Effectiveness of masked autoencoder in vision transformer models for image classification

Zhicheng Li[1,†] Yingze Liu[2,4,†] and Xinran Wang[3,†]

[1]School of Foreign Language, Harbin University of Science and Technology, Harbin, Heilongjiang, 150081, China
[2]College of Information Engineering, Xi’an University, Xi’an, Shannxi, 710000, China
[3]Department of Journalism and Communication, China Youth University of Political Studies, Beijing, 100089, China

[4]1811521134@mail.sit.edu.cn
[†]These authors contributed equally

Abstract. This paper uses the literature reading method to systematically sort out and introduce the basic principles of the three algorithms of the transformer model and their application in the field of image classification, which has high theoretical value and social value, and has strong reference for the development of the transformer model in the future. ViT is simply an innovation in computer vision based on the transformer model. It first separates an image into several local patches (16x16), and then maps each one to a feature vector. These vectors will be delivered to an encoder for polishing. Finally, a special token is appended to these vectors for integrating location information. The final prediction is based on these tokesn. Swin-T is a new Transformer architecture, which is proposed by Microsoft Research to improve the performance of computer vision tasks. It adopts a new windowed feature extraction strategy, which can maintain high accuracy while significantly reducing the amount of computation and memory consumption. It has achieved leading performance in multiple computer vision tasks, becoming one of the most advanced visual Transformer models. In computer vision image classification, the information is highly redundant, the lack of an image piece, may not make the model produce much confusion, the model can be inferred from the surrounding pixel information, masked autoencoder (MAE) is to mask a high proportion of image pieces, create a difficult learning task, the method is simple but extremely effective.

Keywords: image classification, transformer, vision transformer, swin transformer, MAE.

1. Introduction

Image classification aims at categorizing an image to its corresponding pre-defined category label. It is a fundamental mission in computer vision, and plays a significant role in downstream applications, such as behavior analysis, target tracking and target detection. Specifically, it could be applied to medical fields, security field and traffic scene recognition. As a branch of computer vision, image classification is no longer limited to computer vision, but has been extended to deep learning algorithms [1].

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

Among the deep learning algorithms, there are three main deep learning methods: active deep learning image, multi-label image and multi-scale network image. With the fast advancement of it, the algorithm model of deep learning is constantly improved and applied to various fields. For example: driverless cars, face recognition for security, object classification in images, traffic scene recognition and disease diagnosis [2].

The Transformer model was proposed in Attention is All You Need [3] in 2017. Transformer is different from recurrent neural network (RNN), original RNN can only be leveraged for a sequence with fixed length. However, in machine language, sentences often do not have the same length, which is the same as most seq2seq2 models that follow encoder-decoder structure. However, the transformer model contains 6 encoder and decoder modules with the same structure but different parameters. It is required to add location Embedding in the input. The model is based on the self-attention principle. Multiple self-attention forms the Multi-head Self-Attention module in this structure. It is the earliest model constructed merely by attention. The architecture enables it to not only compute faster and better, but also lay a foundation for following machine learning advancement. In this paper, some mainstream vision transformer models are introduced, followed by their representative applications.

2. Method

2.1. ViT

ViT is one of the most classical works applying the transformer for image processing [4]. In today's society, ViT is mostly used to train image recognition tasks with large enough data set, its architecture is demonstrated in Figure 1. First of all, images need to be input into the network. As opposed to conventional convolutional networks, transformer divide an image into several patches at first place. Usually, the patch size is 16x16. After the separation, these patches will be flattened and mapped by some operations to achieve refined feature representations. After the embedding operations, the spatial locations are missing. To mitigate the absence of spatial information, spatial token will be added to each patch and the mixture will be delivered to a transformer. Since all vectors represent a spatial position of the entire image, an extra vector is added in the transformer representing the category of the entire image. The structure of the transformer is displayed in Figure 1, where transformer encoder are stacked for L times.

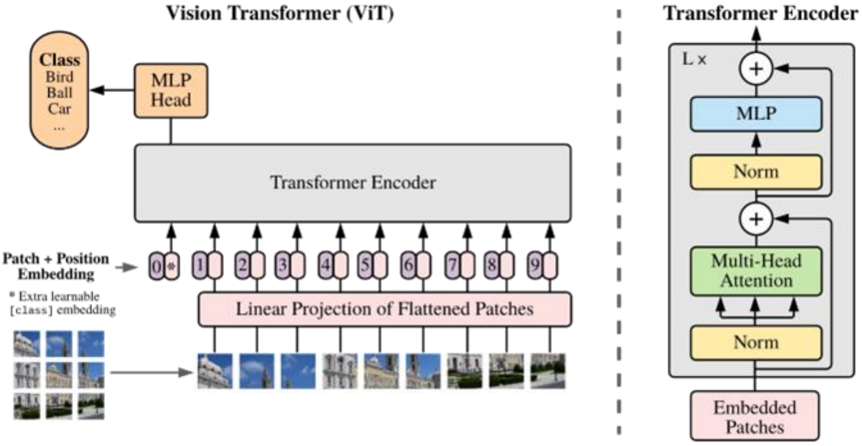


Figure 1. Structure of ViT. (Figure from: https://arxiv.org/abs/2010.11929)

Differently, CNN is also the mainstream method in computer vision. It could produce good performance. However, when transformer based on self-attention structure has enough data for pre-training, the performance of ViT will surpass CNN and break through the limitation of lacking inductive bias in transformer. Better migration effect can be obtained in downstream tasks.

ViT is a standard data stream that can be commonly used in other multimodal learning tasks. At the same time, because ViT will establish the relationship between data and the whole world, its calculation amount will have huge limitations compared with CNN. In the learning task of ViT, the two-dimensional image will be converted to one-dimensional data when the position embedding is installed, so that ViT will lose the perception ability of position when recognizing the image [5].

2.2. Swin transformer

It is a new transformer architecture proposed by Microsoft Research to improve the performance of computer vision tasks [6]. Swin Transformer, uses a new windowed feature extraction strategy that can strongly minimize the amount of computational cost and memory consumption while maintaining high accuracy. Swin-T has achieved leading performance in multiple computer vision tasks, becoming one of the most advanced vision transformer models available.

Today, computer vision is a key method in the development of autonomous driving, along with deep learning. In the study of autonomous driving, there is a way to divide the study into a data-driven computing framework consisting of sensing, planning, decision-making, and control modules.

In terms of sensing modules, Transformer is applied to segmentation of road scenes, text detection and application of traffic information, and object detection. In terms of control modules, Transformer is applied to the prediction of wheel odometry error [1].

2.3. MAE

In computer vision image classification, the information is highly redundant, and the lack of an image piece may not confuse the model much. The model can be inferred through the surrounding pixel information, and Masked Autoencoders (MAE) is to hide the high proportion of image pieces and create difficult learning tasks. The method is simple but extremely effective [7].

Like ViT, both encoder and decoder are composed of Transformer. The input image will first be divided into patches, and then all patches will be embedded to obtain the token of each patch. Then, a high proportion, such as 75% of patches, of these patches will be randomly masked. The encoder will only encode the patch that is not masked. After the encoding, the masked patch is represented by a shared and learnable vector, which is then restored to the original patch order together with the encoded patch, and delivered to decoder. Finally, the output is projected linearly (the number of input elements is the number of pixels in each patch) to obtain the reconstructed pixels. The differences between reconstructed and original input are penalized by the mean square error (MSE) loss function.

In the pre-training phase, MAE designs an asymmetric codec that feeds the visible portion into the encoder through a high proportion of the mask original. In addition, a narrow and shallow lightweight decoder is used to realize the target reconstruction of normalized pixels, accelerate the pre-training speed, and make MAE calculation more efficient. Note that the asymmetric encoding and decoding design allows MAE's encoder to use a complex neural network (stacked with a transformer architecture) and the decoder to use a simple lightweight network.

Thus MAE can be easily adapted to different data patterns. In this report, image classification task is mainly studied. It could also be found that transformer is still the backbone of this approach, but it still has some good training. MAE will prove to be a good start in machine learning.

3. Result and representative applications

3.1. Result comparison of MAE pre-training

Figure 2 shows the MAE pre-training and supervised pre-training conducted by the author through micro-adjustment in ImageNet-1K (224 size), and compared and evaluated the original ViT results of training in IN1K or JFT300M [7]. It can be seen that the accuracy of ViT-H (224 size) was 86.9% using only the IN1K data, but in the IN1K (no external data) benchmark test, by using 448size to fine-tune 87.8% accuracy, based on advanced network, the previous best accuracy reached 87.1%. These results are based on vanilla ViT. And ViT L part fine-tune the results related to the amount of fine-tuning the Transformer under the default Settings block, Tuning 0 blocks is linear probing; 24 is full fifine-tuning. And in IN1K training, ViT-L will degrade with the change of parameters. While it may seem that MAE is behind ViT just by looking at the performances, it is significant to mention that MAE is only pretraining on ImageNet-1K, while ViT is pretraining on JFT (which is an extremely large data set). MAE has two characteristics: simplicity, which means that MAE adds some simple extensions to the vit model, but the model itself is not simple. Good scalability means that you can add data indefinitely. After all, MAE doesn't need labels and larger models could be trained.

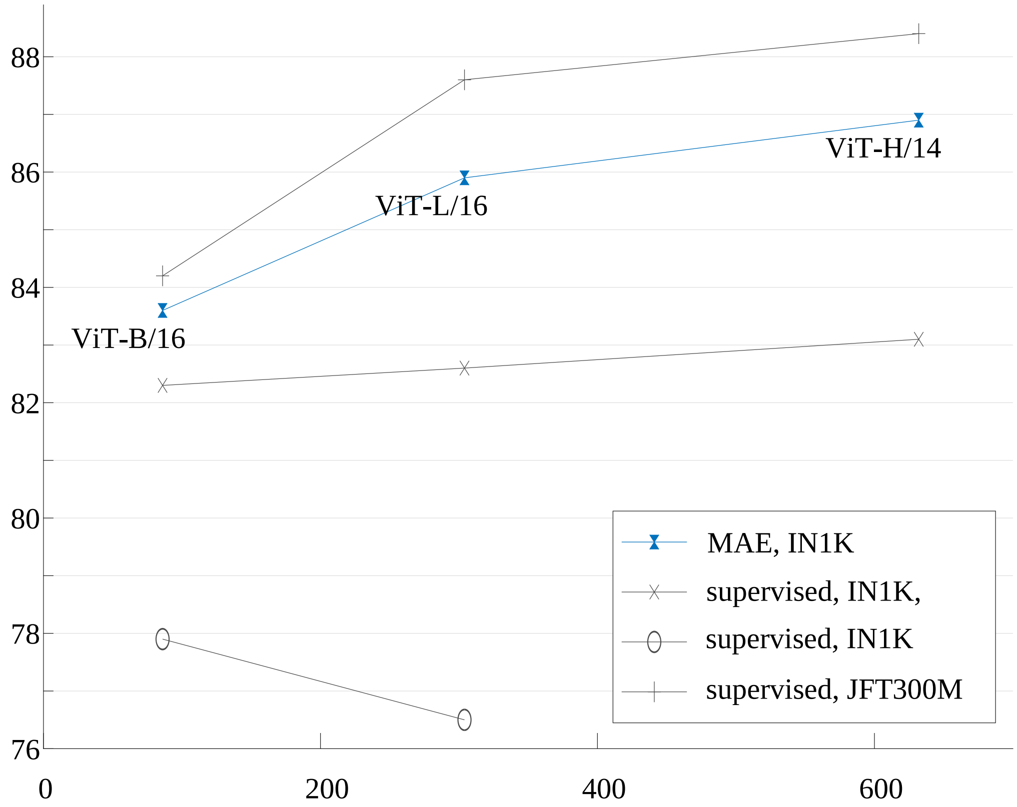


Figure 2. Result comparison using MAE pre-training strategy.

It can be seen from this figure that MAE pre-training results in better results, and for models with higher capacity, the benefit of starting from scratch training is greater, which indicates that MAE can help expand the model size.

3.2. Transformer for autonomous driving

Today, computer vision is a key method in the development of autonomous driving, along with deep learning. In the study of autonomous driving, there is a way to divide the study into a data-driven computing framework consisting of sensing, planning, decision-making, and control modules.

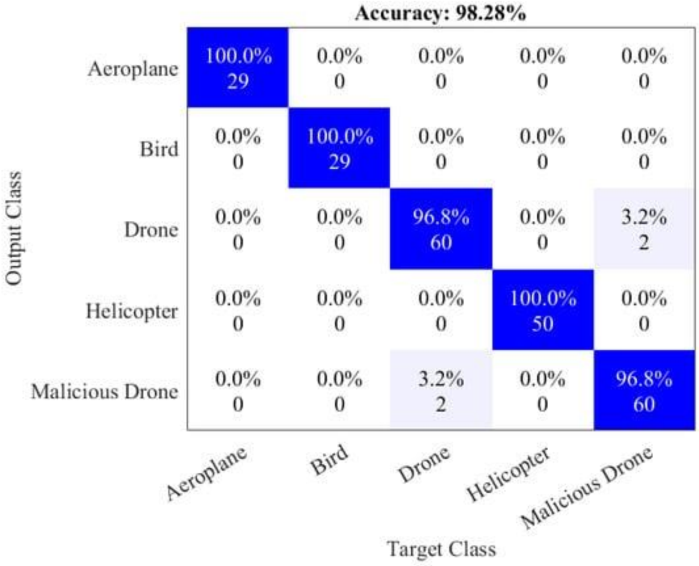
Many scholars have compared different advanced models of computer algorithms in computer vision [8]. In several cases, transformer has shown a better effect than CNN in both performance and speed. For example, in a case, after the author's train, the result shows that the pruned SimpleDist architecture with lightweight vision transformer as encoder achieves higher performance and faster inference speed than fully-convolutional baselines, and compared with SimpleBaselines-ShuffleNetV2, SimpleDist-BoTShuffleNetV2 has similar test AP but with 68.4% fewer parameters and 28.6% improvement of inference speed on MHP dataset.

3.3. Transformer for malicious drones detection

In the utilization of VIT to distinguish malicious drones, it describes a method for identifying malicious drones from other types of drones, as well as aeroplanes, helicopters, and birds, based on their visual characteristics [9]. The images of these objects are fed into a series of handcrafted descriptors, D-CNNs, and a ViT-classifier, which work together to extract features and train a classification model.

The ViT classifier uses a transformer encoder to divide the input images into 14x14 vectors with patches of 16x16, which are then embedded using learnable position vectors and fed into parallel attention heads. The outputs from these heads are concatenated and fed into a neural network to output the ultimate prediction [10].

The model achieves an overall accuracy of 98.28%, with perfect accuracy numbers for aeroplanes, birds, and helicopters. Detailed performances are demonstrated in Figure 3. The accuracy for drones and malicious drones is slightly lower, at 96.8% [10]. Nonetheless, the results indicate that this is an effective and robust method for distinguishing drones and other objects based on their visual characteristics, and may be useful for identifying potential security threats.



Accuracy: 98.28%

Output Class \ Target Class	Aeroplane	Bird	Drone	Helicopter	Malicious Drone
Aeroplane	100.0% 29	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Bird	0.0% 0	100.0% 29	0.0% 0	0.0% 0	0.0% 0
Drone	0.0% 0	0.0% 0	96.8% 60	0.0% 0	3.2% 2
Helicopter	0.0% 0	0.0% 0	0.0% 0	100.0% 50	0.0% 0
Malicious Drone	0.0% 0	0.0% 0	3.2% 2	0.0% 0	96.8% 60

Figure 3. Classification performances. (Figure from: https://www.mdpi.com/2673-2688/3/2/16)

4. Conclusion

This paper systematically combs and introduces the basic principles of three algorithms of transformer model, which are successively leveraged to a series of visual tasks (such as various classification, detection, and segmentation) and data streams (such as images, texts, etc.). For visual tasks specifically the image classification, this paper proposes several specific methods, and evaluates and compares their performance on their models. Therefore, it has high theoretical and social value, and has strong reference value for the development of transformer models in the future. In the era of rapid development of artificial intelligence, transformer algorithm update iteration speed is very fast, so transformer in image classification development still has a lot of room for development. It is expected that this article will help readers better understand the various visual Transformers before deciding to explore them further. According to the current research trend, future research on image classification methods based on vision transformer will focus on how to better introduce local and hierarchical structure, to complement with Transformer's global character and improve model performance. At the same time, self-supervised learning Transformer model is also a hot research direction in the future. Since Transformer and CNN have their own advantages and disadvantages and cannot be replaced by each other, the construction of multi-modal classification algorithm combining the advantages of the two is also one of the hot research spots in the future.

References

[1] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM computing surveys (CSUR), 54(10s), 1-41.

[2] Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352-2449.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et, al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et, al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[5] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270.

[6] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., et, al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, 10012-10022.

[7] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000-16009.

[8] Chen, H., Jiang, X., & Dai, Y. (2022). Shift Pose: A Lightweight Transformer-like Neural Network for Human Pose Estimation. Sensors, 22(19), 7264.

[9] Lykou, G., Moustakas, D., & Gritzalis, D. (2020). Defending airports from UAS: A survey on cyber-attacks and counter-drone sensing technologies. Sensors, 20(12), 3537.

[10] Jamil, S., Abbas, M. S., & Roy, A. M. (2022). Distinguishing malicious drones using vision transformer. AI, 3(2), 260-273.