Couplformer: Rethinking Vision Transformer with Coupling Attention Map Supplementary Material

1. Details on Experiment Setup

1.1. Details on CIFAR-10 and CIFAR-100 Training

For CIFAR-10 and CIFAR-100 datasets, models are trained for 200 epochs with AdamW optimizer. The learning rate is 3e-4, weight decay is 3e-2, and batch size is 128. During the training process, the model is started with 5 epochs of warm-up. Data augmentation includes random crop and random horizontal flip.

1.2. Details on ImageNet-1k Training

For ImageNet-1K, we employ the learning rate of 1e-3, weight decay of 5e-2, batch size of 1024, and train for 300 epochs. All the training employ optimizer AdamW employing cosine decay learning rate scheduler with 20 epochs' linear warmup. We leverage Auto-Augment, Rand-Augment, and random erasing as data augmentation. Among employed augmentation, the parameter of random erase is 0.25, mixup is 0.8, and label smoothing is 0.1.

1.3. Details on MS COCO Training

For MS COCO 2017 dataset, we adopt the same multi-scale training setting: resizing the input image on the shorter side with 800 pixels. The optimizer is AdamW, which learning rate is 1e-4, weight decay is 5e-2, and batch size is 16. The schedule is 3x (9 epochs).

1.4. Inference Speed

In the evaluation of ImageNet-1k dataset, the employed hardware is 8 GeForce RTX 3090Ti, and batch size is 1024. The average time consumption for each batch in a single graphic card is 0.19s via Couplformer-T. The inference time/per image is 1.5ms. For Swin Transformer, its average time consumption for each batch in a single graphic card is 0.29s. Thus, the inference time/per image is 2.3ms.

2. Details on Ablation Studies

In terms of testing more model performance, we simplify the training process and reduce the training epochs. Therefore, the training set for the ablation study of head numbers in Section 4.3 is not the same as the evaluation in Section 4.1, which leads to the result in ablation study having a slight gap between the result in CIFAR-10 experiment. According to the ablation study of head numbers, we also conduct similar experiments in other layers as follows.

Layers	Operation	Head number	Accuracy
10	Concate	256	91.23%
		128	91.01%
		64	90.64%
		32	90.32%
		16	75.91%
		8	38.11%
		4	29.62%
		2	26.39%
		1	22.85%
10	pooling	256	90.98%
		128	91.16%
		64	91.45%
		32	91.27%
		16	91.39%
		8	91.60%
		4	91.29%
		2	91.38%
		1	91.42%
6	Concate	256	90.91%
		128	91.13%
		64	91.25%
		32	91.41%
		16	89.98%
		8	75.31%
		4	43.52%
		2	34.73%
		1	49.87%
6	pooling	256	91.17%
		128	91.22%
		64	91.00%
		32	91.49%
		16	91.21%
		8	91.18.%
		4	90.64%
		2	90.95%
		1	90.29%

Table 1: The comparison of accuracy under different modes of position embedding.

3. Visualization of Attention Masks

In terms of describing the effectiveness of Couplformer, we provide the visualization of attention masks. As shown in Figure, we randomly choose multiple images to present their 8 layers' attention masks in the first three heads. We found that the attention score from the standard Visual Transformer distributes at the diagonal of the matrix. However, in the Couplformer, the attention score presents a row-to-row and column-to-column distribution.

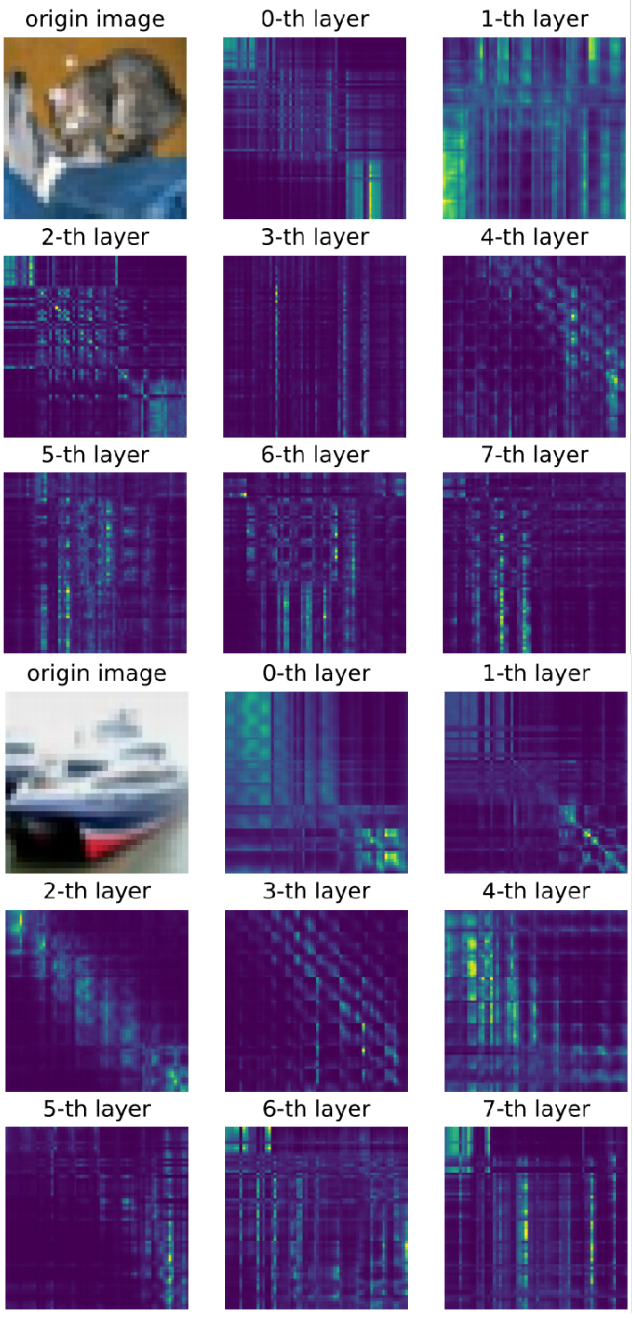

Figure 1: Attention Score from the Standard Transformer

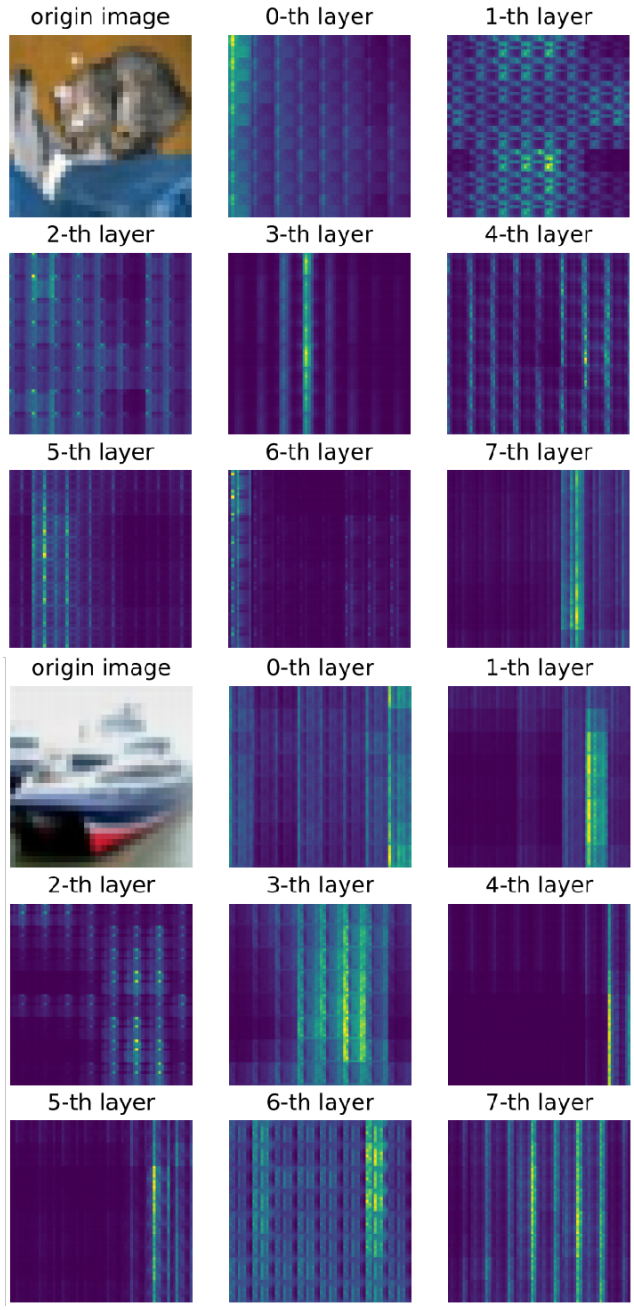

Figure 2: Attention Score from the Couplformer