Lesion-Aware Transformers for Diabetic Retinopathy Grading

Rui Sun[1], Yihao Li[1], Tianzhu Zhang[1], Zhendong Mao[1], Feng Wu[1], Yongdong Zhang[1]
[1]University of Science and Technology of China
{issunrui, luoheliyihao}@mail.ustc.edu.cn, {tzzhang, zdmao, fengwu, zhyd73}@ustc.edu.cn

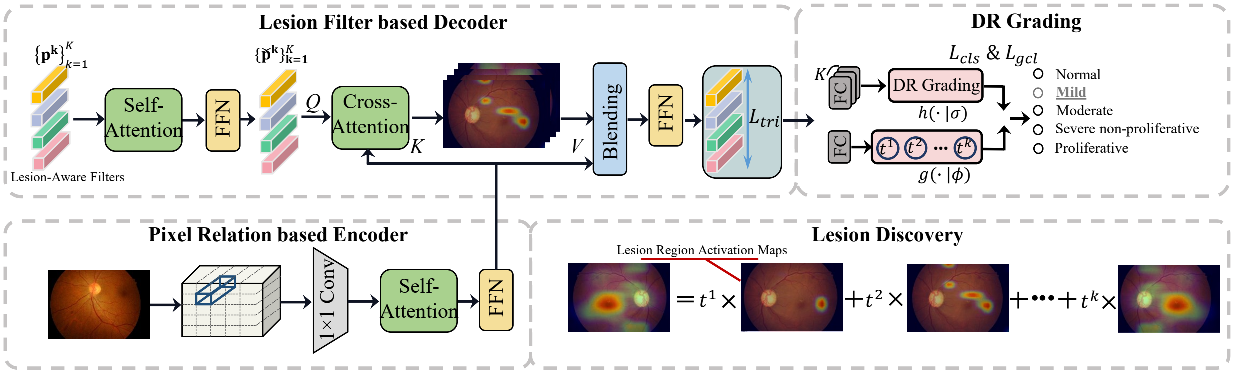

Figure 1. The architecture of our LAT includes an encoder-decoder structure and the classification module. By optimizing the encoder-decoder structure and the classification module jointly, the lesion-aware filters can be learned to identify diverse lesion regions for DR grading and lesion discovery.

In the supplementary material, we first introduce the details about the self-attention, cross-attention and feed-forward network (FFN) in LAT. Then we show more visualization results and analyze them.

1. Model Architecture

The overall architecture of the proposed model is shown in Figure 1, which contains a pixel relation based encoder, a lesion filter based decoder and the classification module. Since our model is based on the transformer architecture [2], in this section, we give more details about the self-attention, cross-attention and the feed-forward network (FFN) in proposed LAT.

1.1. Self-Attention and Cross-Attention Unit

As shown in Figure 2, the difference between self-attention and cross-attention is whether the keys and queries are derived from the same input. The common point is that both types of attention include a multi-head attention mechanism (described in the paper). In addition, there is a residual connection and dropout and layernorm [1] to get the final output.

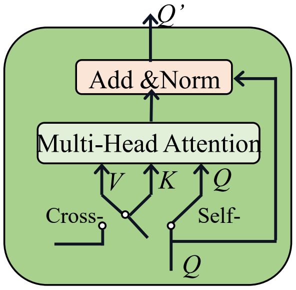

Figure 2. The detailed architecture of the attention unit.

1.2. Feed-Forward Network

The feed forward network (FFN) is a simple neural network containing two fully connected layers with ReLU activations. There is also a residual connection, dropout and layernorm [1] after the two layers. Formally, given $\boldsymbol{X} \in \mathbb{R}^{L\times D}$ as the input, then we can get

$$\hat{\boldsymbol{X}} = \mathrm{ReLU}(\boldsymbol{X}\mathbf{W}_1 + \boldsymbol{b_1})\mathbf{W}_2 + \boldsymbol{b_2}, \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D\times D}$, $\mathbf{W}_2 \in \mathbb{R}^{D\times D}$, $\boldsymbol{b_1} \in \mathbb{R}^{1\times D}$ and $\boldsymbol{b_2} \in \mathbb{R}^{1\times D}$. Then the final output $\tilde{\boldsymbol{X}} \in \mathbb{R}^{L\times D}$ can be calculated

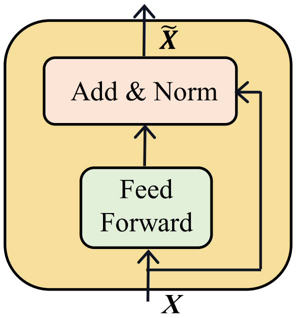

Figure 3. The detailed architecture of the feed-forward network (FFN) in the pipeline of the proposed LAT.

by

$$\tilde{\boldsymbol{X}} = \mathrm{LayerNorm}(\boldsymbol{X} + \mathrm{dropout}(\hat{\boldsymbol{X}})). \tag{2}$$

The detailed architecture of the feed-forward network (FFN) is shown in Figure 3.

1.3. More Visualization Results

With the help of the proposed pixel relation based encoder and the lesion filter based decoder, our model can identify diverse lesions. Although we do not have any lesion information for model training. Even for the most severe level that contain many lesions, LAT can easily identify most of the lesions, which is very meaningful for automatic DR diagnosis based on image-level supervision. LAT can also find other unannotated but important lesions, which often requires specialized equipment to identify such as retinal neovascularizations.

In order to better understand the effectiveness of our method, we show more visualization results of discovered lesion regions in Figure 4. We can observe that in the previous visualization results, the proposed LAT can discover the complete lesion region. However, when the distribution of lesions is very complicated, even when the fundus image is full of lesions, our model may not be able to identify all the lesion regions. As shown in Figure 4, the undiscovered lesion regions are highlighted in red. This is because we do not use any lesion information as supervisory signals for model training. In order to cope with the more extreme distribution of lesions in fundus images, we should further explore the constraint strategy for the lesion-aware filters. This is what we need to solve in the future.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

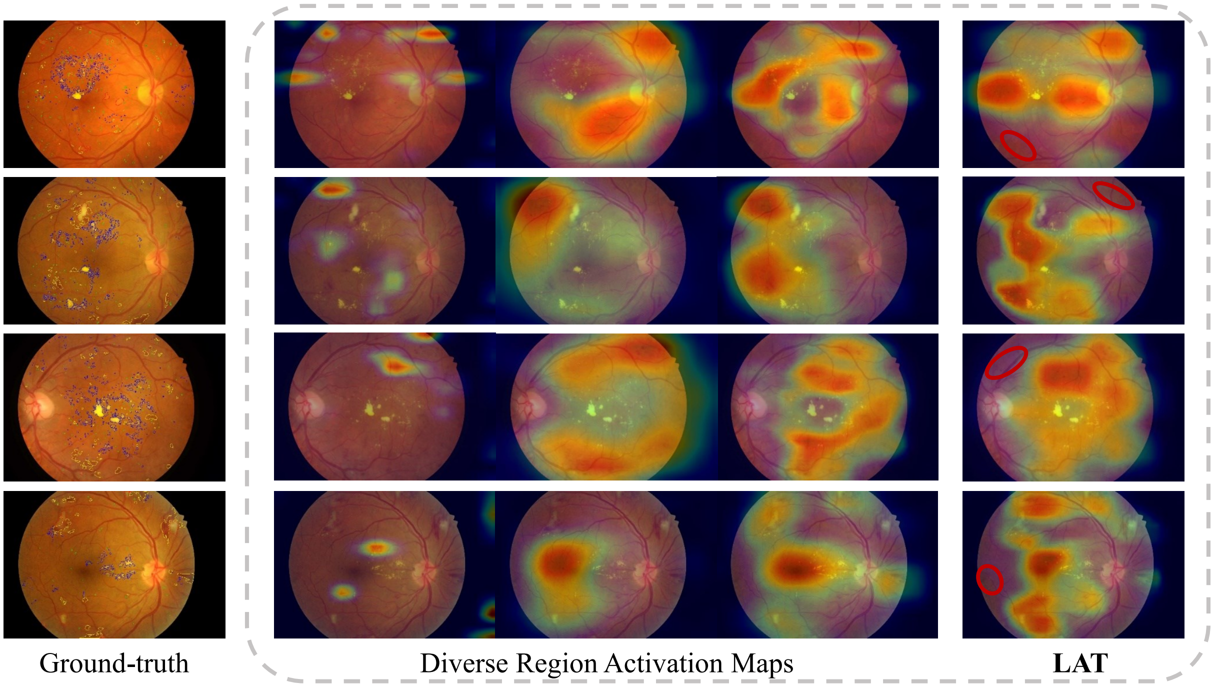

Figure 4. More visualization results. When the distribution of lesions is very complicated, even when the fundus image is full of lesions, our model may not be able to identify all the lesion regions. The undiscovered lesion regions are highlighted in red. And the ground-truth contains microaneurysms, haemorrhages, soft exudates and hard exudates, annotated with green, yellow, green and blue respectively.