Supplementary Material
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

1. Expert Voting strategy

We provide an illustration of the expert voting strategy shown in Figure. 1.

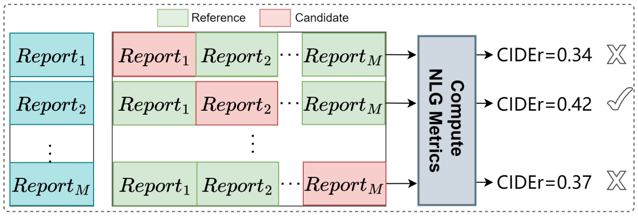


Figure 1. Illustration of expert voting strategy. The $Report_1$ to $Report_M$ represent the reports generated by M expert tokens. Each of the M reports is taken as the candidate report (highlighted in red) alternatively, while the others ($M-1$ reports highlighted in green) are considered as the references for calculating the NLG metrics(highlighted in cyan), and the resulting $M-1$ metrics are averaged as a "voting score" for the candidate report. It is noted that the calculated metric could be any commonly used NLG metrics, such as BLEU_4 [3], ROUGE [2], METEOR [1], CIDEr [4], or a combination of multiple metrics. In this paper, we compute CIDEr as the voting score to select the optimal results.

2. Experiments

Effectiveness of the orthogonal loss. To analyze the impact of orthogonal loss on report generation, we test different weights to the orthogonal loss for training. The experimental results for IU-Xray and MIMIC-CXR are shown in Table. I and Table. II, respectively. We also adopt an overall score to evaluate the performance of the model by considering all metrics by the following literature [5], which is calculated by the Eqn. 1:

$$\mathbf{O} = \frac{1}{4}\left(\frac{B4}{top1(B4)} + \frac{M}{top1(M)} + \frac{R}{top1(R)} + \frac{C}{top1(C)}\right), \quad (1)$$

where O, B4, M, R, and C denote Overall, BLEU_4, METEOR, ROUGE, and CIDEr, respectively, and $top1(\cdot)$ means the highest value of the specific metric in all models. We vary the value of λ to analyze its impact on the model performance. The magnitudes of the two losses are basically balanced when λ is set to 1, and we further increase/decrease λ by 2 and 4 times. As seen, the overall best performance is consistently obtained when λ is increased to 2 on both datasets.

Table I. Hyper-parameter study of λ on IU-Xray dataset.

λ	BLEU_4	ROUGE	METEOR	CIDEr	Overall
0.25	0.166	0.367	0.187	0.439	0.973
0.5	0.170	0.372	0.190	**0.445**	0.989
1	0.167	0.369	0.188	0.442	0.979
2	**0.172**	**0.380**	**0.192**	0.435	**0.994**
4	0.163	0.364	0.180	0.406	0.939

Table II. Hyper-parameter study of λ on MIMIC-CXR dataset.

λ	BLEU_4	ROUGE	METEOR	CIDEr	Overall
0.25	0.121	0.282	0.148	0.357	0.976
0.5	0.124	0.289	0.149	0.361	0.992
1	0.122	0.286	0.147	0.360	0.982
2	**0.124**	**0.291**	**0.152**	**0.362**	**1.0**
4	0.119	0.284	0.145	0.348	0.962

Metrics for expert voting. Our model demonstrates excellent performance even without expert voting, as evident from its top rank in Table 3 of our paper for the CIDEr score metric. Also, Table III verifies voting with other metrics. As seen, CIDEr score can be still improved through voting. CIDEr was chosen as the voting metric because it places less emphasis on common words in reports and thus better captures disease-related information.

Table III. Expert Voting with different metrics, respectively.

Voting Metric	BLEU_4	ROUGE	METEOR	CIDEr
CIDEr	0.124	0.291	0.152	**0.362**
BLEU_4	**0.127**	0.293	0.151	0.355
ROUGE	0.126	**0.296**	0.152	0.357
METEOR	0.122	0.290	**0.155**	0.360

Visualisation of expert tokens. We provide more examples of the visualization of expert tokens in Figure. 2. The attention weights in these cases are obtained by exploring the attention $\hat{\alpha}_s$ between the learned expert token embeddings $\hat{\mathbf{z}}_L^e$ and the visual token embeddings $\hat{\mathbf{z}}_L^v$: $\hat{\alpha}_s = \mathrm{Softmax}(\hat{\mathbf{z}}_L^e(\hat{\mathbf{z}}_L^v)^T)$. Since we use the $\mathrm{Softmax}(\cdot)$ activation function, only the most attended image regions will be shown. As observed, each expert token attends to a distinct and critical image region. For example, the image region attended by expert Token_2 in Figure. 2(a) and expert token_0 in Figure. 2(d) is known as the angle of the rib diaphragm which can provide valuable clinic information. e.g., when the angle of the rib diaphragm is not sharp enough or is even obtuse, it will suggest a small amount of pleural effusion.

References

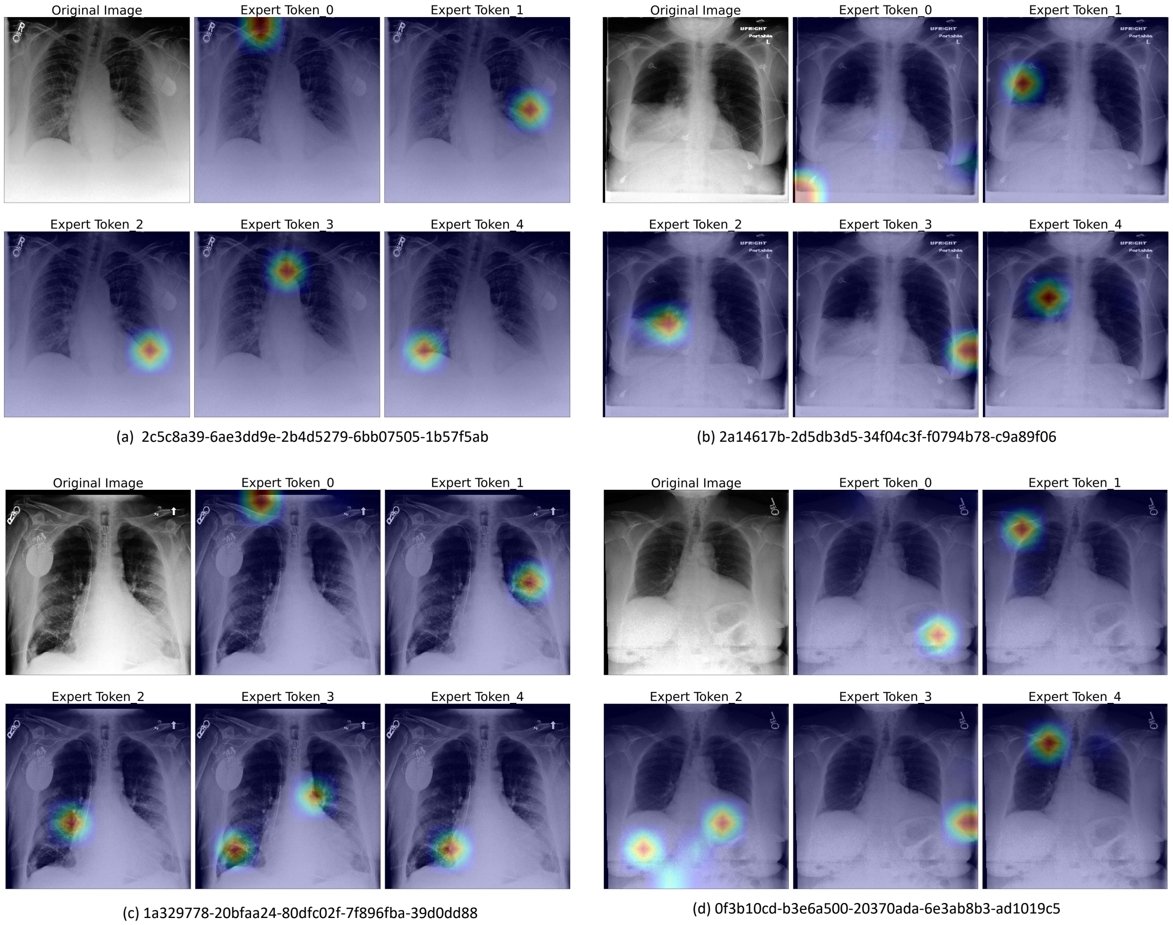
[1] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, 2005. 1

[2] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL*, 2004. 1

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 1

[4] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 1

[5] Wanru Xu, Zhenjiang Miao, Jian Yu, Yi Tian, Lili Wan, and Qiang Ji. Bridging video and text: A two-step polishing transformer for video captioning. 2022. 1



(a) 2c5c8a39-6ae3dd9e-2b4d5279-6bb07505-1b57f5ab

(b) 2a14617b-2d5db3d5-34f04c3f-f0794b78-c9a89f06

(c) 1a329778-20bfaa24-80dfc02f-7f896fba-39d0dd88

(d) 0f3b10cd-b3e6a500-20370ada-6e3ab8b3-ad1019c5

Figure 2. Attention visualization of expert tokens on the image.