{E}_{seg} \in \mathbb{R}^{(N+M) \times D}$ is the segment embeddings.

3.2. Multi-expert Bilinear Attention Encoder

The output of our expert ViT encoder, $\mathbf{z}_L \in \mathbb{R}^{(N+M) \times D}$, combines the visual token embeddings $\mathbf{z}_L^v = \mathbf{z}_L [: N]$ and the expert token embeddings $\mathbf{z}_L^e = \mathbf{z}_L [N : (N+M)]$. They have attended to each other using multi-head self-attention, which is linear attention. Considering the fine-grained nature of medical images and the effectiveness of high-order attention in fine-grained recognition [20], we further enhance both embeddings by expert bilinear attention.

Expert Bilinear Attention. We followed [25] to design our expert bilinear attention (EBA) module. As shown in Figure. 1 Left, we take the enhanced expert token embeddings \mathbf{z}_L^e as query $\mathbf{Q} \in \mathbb{R}^{M \times 1 \times D_q}$, the enhanced visual tokens embeddings \mathbf{z}_L^v as key $\mathbf{K} \in \mathbb{R}^{1 \times N \times D_k}$ and value $\mathbf{V} \in \mathbb{R}^{1 \times N \times D_v}$, a low-rank bilinear pooling [14] if first performed to obtain the joint bilinear query-key \mathbf{B}_k and query-value \mathbf{B}_v by $\mathbf{B}_k = \sigma(\mathbf{W}_k \mathbf{K}) \odot \sigma(\mathbf{W}_q^k \mathbf{Q})$ and $\mathbf{B}_v = \sigma(\mathbf{W}_v \mathbf{V}) \odot \sigma(\mathbf{W}_q^v \mathbf{Q})$, where $\mathbf{W}_k \in \mathbb{R}^{D_B \times D_k}$, $\mathbf{W}_v \in \mathbb{R}^{D_B \times D_v}$, $\mathbf{W}_q^k \in \mathbb{R}^{D_B \times D_q}$ and $\mathbf{W}_q^v \in \mathbb{R}^{D_B \times D_q}$ are learnable parameters, resulting $\mathbf{B}_k \in \mathbb{R}^{M \times N \times D_B}$, $\mathbf{B}_v \in \mathbb{R}^{M \times N \times D_B}$. σ denotes ReLU unit, and \odot represents Hadamard Product. Afterward, we compute attention for input tokens both spatial and channel-wisely. 1) Spatial-wise attention. We use a linear layer to project \mathbf{B}_k into an intermediate representation $\mathbf{B}_{mid} = \sigma(\mathbf{W}_B^k \mathbf{B}_k)$, where $\mathbf{W}_B^k \in \mathbb{R}^{D_B \times D_{mid}}$. Then another linear layer is applied to mapping $\mathbf{B}_{mid} \in \mathbb{R}^{M \times N \times D_{mid}}$ from the dimension D_{mid} to 1 and further followed a softmax function to obtain the spatial-wise attention weight $\alpha_s \in \mathbb{R}^{M \times N \times 1}$. 2) Channel-wise attention. A squeeze-excitation operation [10] to \mathbf{B}_{mid} is performed to obtain channel-wise attention $\beta_c = \text{sigmoid}(\mathbf{W}_c \mathbf{B}_{mid})$, where $\mathbf{W}_c \in \mathbb{R}^{D_{mid} \times D_B}$ is learnable parameters and $\mathbf{B}_{mid} \in \mathbb{R}^{M \times D_{mid}}$ is the average pooling of \mathbf{B}_{mid} . The first layer of EBA is formulated as,

$$\hat{\mathbf{z}}_L^{e(1)} = \text{EBA}(\hat{\mathbf{z}}_L^e, \hat{\mathbf{z}}_L^v) = \beta_c \odot \alpha_s \mathbf{B}_v. \quad (2)$$

Bilinear Encoder Layer. Besides EBA, we also use the “Add & Norm” layer with the residual connection as in a standard transformer in our bilinear encoder layer. The n -th

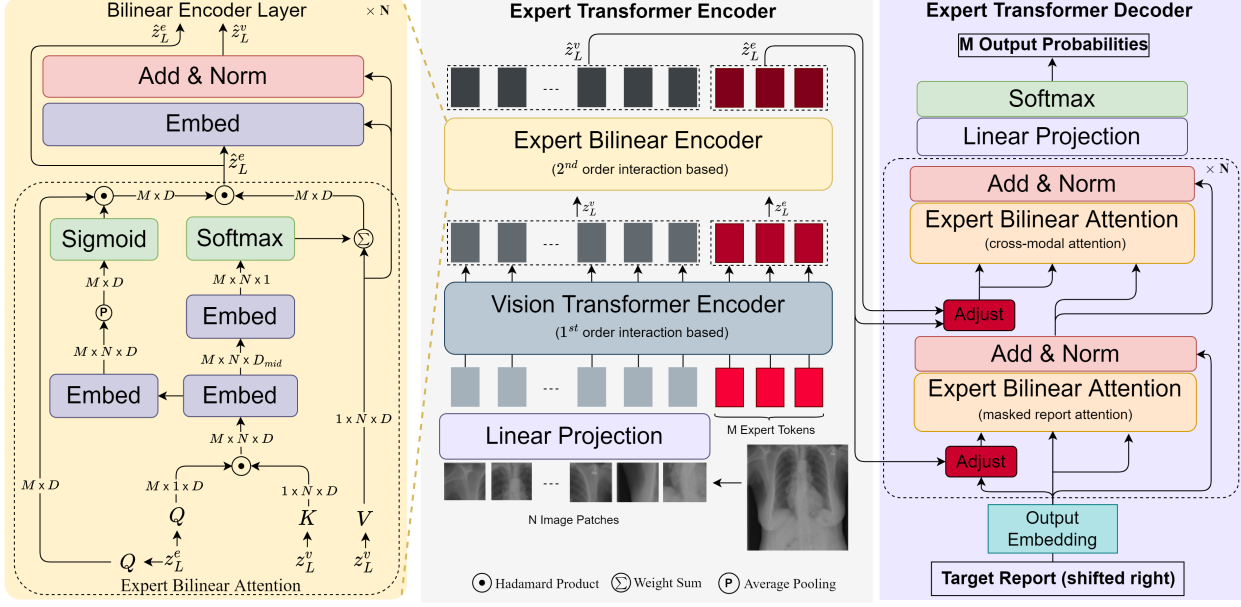


Figure 1. An overview of our proposed METransformer, which includes an Expert Transformer Encoder and an Expert Transformer Decoder. The expert transformer encoder captures fine-grained visual patterns by exploring 1st and 2nd order interactions between input visual tokens and M expert tokens. The expert transformer decoder produces M diverse candidate reports guided by each expert token through an Adjust block. Noted that the “Embed” layer includes a linear projection followed by a ReLU activation function.

layer bilinear encoder is expressed as,

$$\begin{aligned} \hat{\mathbf{z}}_L^{e(n)} &= \text{EBA}(\hat{\mathbf{z}}_L^{e(n-1)}, \hat{\mathbf{z}}_L^{v(n-1)}) \\ \hat{\mathbf{z}}_L^{v(n)} &= \text{LN}(\mathbf{W}_e^n [\hat{\mathbf{z}}_L^{e(n-1)}; \hat{\mathbf{z}}_L^{v(n-1)}] + \hat{\mathbf{z}}_L^{v(n-1)}) \end{aligned} \quad (3)$$

where $n = 1 \dots N$ and N is the number of bilinear encoder layer, $\text{LN}(\cdot)$ denotes layer normalization [2], $\mathbf{W}_e^n \in \mathbb{R}^{(D_q+D_v) \times D_v}$ is learnable parameters and $[\cdot]$ denotes concatenation. We set $\hat{\mathbf{z}}_L^{e(0)} = \mathbf{z}_L^e$ and $\hat{\mathbf{z}}_L^{v(0)} = \mathbf{z}_L^v$ when $n=1$.

3.3. Multi-expert Bilinear Attention Decoder

The output of the last expert bilinear encoder, expert token embeddings $\hat{\mathbf{z}}_L^{e(N)} \in \mathbb{R}^{N \times D_B}$ and visual tokens embeddings $\hat{\mathbf{z}}_L^{v(N)} \in \mathbb{R}^{M \times D_B}$, are sent to the decoder for report generation. The bilinear decoder layer comprises an EBA_{mask} layer to compute attention of the shifted-right reports and an EBA_{cross} to compute the cross-modal attention and an adjust block to regulate the word and visual tokens embeddings by expert token embeddings. For convenience, we denote $\hat{\mathbf{z}}_L^{e(N)}$ and $\hat{\mathbf{z}}_L^{v(N)}$ as \mathbf{f}_e and \mathbf{f}_v .

Adjust block. To incorporate the expert tokens into the report generation process, we propose an expert adjustment block that allows each expert token embedding to influence the embedding of the words and visual tokens, thereby generating the report associated with that expert token. Since our expert tokens are trained orthogonally, we can generate discrepant reports by the different expert tokens. To regulate visual tokens embeddings \mathbf{f}_v by expert token embed-

dings \mathbf{f}_e , the adjust block is calculated as follows,

$$\hat{\mathbf{f}}_v = \text{F}_{\text{adjust}}(\mathbf{f}_e, \mathbf{f}_v) = \sigma(\mathbf{W}_e \mathbf{f}_e) \odot \sigma(\mathbf{W}_v \mathbf{f}_v) \quad (4)$$

where \mathbf{W}_e and \mathbf{W}_v are learnable parameters. σ denotes ReLU unit, and \odot represents Hadamard Product.

Bilinear Decoder Layer. We first perform a masked EBA to word embeddings \mathbf{E}_r and then followed another EBA block to compute the cross attention of word embeddings and visual tokens embeddings. We also employ residual connections around each EBA block similar to the encoder, followed by layer normalization. Mathematically, the i -th layer bilinear decoder can be expressed as,

$$\begin{aligned} \mathbf{E}_{mid}^{(i)} &= \text{LN}(\text{EBA}_{mask}(\hat{\mathbf{E}}_r^{(i-1)}, \mathbf{E}_r^{(i-1)}) + \mathbf{E}_r^{(i-1)}) \\ \mathbf{E}_c^{(i)} &= \text{LN}(\text{EBA}_{cross}(\mathbf{E}_{mid}^{(i)}, \hat{\mathbf{f}}_v) + \mathbf{E}_{mid}^{(i)}) \\ \mathbf{E}_r^{(i)} &= \text{LN}(\mathbf{W}_d^i [\mathbf{E}_r^{(i-1)}; \mathbf{E}_c^{(i)}] + \mathbf{E}_r^{(i-1)}) \end{aligned} \quad (5)$$

where $i = 1 \dots I$ and I represents the total number of decoder layer. $\hat{\mathbf{E}}_r^{(i-1)} = \text{F}_{\text{adjust}}(\mathbf{f}_e, \mathbf{E}_r^{(i-1)})$ and $\hat{\mathbf{f}}_v = \text{F}_{\text{adjust}}(\mathbf{f}_e, \mathbf{f}_v)$. $\mathbf{W}_d^i \in \mathbb{R}^{(D_r+D_r) \times D_r}$ is learnable parameters and $[\cdot]$ denotes concatenation. Specifically, $\mathbf{E}_r^{(0)} \in \mathbb{R}^{T \times D_r}$ is the original word embeddings where T is the total number of words in the reports and $D_r = D_B$ is the dimension of word embedding, and we extend it to M replicates corresponding to M expert tokens to compute parallelly. The final output of the decoder is $\mathbf{E}_c^I \in \mathbb{R}^{M \times T \times D_r}$, which will be further used to predict word probabilities by a linear projection and the softmax activation function.

3.4. Objective Function

Orthogonal Loss. To encourage orthogonality among expert token embeddings, we introduce an orthogonal loss term to the output of expert bilinear encoder $\hat{\mathbf{z}}_L^e$,

$$\mathcal{L}_{OrL}(\hat{\mathbf{z}}_L^e) = \frac{1}{M} \|\ell(\hat{\mathbf{z}}_L^e)^\top \ell(\hat{\mathbf{z}}_L^e) - I\|^2 \quad (6)$$

where $\ell(\cdot)$ denotes L_2 normalization and I represents the identity matrix with dimension M .

Report Generation Loss. We train our model parameters θ by minimizing the negative log-likelihood of $\mathbf{P}(t)$ given the image features:

$$\mathcal{L}_{CE} = -\frac{1}{M} \sum_{m=1}^M \sum_{i=1}^T \log P_\theta(\mathbf{t}_i^{(m)} | \mathbf{I}, \mathbf{t}_{i-1}^{(m)}, \dots, \mathbf{t}_1^{(m)}) \quad (7)$$

where $\mathbf{P}(\mathbf{t}_i^{(m)} | \mathbf{I}, \mathbf{t}_{i-1}^{(m)}, \dots, \mathbf{t}_1^{(m)})$ represents the probability predicted by the m -th expert tokens for the i -th word t_i based on the image \mathbf{I} and the first $(i - 1)$ words.

Our overall objective function is: $\mathcal{L}_{all} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{OrL}$. The hyper-parameter λ simply balances the two loss terms, and its value is given in Section 4.

3.5. Expert Voting strategy

For the M diagnostic reports $R = [r_1, r_2, \dots, r_M]$ produced by M experts, we design a metric-based expert voting strategy to select the optimal one, where the ‘‘metric’’ means the conventional natural language generation (NLG) metrics, such as BLEU-4 and CIDEr (we use CIDEr in our paper). The voting score \mathbf{S}_i for i -th expert’s report can be calculated by the following equation:

$$\mathbf{S}_i = \sum_{j=1, j \neq i}^M \text{CIDEr}(\mathbf{r}_i, \mathbf{r}_j) \quad (8)$$

where $\text{CIDEr}(\cdot, \cdot)$ denotes the function for computing CIDEr with \mathbf{r}_i as candidate and \mathbf{r}_j as reference. In this way, each expert’s report can get a vote score, indicating the degree of consistency of the report with that of other experts. The diagnostic report with the highest score is the winner of the voting. As demonstrated in Table. 3, our voting strategy is more effective than the commonly used ensemble/fusion methods [25]. The possible reasons are that i) our method utilizes NLG metrics as reference scores, thereby the voted results are directly related to the final evaluation metrics; ii) since we ultimately select a single result with the highest score, we can produce a more coherent report than fusing multiple results at the word level.

4. Experiments

4.1. Datasets

In this experiment, two datasets are used for the performance evaluation, i.e., a widely-used benchmark IU-Xray [8] and the currently largest dataset MIMIC-CXR [13] for medical report generation.

IU-Xray Indiana University Chest X-ray Collection (IU-Xray) [8] is the most widely used publicly accessible dataset in medical report generation tasks. It contains 3,955 fully de-identified radiology reports, each of which is associated with frontal and/or lateral chest X-ray images, and 7,470 chest X-ray images in total. Each report is comprised of several sections: Impression, Findings, Indication, etc. In this work, we adopt the same data set partitioning as [5] for a fair comparison, with a train/test/val set by 7:1:2 of the entire dataset. All evaluations are done on the test set.

MIMIC-CXR The recently released MIMIC-CXR [13] is the largest public dataset containing both chest radiographs and free-text reports. In total, it consists of 377110 chest x-ray images and 227835 reports from 64588 patients of the Beth Israel Deaconess Medical Center examined between 2011 and 2016. In our experiment, we adopt MIMIC-CXR’s official split following the work [5] for a fair comparison, resulting in a total of 222758 samples for training, and 1808 and 3269 samples for validation and test.

4.2. Experimental Settings

Evaluation Metrics Following the standard evaluation protocol¹, we utilize the most widely used BLEU-4 [26], METEOR [3], ROUGE-L [19], and CIDEr [31] as the metrics to evaluate the quality of the generated diagnostic reports. To measure the accuracy of descriptions for clinical abnormalities, we follow [4, 5, 22] and further report clinical efficacy metrics. For this purpose, the CheXpert [11] is applied to labeling the generated reports and the results are compared with ground truths in 14 different categories related to thoracic diseases and support devices. We use precision, recall, and F1 to evaluate model performance for clinical efficacy metrics.

Implementation Details For IU-Xray, we use image pairs for report generation as [5]. For both datasets, we use the ‘‘bert-base-uncased’’ model’s tokenizer in huggingface transformer [36] to tokenize all words in the reports. We utilize a pre-trained vision transformer with a patch size of 32 to initialize our Expert ViT encoder. The number of layers of the expert bilinear encoder and decoder is set as (2, 4) for MIMIC-CXR and (2,2) for IU-Xray to reduce the potential overfitting on IU-Xray due to its relatively small data size. The dimensions of the bilinear query-key representation and the transformed bilinear feature (D_B and D_{mid}) in the expert bilinear attention block is set as 768 and 384, respectively. The hyper-parameter λ in Eqn. ?? is set to 2. Our model is trained using Radam optimizer [23] with a mini-batch size of 16. The learning rate is set to be 0.0001 and the model is trained for a total of 20 epochs. We implement our model using Pytorch [27] and Pytorch-lightning library² with two NVIDIA GeForce RTX 3090 GPU cards.

¹<https://github.com/tylin/coco-caption>

²<https://github.com/Lightning-AI/lightning>

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	
IU-Xray	Show-Tell [32]	0.243	0.130	0.108	0.078	0.307	0.157	0.197	
	Att2in [28]	0.248	0.134	0.116	0.091	0.309	0.162	0.215	
	AdaAtt [24]	0.284	0.207	0.150	0.126	0.311	0.165	0.268	
	Transformer [30]	0.372	0.251	0.147	0.136	0.317	0.168	0.310	
	M2transformer [6]	0.402	0.284	0.168	0.143	0.328	0.170	0.332	
	R2Gen [†] [5]	0.470	0.304	0.219	0.165	0.371	0.187	-	
	R2GenCMN [†] [4]	0.475	0.309	0.222	0.170	0.375	0.191	-	
	MSAT [34]	0.481	0.316	0.226	0.171	0.372	0.190	0.394	
	Ours(METransformer)	0.483	0.322	0.228	0.172	0.380	0.192	0.435	
	Results below are not strictly comparable due to different data partition. For reference only.								
	CoAtt [†] [12]	0.455	0.288	0.205	0.154	0.369	-	0.277	
	HGRG-Agent [†] [17]	0.438	0.298	0.208	0.151	0.322	-	0.343	
	KERP [†] [15]	0.482	0.325	0.226	0.162	0.339	-	0.280	
PPKED [†] [21]	0.483	0.315	0.224	0.168	0.376	0.187	0.351		
GSK [†] [41]	0.496	0.327	0.238	0.178	0.381	-	0.382		
MIMIC-CXR	Show-Tell [32]	0.308	0.190	0.125	0.088	0.256	0.122	0.096	
	Att2in [28]	0.314	0.198	0.133	0.095	0.264	0.122	0.106	
	AdaAtt [24]	0.314	0.198	0.132	0.094	0.267	0.128	0.131	
	Transformer [30]	0.316	0.199	0.140	0.092	0.267	0.129	0.134	
	M2Transformer [6]	0.332	0.210	0.142	0.101	0.264	0.134	0.142	
	R2Gen [†] [5]	0.353	0.218	0.145	0.103	0.277	0.142	-	
	R2GenCMN [†] [4]	0.353	0.218	0.148	0.106	0.278	0.142	-	
	PPKED [†] [21]	0.36	0.224	0.149	0.106	0.284	0.149	0.237	
	GSK [†] [41]	0.363	0.228	0.156	0.115	0.284	-	0.203	
	MSAT [†] [34]	0.373	0.235	0.162	0.120	0.282	0.143	0.299	
	Ours(METransformer)	0.386	0.250	0.169	0.124	0.291	0.152	0.362	

Table 1. Comparison on IU-Xray (upper part) and MIMIC-CXR datasets (lower part). † indicates the results are quoted from the published literature: the results of CoAtt [12] and HGRG-Agent [17] on IU-Xray are quoted from [15] while the other results are quoted from their respective papers. For the methods without †, their results are obtained by re-running the publicly released codebase [18] on these two datasets using the same training-test partition as our method.

Methods	Precision	Recall	F1
Show-Tell [32]	0.249	0.203	0.204
Att2in [37]	0.268	0.186	0.181
AdaAtt [24]	0.322	0.239	0.249
Transformer [30]	0.331	0.224	0.228
R2Gen [5]	0.333	0.273	0.276
R2GenCMN [4]	0.334	0.275	0.278
Ours(METransformer)	0.364	0.309	0.311

Table 2. Comparison of clinical efficacy metrics on the test set of the MIMIC-CXR dataset for measuring the accuracy of the description of clinical abnormalities.

4.3. Main Results

Two types of metrics are used in our evaluation: the conventional natural language generation (NLG) metrics and the clinical efficacy (CE) metrics. The results are reported in Table 1 and Table. 2³, respectively.

Specifically, we compare METransformer with 5 state-of-the-art (SOTA) image captioning methods, including Show-tell [32], AdaAtt [24], Att2in [1], Transformer [7]

³It is noted that clinical efficacy metrics only apply to MIMIC-CXR because the labeling schema of CheXpert is designed for MIMIC-CXR.

and M2transformer [6]. For these methods, we use a publicly released codebase [18] and re-run them on both datasets with the same experimental setting as ours. Moreover, eight SOTA medical report generation models are involved in the comparison, including CoAtt [12], HGRN-Agent [17], R2Gen [5], R2GenCMN [4], PPKED [21], GSK [41] and MSAT [34]. It is noteworthy that except R2Gen [5] and R2GenCMN [4], these methods do not have their source code released. CoAtt [12] and HGRN-Agent [17] only report results on IU-Xray in their original paper. For the others, we cite the results from their respective papers. Please note that since IU-Xray does not provide an official training-test partition, the cited performances of these methods on IU-Xray (except R2Gen [5] and R2GenCMN [4] that use the same partition as ours) are not strictly comparable, and they are provided here only for reference. Differently, on MIMIC-CXR, since all these models follow the MIMIC-CXR official training-test partition, their cited performances are comparable.

As shown in Table 1, on both datasets, our METransformer consistently outperforms those “single-expert” based models, including attention-based baselines (Att2in [37], AdaAtt [24]), memory-augmented meth-

#	Models	IU-Xray					MIMIC-CXR				
		BLEU_4	ROUGE	METEOR	CIDEr	AVG. Δ	BLEU_4	ROUGE	METEOR	CIDEr	AVG. Δ
1	BASELINE	0.161	0.357	0.183	0.337	-	0.109	0.277	0.143	0.275	-
2	+BE	0.163	0.360	0.185	0.346	+1.5%	0.111	0.279	0.144	0.287	+1.9%
3	+BE+ETs	0.168	0.372	0.188	0.402	+7.6%	0.117	0.284	0.147	0.334	+8.5%
4	+BE+ETs+OrL	0.170	0.377	0.190	0.426	+10.4%	0.121	0.287	0.150	0.352	+11.9%
5	+BE+ETs+OrL+EV	0.172	0.380	0.192	0.435	+11.8%	0.124	0.291	0.152	0.362	+14.2%

Table 3. Ablation studies. The “BASELINE” uses ViT encoder and bilinear-attention decoder with the traditional cross entropy-loss. Here “BE”, “ETs”, “OrL”, “EV” stand for “Bilinear-attention Encoder”, “Expert Tokens”, “Orthogonal Loss”, and “Expert Voting”. The average improvement over all NLG metrics compared to baseline is also presented in the “AVG. Δ ” column.

Models	Params	Training Time	BLEU_4	ROUGE	METEOR	CIDEr
METransformer(num_expert=1)	152M	0.48h	0.163	0.362	0.183	0.346
Stochastic Model Averaging	152M x 7=1064M	0.48h x 7=3.36h	0.168	0.373	0.187	0.389
Random Initialization	152M x 7=1064M	0.48h x 7=3.36h	0.170	0.376	0.189	0.408
Multiple Decoder	483.3M	1.85h	0.166	0.368	0.186	0.378
METransformer(num_expert=7)	152.007M	0.55h	0.172	0.380	0.192	0.435

Table 4. Comparison with the ensemble models on the IU-Xray dataset.

ods (R2Gen [5], R2GenCMN [4]) and models introducing external information (PPKED [21], MSAT [34]). On MIMIC-CXR, METransformer is the best performer across all metrics. Especially, our CIDEr score is up to 0.362, which is to date the best performance and makes a significant improvement over the second best score of 0.299 of MSAT [34]. These improvements demonstrate the advantage of our framework which is conceptually “multiple expert co-diagnosis”. Ours METransformer also surpasses these methods on IU-Xray over most metrics, while it is again worth mentioning that the cited performances on IU-Xray may not be strictly comparable as obscure training-test partitions were used by different methods.

4.4. Ablation Study

Contribution of each component. We conduct an ablation study to single out the contributions of each model component, as presented in Table 3. We build a baseline by using ViT transformer encoder and bilinear-attention decoder to verify the performance improvements brought by multiple expert tokens, orthogonal loss, and our metric-based expert voting strategy. In Table 3, there are four components: “BE”, “ETs”, “OrL”, and “EV”, representing Bilinear-attention Encoder, Expert Tokens, Orthogonal Loss, and Expert Voting, respectively. The symbol “+” indicates the inclusion of the following component based on the “BASELINE” model. It can be observed that each component of METransformer has a positive effect on performance. By comparing #3 and #1 in Table 3, it can be found that extending the expert tokens and enhancing them with bilinear-attention encoder can increase the overall performance by 7.6% on IU-Xray and 8.5% on MIMIC-CXR. Training with the orthogonal loss on expert tokens can further enhance model performance (by comparing #4 and #3). For experiments #3 and #4 where multiple expert tokens

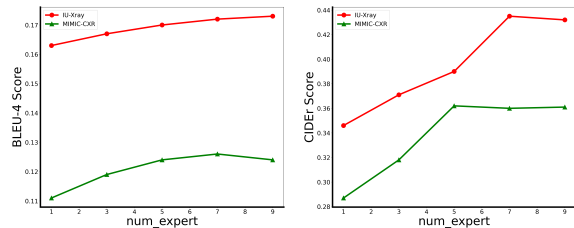


Figure 2. Bleu_4 and CIDEr scores by using different numbers of expert tokens on IU-Xray and MIMIC-CXR dataset.

are used, the final prediction result is obtained by averaging the probability of words predicted by the multiple experts. Comparing #5 and #4 shows that our voting strategy is more effective than the averaging method above.

Impact of expert tokens. To show the impacts of the expert tokens, we train METransformer with different numbers of expert tokens, i.e., $num_expert \in \{1, 3, 5, 7, 9\}$ and the results on IU-Xray and MIMIC-CXR are shown in Figure 2. We observe the following. First, increasing the expert tokens can significantly improve the overall performance of the model. This validates the effectiveness of our motivation that by focusing on different image regions with multiple experts, the model can learn more diversified information and thus produce more accurate and diverse diagnostic reports. Second, when the number of expert tokens exceeds a threshold, increasing num_expert is not able to continue promising a better outcome, for example, when num_expert is increased from 7 to 9, the CIDEr score on the IU-Xray dataset decreases. A possible explanation is that our model forces each expert token to focus on a different image region (via orthogonal loss), which may cause some experts to attend to irrelevant image regions when there are too many experts, thus may negatively affect the generation process.

Comparison with ensemble models. By using multiple experts, our model is conceptually analogous to an ensemble model. We thus compare with three ensemble models

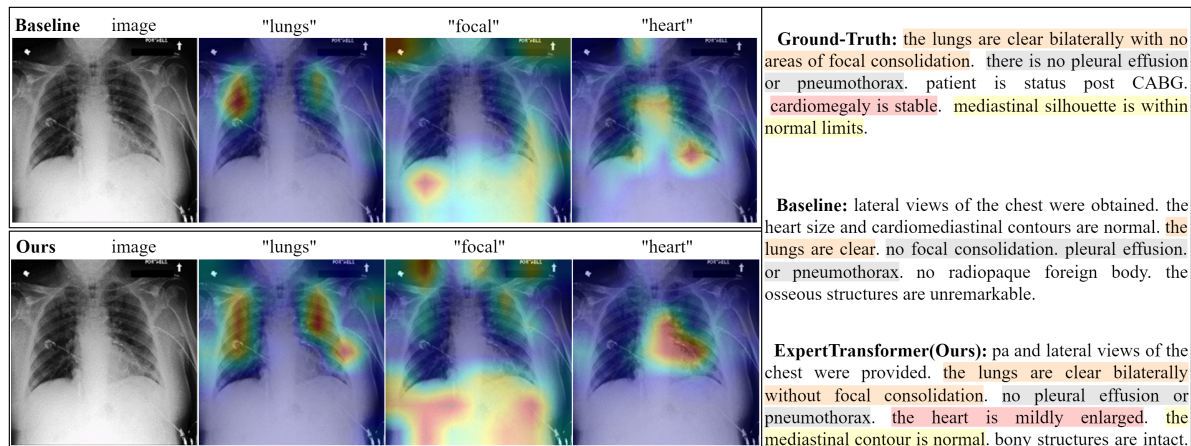


Figure 3. An example of the generated reports and their attention-mapping visualization of three key medical terms from BASELINE and our METransformer. For better illustration, different colors in the generated reports highlight different medical information.

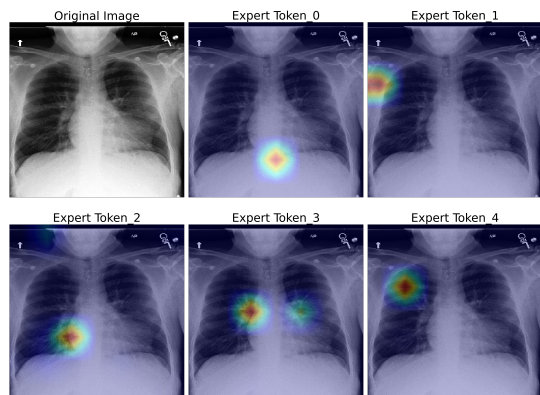


Figure 4. Attention visualization of expert tokens on the image.

in Table. 4. “Random Initialization” trains randomly initialized METransformer(num_expert=1) model for 7 times, corresponding to random-seed based ensemble. “Multiple Decoder” trains the model with the encoder of METransformer(num_expert=1) and 7 randomly initialized METransformer’s decoders, corresponding to late fusion. We also compare with a stochastic model averaging ensemble method. All methods ensemble the results using our proposed Expert Voting strategy. As observed, our METransformer(num_expert=7) performs significantly better with much fewer trainable parameters. We attribute this to the sophisticated interaction between expert tokens through our compact design. Compared with using a single expert token, our method using 7 expert tokens only incurs 0.007M (0.05%) extra parameters, demonstrating its scaling ability.

4.5. Qualitative analysis.

Visualisation of expert tokens. To gain insight, we visualize the image regions mostly attended by each learned expert token in Figure. 4 via exploring the attention $\hat{\alpha}_s$ between the learned expert token embeddings $\hat{\mathbf{z}}_L^e$ and the vi-

sual token embeddings $\hat{\mathbf{z}}_L^v$: $\hat{\alpha}_s = \text{Softmax}(\hat{\mathbf{z}}_L^e (\hat{\mathbf{z}}_L^v)^T)$ Using $\text{Softmax}(\cdot)$, only the mostly attended image regions are shown. As observed, each expert token attends to a distinct and critical image region. For example, the image region attended by expert Token_2 is known as the angle of the rib diaphragm which can provide valuable clinic information.

Qualitative results. We show the generated results of METransformer compared with the Baseline method (“#1” in Table. 3) in Figure. 3. On the left, we visualize the image-text attention mapping from the last layer of the expert transformer decoder with three key generated medical terms. As observed, METransformer better aligns the locations with the related text. On the right, we present the corresponding generated reports and the ground truth. For better illustration, we differently color sentences containing different medical information. It is observed that METransformer is able to generate descriptions better aligned with that written by radiologists. For example, METransformer can diagnose anomalies in the heart part, while the Baseline model fails. This is consistent with the fact that our model can better attend to the heart part of the image (see the attended area of “heart” on the left side of the figure).

5. Conclusions

We present an effective approach for radiology report generation from a new perspective orthogonal to existing research efforts in this field. Our approach follows the concept of multi-specialist consultation to improve the quality of generated reports by introducing multiple learnable expert tokens into a transformer-base framework. Despite its promising performance and properties demonstrated in the experiment, our METransformer is still **limited** in being a basic framework that could be further enhanced by integrating medical domain knowledge, as seen in “single-expert” based methods, which will be explored in our future work.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 2, 6
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3, 4
- [3] Satyanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, 2005. 5
- [4] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. 2022. 1, 2, 5, 6, 7
- [5] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, 2020. 1, 2, 3, 5, 6, 7
- [6] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 1, 2, 6
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019. 6
- [8] Demner Fushman Dina, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Rodriguez Laritza, Antani Sameer, George R Thoma, and Clement J Mcdonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association Jamia*, 2015. 5
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3
- [11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. 5
- [12] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In *ACL*, 2018. 2, 6
- [13] Alistair E. W Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih Ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *CoRR*, 2019. 5
- [14] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 3
- [15] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019. 1, 6
- [16] Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. 2020. 1, 2
- [17] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NeurIPS*, 2018. 1, 6
- [18] Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3799–3802, 2021. 6
- [19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL*, 2004. 5
- [20] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. 2, 3
- [21] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, 2021. 1, 2, 6, 7
- [22] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 2021. 1, 2, 5
- [23] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 5
- [24] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2, 6
- [25] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*. 1, 2, 3, 5
- [26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5
- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 5
- [28] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 2, 6
- [29] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3, 6
- [31] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 5

- [32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2, 6
- [33] Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. Automated radiographic report generation purely on transformer: A multi-criteria supervised approach. *IEEE Transactions on Medical Imaging*, 2022. 1, 2
- [34] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In *MICCAI*, 2022. 1, 2, 6, 7
- [35] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *CVPR*, 2021. 1, 2
- [36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 5
- [37] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2, 6
- [38] Yuan Xue and Xiaolei Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *IPMI 2019*. 2
- [39] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *MICCAI*, 2018. 2
- [40] S. Yang, X. Wu, S. Ge, X. Wu, S. K. Zhou, and L. Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Image and Video Processing*, 2021. 2
- [41] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao. Knowledge matters: Radiology report generation with general and specific knowledge. *Medical Image Analysis*, 2021. 2, 6
- [42] Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. Review networks for caption generation. In *NIPS*, 2016. 2
- [43] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xi-anli Zhang, Yang Li, and Qinghua Zheng. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *ICDM*, 2020. 2
- [44] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *MICCAI*, 2021. 1, 2
- [45] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2
- [46] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *MICCAI*, 2019. 2
- [47] Y. Zhang, X Wang, Z. Xu, Q. Yu, and D. Xu. When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1, 2
- [48] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *CVPR*, 2020. 1, 2