

Flower Classification and Key Parameter Analysis Based on Vit

Ruo Chen Deng^a

School of Computer Science and Engineering, Sun Yat-sen University, Guangdong, China

Keywords: Flower Classification, Vision Transformer, Multi-Head Attention Mechanism, Self-Attention Mechanisms.


Abstract: Flower classification holds significant implications for various fields, including plant resource survey and plant taxonomy education. This paper proposes employing the Vision Transformer (ViT) model for flower classification tasks. The study aims to investigate the impact of varying depth and head parameters in ViT model on their performance. Through an analysis of accuracy performance and attention properties, the research explores optimal strategies for setting depth and head parameters. Additionally, it delves into the phenomenon of attention collapse within the multi-head attention mechanism, utilizing mean attention distance plots for in-depth analysis. Results reveal a positive correlation between model depth, number of heads, and classification accuracy. Moreover, insights gleaned from attention collapse observations provide valuable guidance for optimizing depth and head parameter settings. This study offers valuable insights into the performance of ViT models in flower classification tasks, while also contributing to the understanding of depth and head parameters in self-attention mechanisms for future research endeavors.

1 INTRODUCTION

On the plant evolutionary stage, the number of flowering plants grows considerably to about 250,000 living species, classified into nearly 350 families, which makes them one of the most flourishing creatures on the planet (Kenrick, 1999). Flower classification is a groundwork in Botany. The most primitive way to classify flowers is fully by man-made observation with botanical expertise. In the modern time, the application of flower classification is reflected in many aspects, including management of a query index system based on image content for flower databases (Das, 1999), plant resource survey, and education on plant taxonomy (Chi, 2003). In these scenarios, manual classification can be not practical enough. With the rapid development of computer vision science, flower classification by computer becomes an obvious tendency.

Classifying flower plants always appears to be tougher than other image classification job. Even for a real person, it can be hard to tell the differences between two species of flowers, compared with a 'car, bike, human' task (Nilsback, 2006). In the field of computer vision, the flower classification presents an additional hurdle due to the significant similarities

between classes, and furthermore, flowers are non-rigid entities capable of various deformations, leading to considerable intra-class variation (Nilsback, 2008). Traditional flower classification techniques typically use extracted features such as color, texture, and shape from images to enhance classification performance. While Support Vector Machines (SVM) can classify flowers based on these features, the robustness of this traditional method is not guaranteed. This is primarily because the conventional ways highly rely on specific manually crafted features, which may not generalize well under varying conditions like changes in lighting, flower poses, or surrounding objects (Nilsback, 2006)(Nilsback, 2008)(Hiary, 2018). With the introduction of deep learning technology in the direction of image recognition, especially the use of convolutional neural networks (CNN), automatic learning of invariant features of flower images demonstrates superior accuracy compared to traditional hand-made methods. Besides, after the transformer architecture was proposed and achieved significant success in the realm of natural language processing (NLP), the seq2seq architecture has also been applied in the field of computer vision. The ResNet (He, 2016) and EfficientNet (Tan, 2019) which are based on CNN are often considered to

^a <https://orcid.org/0009-0004-2986-3886>

dominate the field, but now the Vision Transformer (ViT) shows its potential replacement (Dosovitskiy, 2020).

The main objective of this research is to assess the effectiveness of flower classification utilizing transformer architecture, particularly the ViT model renowned for its unique self-attention mechanism (Vaswani, 2017). This mechanism enables the establishment of global relationships among image patches, facilitating the learning of intricate feature correlations within flower image datasets. The research focuses on examining the impact of varying head quantities and depth of encoder layers on the prediction accuracy curves of the ViT model with pre-training. Furthermore, analyzing the attention map distribution across different heads and transformer layers is vital for understanding the model capability to establish relationships between image patches and extract meaningful features from complex flower images. The analysis also highlights that while increasing model depth can lead to performance improvements, there's a point of saturation. Through simulations of the transformer's receptive field to measure attention distribution, this study provides insights into optimal trade-offs. Ultimately, it suggests that while augmenting the number of heads and depth in ViT models generally enhances performance, the highest values may not always be optimal, especially in intricate tasks like flower classification. Careful consideration of these trade-offs is essential for achieving optimal results in flower classification tasks.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

The dataset used in this work is called `tf_flowers`, sourced from TensorFlow Datasets (TFDS), containing 3670 images of flowers (Luo, 2022). All original images are sourced from Flickr. Each image varies in size, the number of flowers, shapes, proportions within the frame, etc. This flowers dataset contains five categories: daisy, dandelion, roses, sunflowers, and tulips. A sample is shown in the Figure 1.

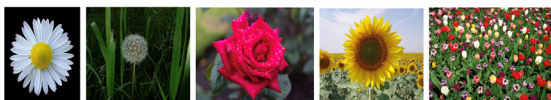


Figure 1: Images from `tf_flowers` dataset (Photo/Picture credit: Original).

With no predefined splits, in this work, 20% of images are randomly sampled for validation, the rest for training. About data preprocessing, the main task is to resize the images to a consistent size, 224x224 pixels. Specifically, for the training set, to enhance data diversity and complexity, images are randomly cropped to the specified size, with a 0.5 probability of horizontal flipping. For the validation set, to maintain consistency and comparability in evaluation, the image's shorter is resized to 256 pixels and cropped into a 224x224 pixel region from the centre. Finally, all image data is converted into tensor and normalized.

2.2 Proposed Approach

This study primarily focuses on implementing the classic ViT model for flower image classification tasks, with a specific emphasis on two key hyperparameters: depth and head. The architecture of the ViT model comprises three main components: the Embedding layer, the transformer encoder, and the MLP head. The investigation employs various parameter analysis methods, including accuracy curves and visualization of attention maps (both self and class token), along with mean attention distance dot diagrams. These methodologies are employed to examine how variations in depth and head influence the model's performance, thus offering valuable insights into the effectiveness of the ViT model for flower classification tasks. The pipeline is illustrated in Figure 2, providing a visual representation of the process.

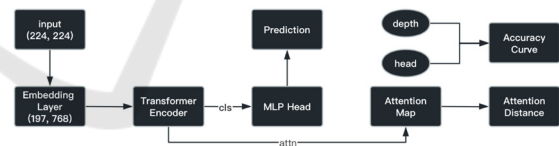


Figure 2: The pipeline of the model and analysis method (Photo/Picture credit: Original).

2.2.1 Embedding Layer

The model converts the image, represented as a three-dimensional matrix $[H, W, C]$, into patches using a simple convolutional process. With a kernel size of 16×16 , a stride of 16, and 768 filters, an input image shape of $[224, 224, 3]$ transforms into $[14, 14, 768]$. After this process, the output can be a set of tokens with a shape of $[196, 768]$. Furthermore, before these tokens proceed to the next parts, a position embedding process is applied to keep the sequential information among the patches. Besides, a `[class]`

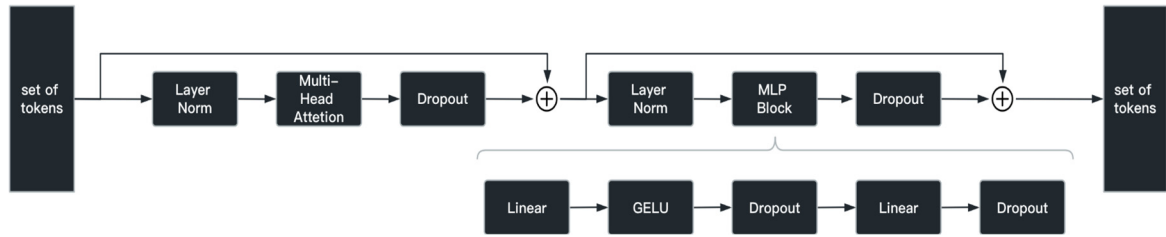


Figure 3: Encoder Block (Photo/Picture credit: Original).

token is added as well to store classification information, resulting in [197, 768] as output shape.

2.2.2 Transformer Encoder

This part is the core of the model, and the hyperparameters, depth and head mainly discussed in this work, also come from this part. Transformer Encoder consists of stacking Encoder blocks, the precise number of Encoder block represents the value corresponding to the depth parameter.

Each Encoder block contains Layer Normalization, normalizing tokens, not the batches, and Dropout process and MLP block are also used here, which is shown in the Figure 3.

The most critical part is multi-head attention. The head parameter comes from the number of head here. Compared with conventional self-attention, multi-head self-attention performs the same process on multiple "heads" in parallel. Each head has independent, learnable matrices for generating key, query, and value vectors. In this work, the division of multi-head is by equally dividing the dimensions of the Query vector, Key vector, and value vector to each head. Then different heads use their own token information to complete the self-attention mechanism. Eventually, directly adding a fully connected layer can aggregate the output of different heads. The formula is as follows (Dosovitskiy, 2020):

$$Attention(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

$$MultiHead(Q, K, V) = \text{Concate}(head_1, head_2, \dots, head_n)W^o \quad (2)$$

Where:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

2.2.3 Mlp Head

After the Transformer Encoder, the output shape remains unchanged compared to the input. In this work, only classification information is needed. By completing the full connection layer of the class token

input in MLP, the probability distribution of each corresponding classification item can be formed, based on which the prediction can be made.

2.2.4 Attention Map Visualization and Attention Distance

To examine the impact of the hyperparameters depth and head on the self-attention mechanism, it is necessary to extract specific attention information from the model. In this work, the ‘Visualizer’ (Luo, 2022) tool helped extract the Attention Map nested deep in the model. When the model predicts the test image, it will store ‘depth’ pieces of tensor information with shape of [index, head_num, 197, 197] in the cache, in which depth is the layer number of the transformer encoder, index is the picture number, head_num is the head number. the values in row i and column j in the 197*197 two-dimensional matrix represent the attention value of the i-th patch to the j-th patch, and among 197 tokens, the one at position 0 is [class] token. Based on this, deeper attention information is accessible.

Attention distance (Dosovitskiy, 2020) is proposed to explain the ‘receptive field’ of the ViT model and is used to characterize the model capability to perceive data. As to the ViT model, if the attention mechanism tends to integrate global information, it means that its ability to perceive data is stronger and the attention distance is larger. In this work, it is concerned that the average Attention distance of all patches corresponding to a certain head in a certain layer, which is the mean attention distance.

2.2.5 Loss Function

In this task, the cross-entropy loss function is utilized as the Loss function, which has been proven to be extremely effective in image classification tasks many times.

Specifically in this task, flower classification is essentially a single-label classification task. Each sample has only one label. For such a single sample, assume that the real distribution is y, the model output

distribution is \hat{y} , and the number of categories is n . Then there is the following counting formula:

$$Loss = -\sum_{i=1}^n \log \hat{y}_i \quad (4)$$

The lower the loss value, the closer probability distribution output by the model is to the real one.

After the calculation of Loss is completed, the back propagation algorithm will be utilized to calculate the parameters' gradients. Then the pre-designed optimizer, with the learning rate through the cosine annealing algorithm, will promote the process of stochastic gradient descent, updating the parameters. At last, the previous gradient information needs to be cleared in time.

2.3 Implementation Details

When specifically training the ViT model, it is not recommended to only use this flower data set for training. In this case, this model does not work as well as it can. No matter how to adjust the epoch, learning rate algorithm, or other hyperparameters, the accuracy is always tough to rise to a satisfactory level. The ViT model is similar to other transformer models. Only by pre-training on very large-scale data sets, accumulating generalization capabilities, can the model reflect its own effects and advantages. This work mainly uses the weight file formed by the model pre-training on ImageNet-21k as a base for further model training on flower classification tasks.

3 RESULTS AND DISCUSSION

By utilizing the ViT model on flower classification, this study firstly analyzes the performance of changing depth and head hyperparameter on model training, and then discusses the logic behind the design of the multi-head mechanism. Based on these analyses, this work explores the shallow and deep trade-offs on hyperparameter settings.

As shown in the Figure 4, the accuracy curve of the model without pre-training weights was examined when depth is set to 1, 2, 4, 8, 12, 24, and 32. In this part, the accuracy is generally positively correlated with the depth. Deepening the number of layers to achieve better and more stable training is the basic idea in the deep learning field. More encoder blocks extract data information more deeply. The improvement brought by the depth from 1 to 12 is considerable. However, it was found that the improvement brought by the depth from 12 to 32 is not significant enough, and the cost-effectiveness of

deepening was greatly reduced. A trade-off based on this is necessary. In this work, images of flower often have sufficient space for understanding, and when the depth is 12, the ViT model can work well enough.

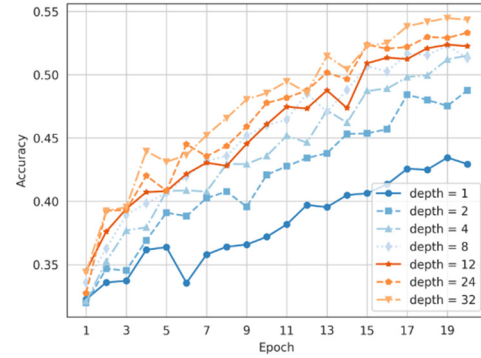


Figure 4: Accuracy curve of varying depth (without pre-training) (Photo/Picture credit: Original).

As shown in the Figure 5, with pre-training weights and fixing the depth to 12, the accuracy curves of the model are examined when head is set to 1, 2, 4, 8, and 12. With the increase in head quantity, the model's accuracy under the same epoch improves, and the enhancement is more stable. In this work, the division of heads is by equally dividing the dimensions of the token into each head. It can be considered that different heads understand the information from different aspect. Q and K in the attention mechanism are essentially used to measure the association between different patches, and this association may be very abstract and complex. To better understand it, different heads will describe these associations in different vector spaces.

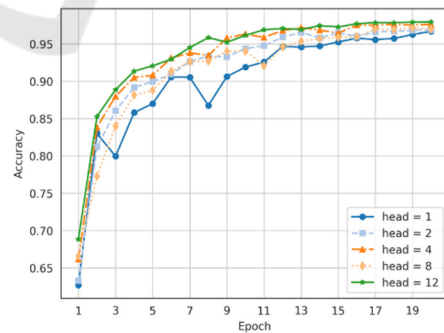


Figure 5: Accuracy curve of varying head (with pre-training) (Photo/Picture credit: Original).

Figure 6 shows the attention map in the first layer in the single-head and multi-head case. When the number of heads is small, the model may excessively focus on its own position, appearing an obvious diagonal on the attention map. But for multiple heads,

the attention maps formed by different heads in the same layer vary a lot. Some are also obvious diagonals, and some are not obvious at all. The multi-head design essentially offers the model some chances to avoid the trap of excessive self-focus, so it can improve the model to be better and more stable.

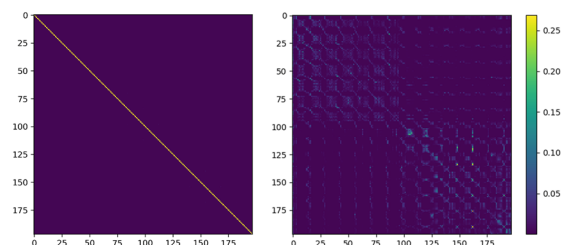


Figure 6: Attention map of one head and multi-head (Photo/Picture credit: Original).

As shown in Figure 7, the test image is shown on the left. The 12 Figures in the upper two rows on the right are the class token attention of the 12 heads in the first layer. Some focus on the local part of the flower, some focus on the overall flower, and some focus on global parts of the background. From the perspective of this diversity effect of attention, the significance of multi-head is profound. The 12 figures in the next two rows are from the 12th layer, the last layer. It is found that different heads show strong similarity in the distribution of attention, which shows that as the model deepens, the self-attention mechanism of each head becomes less effective in generating different attention to capture the features and connections between patches. In other words, attention collapse (Dosovitskiy, 2020) occurs.

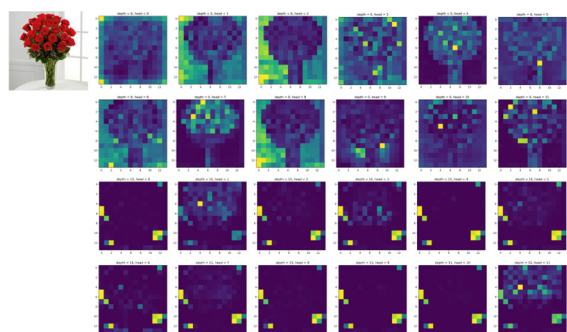


Figure 7: Attention visualization of each head on layer 0 and 11 (Photo/Picture credit: Original).

In addition to the similarity in attention distribution, it was also found that each head seems not to focus on the foreground information about flowers.

Based on this, this part introduces attention distance for further investigation. As shown in Figure 8, in the first few layers, the mean attention distance varies a lot from head to head. As the layer goes deeper, the mean attention distance of different heads begins to show convergence at high values. This result shows that at the beginning, some heads are responsible for local information, and some are responsible for global information. As the depth continues to increase, all the heads will all tend to integrate global information. After that, the benefits of layer training begin to decrease. The inflection point where benefits begin to decrease provides a vital reference for deep trade-off on the settings of depth and head hyperparameters.

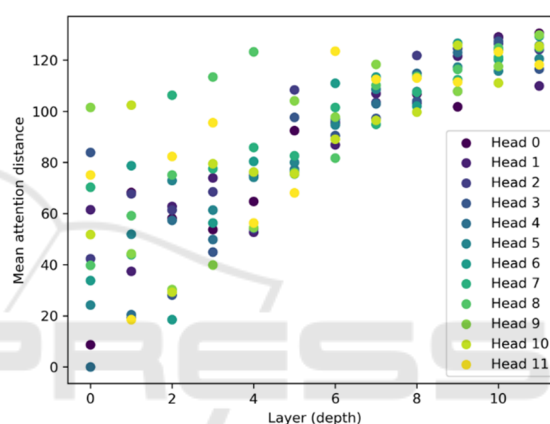


Figure 8: Mean attention distance distribution of heads on each layer (Photo/Picture credit: Original).

4 CONCLUSIONS

This study delves into the exploration of the crucial hyperparameters, namely head and depth, and their profound impact on the performance of the ViT model within the context of flower classification tasks. Beyond mere accuracy assessments, the investigation employs sophisticated techniques such as attention visualization and distance values to delve deeper into the significance of the multi-head design. Extensive and meticulously designed experiments were meticulously conducted to comprehensively analyse the intricate relationships between hyperparameters and model architecture. The findings reveal a compelling and unmistakable positive correlation between the depth of the model and its accuracy, as well as between the number of heads and the resulting accuracy. Furthermore, the study uncovers a critical issue within the ViT model, namely, the occurrence of performance saturation

attributed to attention collapse as the model's depth increases. This research endeavours to thoroughly dissect the underlying causes and consequences of such saturation and utilizes the inflection points derived from the mean attention distance distribution to navigate the intricate trade-offs involved in setting depth and head parameters. Looking ahead, future endeavours will place a central focus on addressing attention collapse as a primary research objective, striving to develop innovative methodologies to transcend its limitations and further refine the ViT model for enhanced performance in flower classification tasks.

REFERENCES

- Chi, Z. (2003). Data management for live plant identification. *Multimedia information retrieval and Management*, pp: 432–457.
- Das, M., Manmatha, R., & Riseman, E. M. (1999). Indexing flower patent images using domain knowledge. *IEEE Intelligent Systems and their Applications*, vol. 14(5), pp: 24-33.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp: 770-778.
- Hiary, H., Saadeh, H., Saadeh, M., & Yaqub, M. (2018). Flower classification using deep convolutional neural networks. *IET Computer Vision*, vol. 12(6), pp: 855–862.
- Kenrick, P. (1999). The family tree flowers. *Nature*, vol. 402(6760), pp: 358–359.
- Luo Y. (2022). Visualizer. <https://github.com/luo3300612/Visualizer>
- Nilsback, M.-E., & Zisserman, A. (2006). A Visual Vocabulary for Flower Classification. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp: 1447–1454.
- Nilsback, M.-E., & Zisserman, A. (2008). Automated Flower Classification over a Large Number of Classes. 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp: 722–729.
- Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp: 6105-6114.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, vol. 30.