multi-head LDSA

February 9, 2021

Background

Recently, Tay *et al.* [1] proposed dense synthesizer attention (DSA) to simplify the expensive dot-product self-attention. In [2], we proposed local dense synthesizer attention (LDSA) which adapts DSA to ASR task by restricting the current frame to interact with its finite neighbouring frames only. In this blog, we describe the formula of multi-head LDSA which is missing in [2]. It should be noted that this is not a formal publication, but rather a supplement to the original work.

The calculation of multi-head LDSA

Assume that there is h heads in each multi-head LDSA block. We first calculate the attention weights for the i-th head by :

$$\mathbf{B}^i = \text{Softmax}(\sigma_{\text{R}}(\mathbf{X}\mathbf{W}_1^i)\mathbf{W}_2^i) \tag{1}$$

where σ_{R} is the ReLU activation function, and $\mathbf{W}_1^i \in \mathbb{R}^{d\times d_k}$ and $\mathbf{W}_2^i \in \mathbb{R}^{d_k\times c}$ are learnable weights of the i-th head. Note that $d_k = d/h$ is the dimension of the feature vector for each head and c is the predefined context width.

Then, we calculate the "value" and the output of the i-th head by:

$$\mathbf{V}^i = \mathbf{X}\mathbf{W}_3^i \tag{2}$$

$$\mathbf{Y}_t^i = \sum_{j=0}^{c-1} \mathbf{B}_{t,j}^i \mathbf{V}_{t+j-\lfloor \frac{c}{2} \rfloor}^i \tag{3}$$

Finally, we concatenate the outputs of all the h heads and calculate the output of the multi-head LDSA block:

$$\text{MH-LDSA}(\mathbf{X}) = \text{Concat}\left(\mathbf{Y}^1, \cdots, \mathbf{Y}^h\right)\mathbf{W}^{\text{O}} \tag{4}$$

where $\mathbf{W}_3^i \in \mathbb{R}^{d\times d_k}$ is learnable projection parameter of the i-th head and $\mathbf{W}^{\text{O}} \in \mathbb{R}^{d\times d}$ is the learnable weight matrix of the final linear projection layer.

References

[1] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng. Synthesizer: Rethinking Self-Attention in Transformer Models. *arXiv preprint arXiv:2005.00743*, 2020.

[2] M. Xu, S. Li, and X.-L. Zhang. Transformer-based End-to-End Speech Recognition with Local Dense Synthesizer Attention. *arXiv preprint arXiv:2010.12155*, 2020.