SUTD 50.039 Theory and Practice of Deep Learning (2024)
HW4
Last update: 2-Apr, 2024

Part I: Coding question

In this question, you will write proxies to predict the ranking of Vision Transformer architectures. In particular, you will compute the proxy scores using (i) grad_norm and (ii) SNIP (details below), based on the initialized weights, and using only one forward and backward pass of one minibatch of samples. No full training is performed on the Vision Transformers to predict the ranking.

We have already prepared 50 Vision Transformer architectures for this homework. For evaluation of the performance of the proxies in ranking, we also provide test accuracy for these architectures. Note that the test accuracy is not used in the proxies. The test accuracy is used only to evaluate the effectiveness of the proxy scores for ranking.

- grad_norm [3]:

This is the Euclidean norm of the gradient vector in the backward pass:

$$\|\nabla_\theta L\| = \sqrt{\sum_{i=1}^{N} \left(\frac{\partial L}{\partial \theta_i}\right)^2}$$

Here, L is the loss function, θ_i is the network weight, and N is the total number of network weights.

- SNIP [3]:

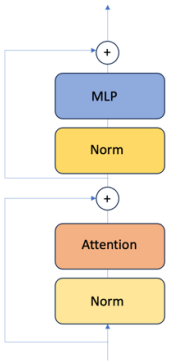
This is the sum of the absolute values of the products between network weights and their gradients at initialization.

$$\sum_{i=1}^{N} \left| \frac{\partial L}{\partial \theta_i} \cdot \theta_i \right|$$

Part II: Concept questions

Question 1:

The following figure shows a Transformer layer as in Vision Transformer ViT [1,2].



The following sequence of two embedding vectors are input to the Multihead Self-Attention block (i.e. after normalization)

[1 0 1 2 1 1], [1 2 0 1 1 1]

The model parameters of the attention block are given as follow:

Number of attention head = 2

Head index i	1	2
W_i^Q	[1 0 0; 0 1 0; 0 0 1; 1 1 0; 0 1 1; 1 1 1]	[0 1 2; 2 1 2; 1 2 1; 1 1 0; 0 0 1; 0 1 0]
W_i^K	[1 1 0; 0 1 1; 1 1 1; 0 1 2; 2 1 0; 3 3 1]	[0 0 0; 0 0 0; 0 0 0; 0 0 1; 1 0 0; 1 3 1]
W_i^V	[3 2 1; 1 2 3; 1 2 1; 2 1 2; 1 1 1; 3 1 3]	[3 2 4; 4 2 3; 0 0 1; 1 1 0; 1 1 1; 3 1 3]
W^O	[1 0 0 1 0 0; 0 1 0 0 2 0; 1 1 1 1 1 2; 2 2 2 1 1 0; 2 1 0 1 2 0; 0 0 0 0 1 0]	

Calculate:

1. The output vectors of the first attention head.
2. The output vectors of the second attention head.
3. The final output embedding vectors of the Multihead Self-Attention block.

Question 2:

As discussed in the class, the complexity for Multihead Self-Attention (MSA) as in Vision Transformer ViT [1,2] is given by:

$$\Omega(4N'D^2 + 2N'^2D)$$

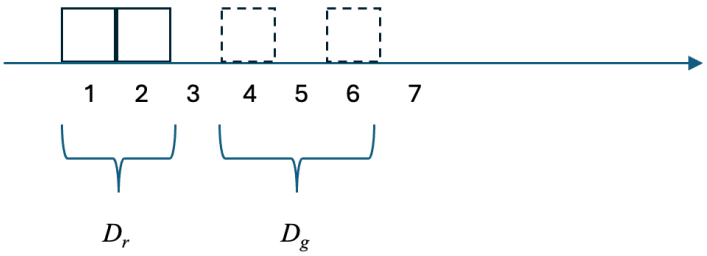
Here, N' is the number of input embedding vectors for MSA, D is the dimension of the input embedding vectors.

Note that complexity for softmax and scaling is not considered.

Explain this complexity expression. State any assumption in queries', keys' and values' dimensionality used in this complexity expression.

Question 3:

Compute the Earth Mover's Distance (EMD) for transporting the distribution of masses D_r to another distribution of masses D_g. Assume that we consider transporting integer amounts of mass, and Euclidean distance is used as the distance metric. Show your calculation, your justification, and your optimal transport plan.



[1] A. Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale" ICLR-2021.
[2] A. Vaswani et al. "Attention is all you need" NeurIPS-2017.
[3] M. S. Abdelfattah et al. "Zero-Cost Proxies for Lightweight NAS" ICLR-2021.