The Investigation of DeiT model Based on PaddlePaddle Framework on CIFAR-10 Dataset Image Classification

Yuda Li

Computer Science and Technology, Beijing Jiaotong University (Weihai Campus),
Shandong, 264401, China
20722089@bjtu.edu.cn

Abstract. Image classification is one of the important classifications in the field of computer vision, and the development of deep learning models has brought historic breakthroughs to the development of image classification. Transformer model, as a powerful sequence modeling tool, has achieved great success in natural language processing. Recently, the application of the Transformer model to image classification tasks has also achieved significant results. Distilled-Enhanced-Transformer (DeiT) model is one of the representative models, which realizes the goal of efficient image classification on small data sets by introducing self-attention mechanism and Transformer architecture. The core idea of DeiT model is to use self-attention mechanism to establish the global context of input image. Traditional convolutional neural networks capture image features through local receptive field and hierarchical structure when processing images. The DeiT model, on the other hand, processes the image in chunks, breaking it into small patches and feeding them into Transformer as a sequence. In this way, the DeiT model is able to model each patch using self-attention mechanisms to capture more global image features. In the experiment, this study used the DeiT model provided in the PaddlePaddle 2.0 framework to perform an image classification task on the CIFAR-10 dataset. The CIFAR-10 dataset contains 60,000 32x32 color images from 10 different categories, 50,000 for training and 10,000 for testing. This study trained the model using a stochastic gradient descent (SGD) optimizer and a cross entropy loss function. In order to improve the generalization ability of the model, this study also use data enhancement techniques such as random cropping, flipping and rotation.

Keywords: DeiT, CIFAR-10, PaddlePaddle

1 Introduction

Machine learning and computer vision are the most concerned research directions in the field of artificial intelligence [1-3]. In image classification tasks, LeNet is an early classical convolutional neural network model [4], which extracts image features through convolution layer, pooling layer and fully connected layer, and is trained by gradient descent algorithm. The emergence of AlexNet has attracted a lot of attention

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_106

[5], with its breakthrough results at the 2012 ImageNet competition, which greatly improved image classification performance by introducing ReLU activation function and Dropout regularization technology, as well as deeper network structure and larger data sets. The VGG model further deepens the network structure by using smaller convolution kernel sizes and pooling layers, making it more nonlinear expressive [6]. ResNet introduces residual connections to solve the problem of disappearing gradients and exploding gradients in deep network training [7], making deeper network structures possible. The SENet model adds a Resnet-based attention mechanism [8], which enables the model to automatically focus on important features in the input, further improving the performance of image classification tasks. In the field of natural language processing, the advent of the NLP Transformer model is revolutionary [9]. It uses a self-attention mechanism to capture long-distance dependencies in text, resulting in significant performance improvements in tasks such as machine translation and text generation. More recently, with the continuous integration of the interdisciplinary fields of computer vision and Natural language processing, Vision Transformer (VIT) and DeiT models introduced the Transformer architecture into the field of computer vision [10, 11], and made pioneering research achievements in image classification task processing. These models achieve a global understanding of the entire image by splitting the image into small chunks and modeling those chunks using a self-attention mechanism. In addition, the DeiT model introduces Knowledge Distillation, which enables small models to achieve performance that approaches or even exceeds that of the original large model by learning from the larger model. Delving into these models and techniques will not only help people better understand the evolution of machine learning and computer vision, but will also guide people in selecting and optimizing suitable models for practical applications. As technology continues to advance, it can be expected that more innovative models and algorithms to emerge, bringing greater breakthroughs and advances in the field of machine learning and computer vision.

The Transformer model can better model global information by introducing a self-attention mechanism to capture long-distance dependencies. In the PaddlePaddle 2.0 framework, the Distilled-Enhanced-Transformer (DeiT) model is provided, which is an improved model based on the Transformer architecture. The DeiT model is distilled to transfer the knowledge of the large Transformer model to a smaller model, thereby reducing the number of model parameters while maintaining performance. The model was trained and optimized on the CIFAR-10 dataset to improve the accuracy of the image classification task. The CIFAR-10 dataset is a commonly used image classification benchmark dataset containing 60,000 32x32 pixel color images in 10 different categories. This dataset is difficult for algorithms because the image resolution is relatively low and there are certain similarities and overlaps between the categories. The purpose of this study is to study the image classification task on the CIFAR-10 dataset using the DeiT model provided in the PaddlePaddle 2.0 framework. By comparing the DeiT model with the traditional CNN model in terms of accuracy, computational efficiency and model complexity, this study will evaluate the potential of DeiT model in improving the performance of CIFAR-10 image classification tasks. This research contributes to the development of image classification and provides new ideas and methods for applying Transformer model to image processing tasks.

2 Methodology

2.1 Dataset Preparation

The CIFAR-10 that is a widely used dataset in the computer vision task, is chose for training and evaluating and DeiT model in the PaddlePaddle 2.0 framework. The CIFAR-10 dataset contains 10 categories: "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", "truck". There are 60,000 images in the dataset, with 6,000 images per category. Each image is 32×2 in size and is a color image in RGB format. In terms of preprocessing, the methods employed in this study are image normalization and data enhancement. Normalization scales the image pixel values to between 0 and 1 to better train the model. Data enhancement includes random clipping, random flipping, random rotation and other operations, which can increase the diversity of data and improve the generalization ability of the model.

2.2 DeiT

VIT is an image classification model based on Transformer architecture. It divides the image into fixed-size blocks and captures the global relationship between the blocks through a multi-layer self-focusing mechanism, thus achieving the image classification task. In PaddlePaddle 2.0 framework, the DeiT model is provided, which is an improvement and optimization of the VIT model. DeiT model uses distillation technology in the training process, combined with the knowledge of large-scale teacher model, guides the training of small-scale student model, thus improving the performance and generalization ability of the model. DeiT model combines Transformer's self-attention mechanism and extraction technology and has the following characteristics :1) Self-attention mechanism: DeiT model uses self-attention mechanism to capture the global relationship between image blocks, so as to effectively model the dependency relationship between image features. 2) Distillation technology: Through distillation technology, the DeiT model can learn more knowledge from the large Teacher model and pass this knowledge to the small Student model, thereby improving the performance and generalization ability of the model. 3) Image classification task: The DeiT model is mainly applied to tasks such as image classification, which can classify input images and identify object attributes or scene types in the images.

2.3 Implementation Details

The implementation details of the image classification task on the CIFAR-10 dataset using the DeiT model provided in the PaddlePaddle 2.0 framework are as follows: The details of the image classification task for the CIFAR-10 dataset using the DeiT model in the PaddlePaddle 2.0 framework are as follows: 1) Loss function: The cross entropy loss function (CrossEntropyLoss) is used as the optimization objective in the training process. 2) Evaluation index: Accuracy is used as an indicator to evaluate the performance of the model, and accuracy represents the proportion of correctly classified samples in the total number of samples. 3) Optimizer: Select optimizers to update the parameters of the model, common optimizers are stochastic gradient Descent (SGD) and Adam, etc., which can be selected and configured according to needs. 4) Batch size (batch_size) : Specifies the number of image samples contained in each training batch, which can be adjusted based on computational resources and model complexity. In this experiment this study got a value of 64 5) learning_rate: Sets the learning rate of the optimizer and controls the pace of parameter update. It can be adjusted according to the actual situation. In this experiment this study took a value of 0.0011 6) Epoch: specifies the number of times the entire training set is traversed during the training. Each epoch represents one complete training cycle. In this experiment, this study took a value of 10.

3 Results and Discussion

3.1 Experimental Results and Analysis

Following the completion of the training phase, this study proceeded with a comprehensive evaluation of the model shown in Table 1 by employing an independent test set, thereby facilitating an accurate measurement of the classification accuracy. Through rigorous experimentation and meticulous analysis, the investigations culminated in the determination that the DeiT model yielded a classification accuracy of 94.32% when applied to the CIFAR-10 dataset. These findings not only serve to validate the model's effectiveness in accurately categorizing the diverse range of images encompassed by the CIFAR-10 dataset but also underscore the robustness and reliability of the approach.

Table 1. Accuracy of perturbation attack in different filter effects on the CIFAR-10.

Model	Test set classification accuracy
DeiT	94.32%

The experimental results provide compelling evidence that the DeiT model exhibits commendable classification performance when confronted with the CIFAR-10 dataset, as evidenced by the attainment of an impressive classification accuracy of 94.32%. Such outcomes suggest that the DeiT model, leveraging the power of visual Transformers, demonstrates remarkable generalization capabilities and adaptability when confronted with image classification tasks.

4 Conclusion

The DeiT model, which is built upon the visual Transformer architecture, has demonstrated notable performance and remarkable generalization capabilities in image classification tasks, thereby exhibiting substantial potential and prospects for future research endeavors. This study delves deeply into the image classification task involving the DeiT model based on visual Transformer applied to the CIFAR-10 dataset. Empirical findings reveal that the DeiT model exhibits favorable classification performance and generalization abilities, yielding a classification accuracy of 94.32% on the CIFAR-10 dataset. In contrast to conventional convolutional neural networks, the DeiT model, leveraging its visual Transformer foundation, offers advantages such as global feature modeling, scalability, generalization abilities, and interpretability. Through the utilization of multiple Transformer encoders, the DeiT model facilitates feature extraction and image classification. Moreover, the model incorporates techniques such as tag smoothing, Dropout regularization, and random data augmentation to enhance generalization abilities and mitigate overfitting concerns. Future investigations will focus on exploring the performance and application efficacy of the DeiT model on larger datasets. Additionally, endeavors will be made to integrate other deep learning models and algorithms to further enhance the accuracy and efficiency of image classification tasks. Concurrently, efforts will be devoted to applying the DeiT model in practical scenarios, optimizing the model's structure and parameters to meet the demands of real-world applications. The continuous exploration and innovation in these areas will contribute to the advancement and widespread utilization of deep learning technology in the field of image classification.

References

1. Voulodimos, A., Doulamis, N., Doulamis, A., et al.: Deep learning for computer vision: A brief review. Computational intelligence and neuroscience, (2018).
2. Yu, Q., Wang, J., Jin, Z., et al.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72: 103323 (2022).
3. Chai, J., Zeng, H., Li, A., et al.: Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications, 6: 100134 (2021).
4. Le, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 (1998).
5. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25 (2012).
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, (2014).
7. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016).

8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 7132-7141 (2018).
9. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in neural information processing systems, 30 (2017).
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
11. Touvron, H., Cord, M., Douze, M., et al. Training data-efficient image transformers & distillation through attention, International conference on machine learning. PMLR, 10347-10357 (2021).