What does the multi-head do?

Jungyun Byun, Sungkuk Kim
Yonsei University

Abstract

After the introduction of the Vision Transformer, several methods have been developed that utilize multi-head attention for processing images. However, the role of multi-head attention has not been clarified. In this paper, data is analyzed from the perspective of similarity, attention covering range, and image frequency to understand the role of multi-head. In addition, head masking was performed during the test stage to identify head similarity and the importance of each head. An experiment was recently conducted on the test stage using the Swin Transformer and Vision Transformer as models.

I. Introduction

As Transformers(Vaswani et al., 2017) have shown state of art performance in NLP tasks, it has been applied in image tasks, such as Vision Transformer(Alexey Dosovitskiy., 2021). After Vision Transformer had introduced attention to image tasks, Many researches such as Swin Transformer(Ze Liu., 2021), DeepViT(Daquan Zhou., 2021) have been proposed to overcome the limitations of ViT. Although all techniques use multi-head attrition in common, few researches had been conducted on what role multi-head or attention plays in vision. NLP paper 'Multi-head or Single-head? An Imperial Comparison for Transformer Training'(Liyuan Liu., 2021) studies have shown that multi-head and multi-dimensional single-head eventually show the same performance.

So what is the role of multi-head? In addition to the performance index, this research was conducted to define the similarity of heads and understand the effect of head masking of each layers, increasing the number of head masked and tracking the score of model. Part 1 discusses how to measure the similarity between attention heads using SVCCA, attention covering range, and image frequency. In Part 2, the significance of heads and layers, as well as the function of heads, was discussed through the concept of head masking. I used Vision Transformer (ViT), and because there are multiple heads (12 layers with 12 heads), I implemented head masking by either masking all heads in each layer, masking heads in the same index across all layers, or masking random heads.

II. Related work : Attention, SVCCA

1. Attention

Attention is a technique first introduced in the NLP paper "Attention is all you need"(Ashish Vaswani., 2017). The formula for single-head attention (SHA) is as follows.

$$Att_{W_k,W_q,W_v,W_o}(x,q) = W_o \sum_{i=1}^{n} softmax(\frac{q^T W_q^T W_k x_i}{\sqrt{d}}) W_v x_i \quad (1)$$

Vectors of n d-dimensions $x = x_1, \dots, x_n \in \mathbb{R}^d$ is given as an input, and the attention value is obtained through the above formula (1) with the learnable parameters $W_k, W_{q,}, W_v, W_o \in \mathbb{R}^{d \times d}$. This equation is called an equation performed by single head, and the equation performed with a multi-head is as follow formula (2).

$$MHAtt(x,q) = \sum_{h=1}^{N_h} Att_{W_k,W_q,W_v,W_o}(x,q) \quad (2)$$

Here, N_h is the number of heads. As the equation is developed through N_h heads, the dimensions of each weight are also progressed to $W_k, W_{q,}, W_v, \in \mathbb{R}^{d \times d_h}$, $W_o^h \in \mathbb{R}^{d \times d_h}$. In this case, it is usually adjusted to $d_h = \frac{d}{N_h}$, and when it proceeds in this way, there is an effect that looks like an ensemble of SHA.

2. SVCCA : Singular Vector Canonical Correlation Analysis

SVCCA(Maithra Raghu., 2017) is a proposed tool to measure the similarity between two neurons in the Deep Neural Network.

Input : $l_1 = \{z_1^{l_1}, \dots, z_{m_1}^{l_1}\},\ l_2 = \{z_1^{l_2}, \dots, z_{m_2}^{l_2}\}$. As shown in , two neuron sets are received as input.

Step1: Extract subspaces $l_1' \subset l_1,\ l_2' \subset l_2$ representing the most important direction in l_1 and l_2 using the Singular Value Decomposition.

Step2: Calculate the Canonical Correlation similarity for $l_1',\ l_2'$. In this case, the goal is to find the transformed subspace of $\tilde{l}_1 = W \times l_1',\ \tilde{l}_2 = W \times l_2'$ that maximize the correlation value.

Output : Returns the correlation value as a result of SVCCA and the aligned direction $(\tilde{z}_i^{l_1},\ \tilde{z}_i^{l_2})$.

In conclusion, it is a technique of measuring the similarity between two neuron vectors, and it can be seen as a technique of finding a subspace that maximizes the correlation value of two of the subspaces of the vectors using SVD and CCA.

Part 1. Characteristics and similarity of the heads

In order to analyze the use of multi-head in terms of similarity, attention covering range, and frequency, we measured these factors during the testing step. At this time, the Swin Transformer-Base used as a model.

I. How to measure SVCCA similarity

It is ambiguous to define what constitutes similarity. Here, we used SVCCA to quantify the degree of similarity, and instead of stating that heads perform a similar function, we observed similarity in terms of vectors.

In order to apply SVCCA, two vectors need to be selected. First, we select vectors after applying the attention operation. The process of selecting two vectors can be classified as follows.

1.1. Coef idx, and coef value graph

The correlation value is expressed in a graph according to the CCA Coef value and CCA Coef idx. Figure 1-(a) is a graph comparing similarities between blocks of 3rd Layer based on 1st block, and Figure 1-(b) is a graph comparing similarities between blocks of 4th Layer based on 1st block. Figure 1-(c) is a graph comparing similarities between the last blocks of each Layer based on the last block of 1st Layer.

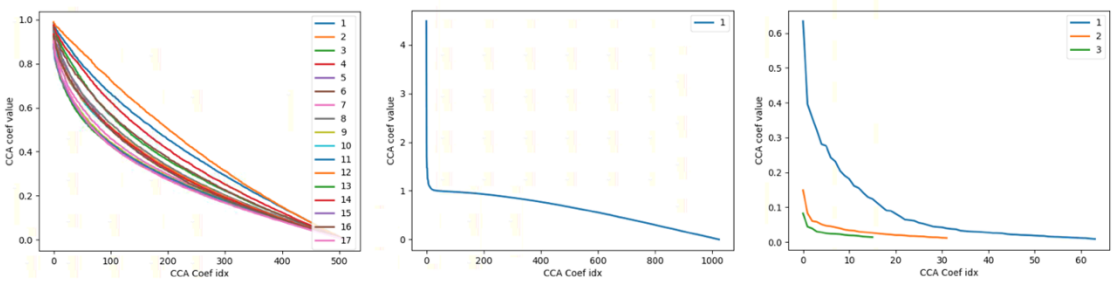


figure 1. SVCCA correlation graph

All three graphs exhibit a convex shape, indicating that similarity decreases as the diversity of the projection direction increases. Therefore, it can be predicted that measuring similarity solely based on the mean value of SVCCA will be challenging.

However, because the similarity between blocks and the similarity between heads may differ, we measure the similarity for each layer's head and each block.

1.2. Each Layer, Last block

First, the similarity between the heads at the last block stage in each layer was compared. Although it was difficult to find the tendency of similarity through distribution charts, it was possible to see how similar they were in each Layer through the maximum and minimum values of the SVCCA values.

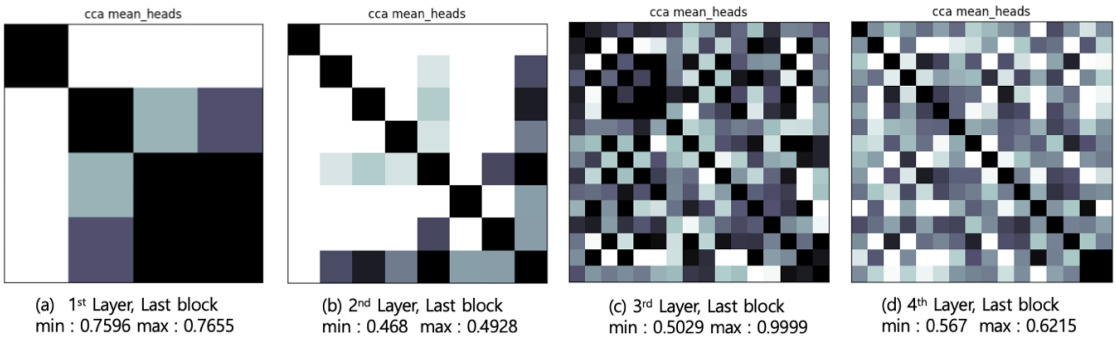


figure 2. head similarities of each Layer, last block

At this time, the number of blocks per layer was [2, 2, 18, 2], and the maximum similarity of the 3rd Layer was 0.99999 even though the number of blocks was 18, and it was found that the SVCCA similarity was very high as the dark color was generally found in the graph.

1.3. 3rd Layer, each block

The similarity was measured high in the 3rd Layer, so the similarity in each block was measured to see if the similarity was high in all blocks.

Block num	0	1	2	3	4	5	6	7	8
Min_correlation	0.5226	0.5094	0.4225	0.4206	0.3845	0.3986	0.3822	0.3641	0.3593
Max_correlation	0.9999	0.5364	0.4615	0.4656	0.425	0.9999	0.9999	0.3967	0.3926
Block num	9	10	11	12	13	14	15	16	17
Min_correlation	0.357	0.3586	0.3679	0.3923	0.404	0.434	0.468	0.4935	0.5029
Max_correlation	0.3834	0.3886	0.401	0.4188	0.4421	0.9999	0.5012	0.5192	0.9999

Table 1. max, min correlation of head similarities on 3rd Layer, each block

The maximum/minimum correlation value between heads in each block is shown as above. While there are blocks with a maximum similarity of about 0.99999, there is a similarity that stays around 0.4.

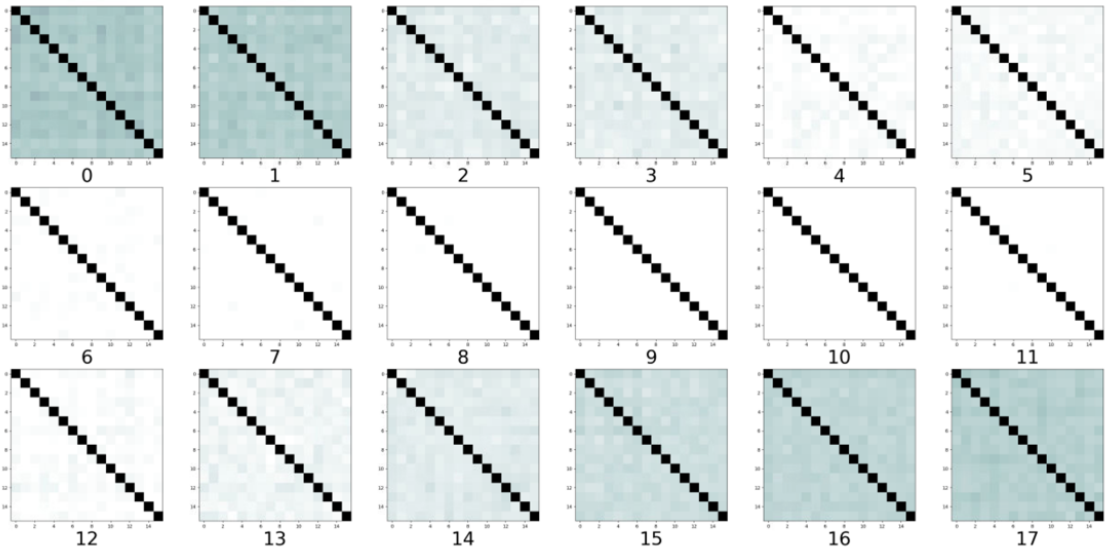


figure 3. applied vmax=1, vmin=0.4 on every map

In the case of Figure 3, the maximum and minimum values are applied equally to all maps. So in the case of dark indexes, we could figure the overall similarity between each other is high. As the block goes backward, it is not gradually darkened or brightened, but it can be seen that it shows a symmetrical appearance.

Among the existing studies, there was an analysis that the similarity between blocks was higher toward the back of the blocks. But it could be seen that the similarity between the heads was different.

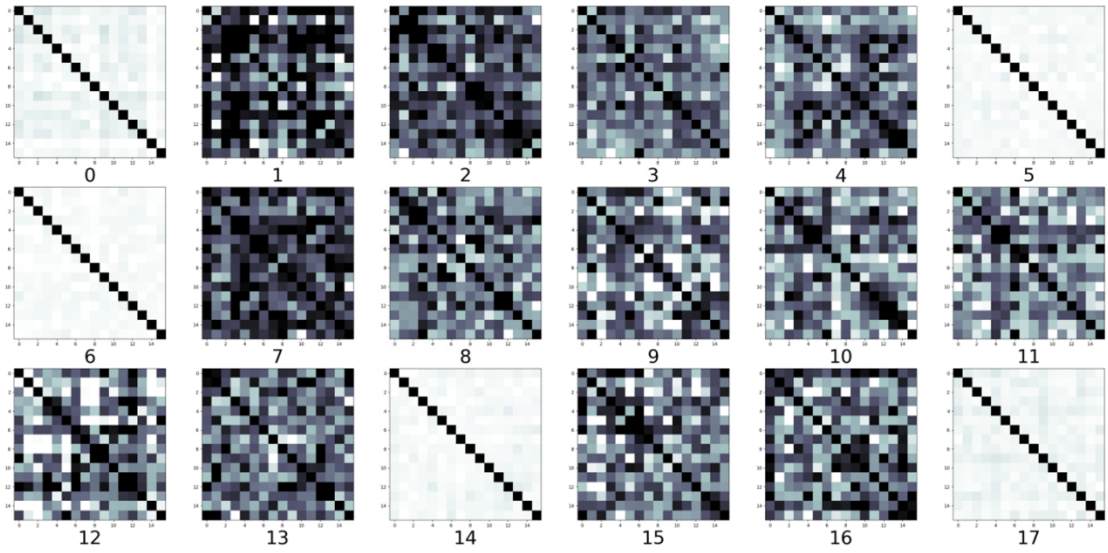


figure 4. For each map, vmax, vmin is applied to each correlation maximum and minimum

Figure 4 shows the application of the max and min values presented in Table 1 to the vmax and vmin values of the plot for each map. In the case of 0, 5, 6, 14, and 17 blocks in which the max value was 0.99999 (as a result excluding similarity to oneself), it was found that only a few parts showed dark colors, and the rest showed similar similarity overall.

2. How to measure attention-covering range

By using the attention output of the model, the attention mask value can be derived to identify where the focus was directed.

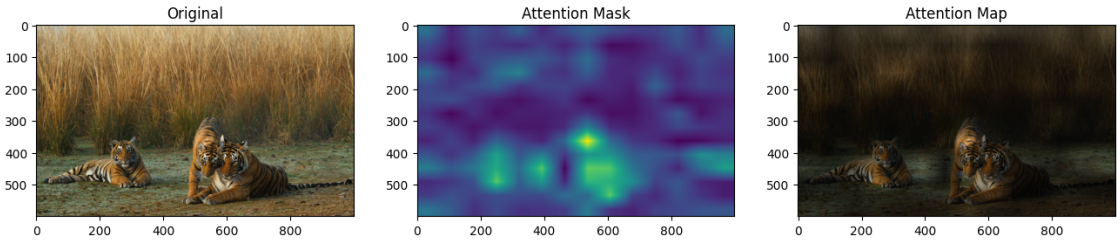


Figure 5 Attention Mask, Attention Map

The attention covering range of each image may be determined by creating a distribution map around the maximum value of the attention mask.

3. How to measure frequency

While looking for several ways to analyze multi-head, we analyzed the correlation between head and frequency after seeing that CNN and Attention may be acting as high pass filters and low pass filters, respectively.

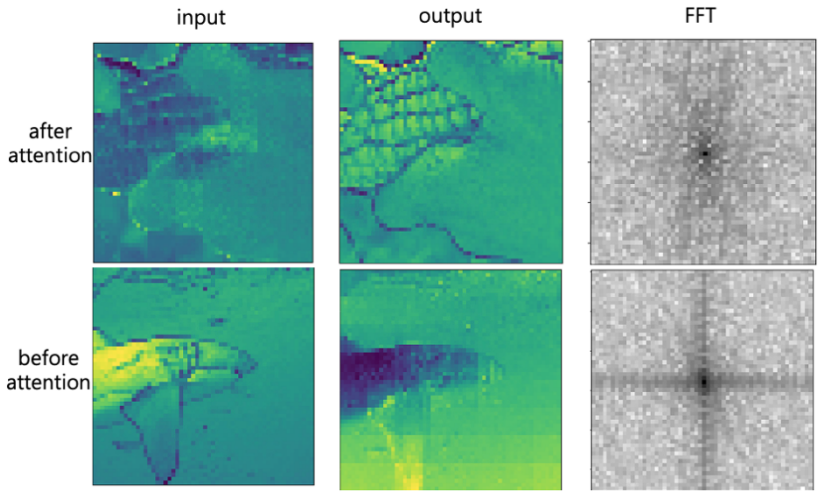


Figure 6. images of inputs, outputs, and Fast Fourier Transform applied to outputs before and after attention is applied.

Although a change in frequency form is seen on figure 5, it is not applied to the same images, and additional quantitative analysis is required. A quantitative analysis will be added later by referring to https://github.com/xxxnell/how-do-vits-work/blob/transformer/fourier_analysis.ipynb (Namuk Park., 2022)

Part 2. Head masking

As it is a study to determine the influence of the number of heads, in this part, I conducted on head masking. For this part, I used pretrained ViT model and finetuned on cifar10 dataset. Head masking is conducted during testing step.

1. What effects the score : Head or Layer

Because the number of heads was too large in the model used, all cases could not be considered in head dropping. As shown in Figure 6, it was divided into three main methods.

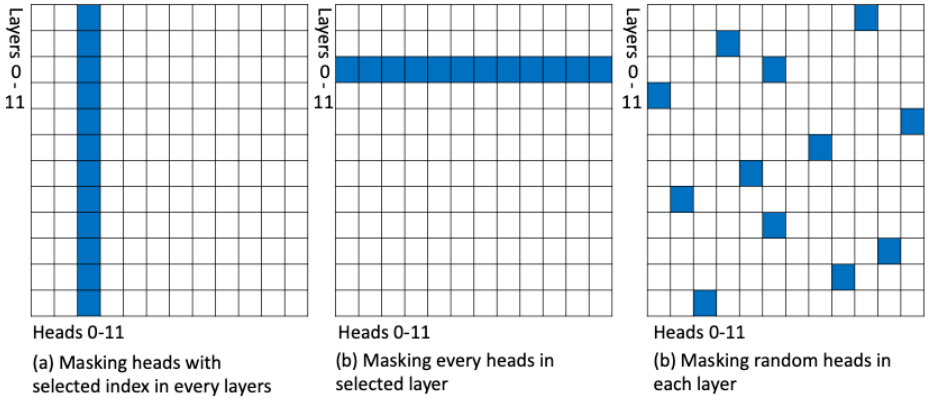
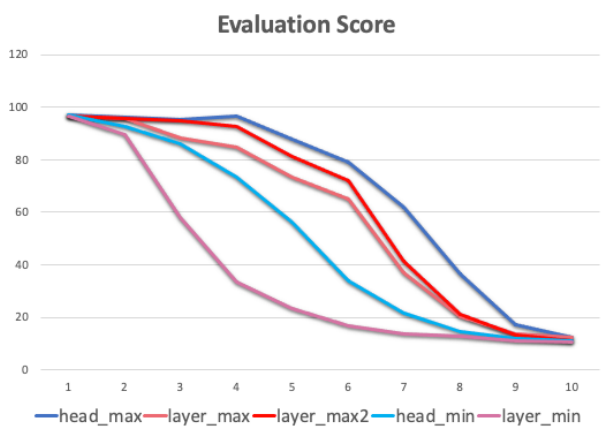


Figure 7. 3 ways of masking 12 heads

The case of Figure 7-(a) is when the head of the selected index is masked in all layers. Figure 1-(b) is when all heads of the selected layer are masked, and Figure 7-(c) is when a specific number of heads are randomly pulled out and masked in each layer. The criterion for selecting a head or layer is based on the value that produces the greatest or worst result when each remaining index is masked one by one in the existing masked state. For example, in case (a), if the head of [2,4,6,7] is masked, the remaining indexes [0,1,3,5,8,9,10,11] are masked one by one to observe the evaluation result, and the index with the highest result is selected. Then we repeated this step.



(a) graph

step	head_max	layer_max	layer_max2	head_min	layer_min
1	97.025	96.46	96.4	96.9	96.46
2	96.22	95.74	95.68	92.63	89.5
3	95.09	88.09	94.95	86.14	58.02
4	96.7	84.88	92.65	73.46	33.43
5	87.8	73.52	81.44	56.11	23.35
6	79.25	65.17	71.88	34.01	16.7
7	61.73	37.11	41.14	21.78	13.85
8	36.7	20.26	21.12	14.66	12.78
9	17.075	13.7	13.06	11.78	11.17
10	12.23	12.33	11.25	10.9	10.6

(b) chart

Figure 8 Evaluation Score after head masking

The head_min value follows the same format as in Figure 7-(a), and the index is chosen to maximize the evaluation value of the remaining heads. The head_min value follows the same format as in Figure 7-(a), but the index is chosen to minimize the evaluation value of the remaining heads. Layer_max and layer_min

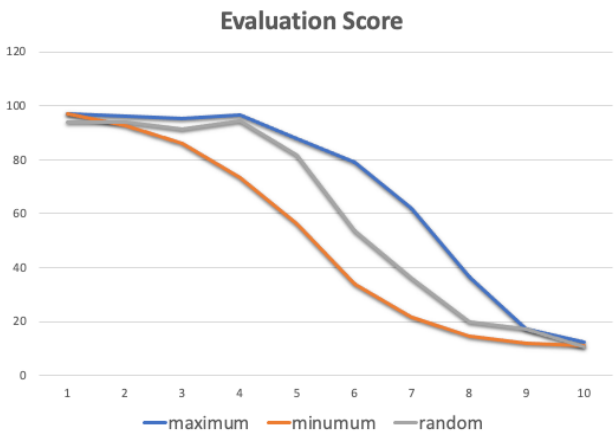
are also derived using the same method, but the head masking followed the format shown in Figure 7-(b). In the case of layer_max2, the evaluation score of the last layer was not included. It was deemed appropriate to use the value of layer_max2 when comparing with head_max because the value of the last layer turned out to be exceptionally small.

The masking process yielded unexpected results, particularly with head_max, as there was no significant difference in performance even in Step 4 (when 48 heads were masked). The performance showed a decline from Step 5. Comparing the similarity between the heads masked up to Step 4 and the difference between the heads masked in Step 5 appears to be helpful in defining the similarity between the heads.

Overall, it is evident that when the head was masked as shown in Figure 7-(a), the performance was higher compared to when the head was masked as shown in Figure 7-(b). However, even after removing all the heads of the layers, layer_max2 did not show a significant performance change until step 3. This suggests that the overall structural inefficiency of the ViT model could be identified.

2. Is there important head?

To observe the impact of the masked head on performance, the results of masking were compared in three different ways.



(a) graph

step	maximum	minumum	random
1	97.02	96.9	94
2	96.22	92.63	94
3	95.09	86.14	91.1
4	96.7	73.46	94.3
5	87.8	56.11	81.8
6	79.25	34.01	53.4
7	61.73	21.78	36.2
8	36.7	14.66	20
9	17.07	11.78	17.4
10	12.23	10.9	10.4

(b) chart

As a result, the minimum masking experimental value showed a significant drop compared to other values from step 4 (i.e., when the head was masked 48 times). In the case of random masking, it was observed that the value decreased compared to the maximum masking from step 6.

Therefore, it can be seen that the evaluation value varies greatly depending on which head is selected, and it will be possible to select or classify important heads from those that are not. Then we could find the characteristics by comparing important heads or similar heads by this result.

In this experiment, maximum masking was determined by simply selecting the head index that lowers the evaluation score value to the minimum. However, if the maximum value is selected with another equation in the future, a better value can be derived.

3. Others

Experiments were also conducted to mask the heads one by one or to examine the correlation between each head and class in CIFAR-10. However, when the heads were masked individually, the resulting values fell within the error range, and there was no clear data indicating the correlation between the head and class, so it was not included.

Discussion and Conclusions

In order to analyze the role of multi-head, research has been conducted from the perspective of SVCCA, attention covering range, and frequency. Through head masking, it was concluded that even with 30% of head masked, there was no direct effect on the score and that the importance of each head would vary.

Through head masking, we were able to select significant features and identify heads that are assumed to have similarities. During the experiment, the GPU RAM capacity was limited, so it was unable to apply the measures of Part 1 to this dataset. However, if given the opportunity, I would like to continue this study.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, arXiv:2103.14030

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, Jiashi Feng, DeepViT: Towards Deeper Vision Transformer, arXiv:2103.11886

Liyuan Liu, Jialu Liu, Jiawei Han, Multi-head or Single-head? An Empirical Comparison for Transformer Training, arXiv:2106.09650

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, arXiv:1706.03762

Maithra Raghu, Justin Gilmer, Jason Yosinski, Jascha Sohl-Dickstein, SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability, NIPS 2017

Namuk Park, Songkuk Kim, How Do Vision Transformers Work?, ICLR 2022