中国科学: 信息科学

SCIENTIA SINICA Informationis

AI Course Project

Few shot Image Classification

李晓畅, 龚宇昊, 黄润, 孙德恺, 李咸若

摘要 This report presents our attempts of utilizing transfer learning for few-shot image classification. The project is structured around two main tasks. We first conducted extensive measurements on various classifier, optimizer, and loss functions to determine the optimal combination and fine-tune the model parameters; secondly, we employed a variety of techniques to augment the sample data to make the best use of the available information, thus improving the generalization and overall performance of our model. By combing the ability of model and with features of specific task, we achieved notable results in all three phases of the experiment, surpassing many of our classmates. Most importantly, we gain a comprehensive understanding about how to address the challenges of limited labeled data by utilizing the knowledge learned from other data-sets.

关键词 Transfer Learning, Visual Transformer, Data Augmentation

1 Introduction and Design Ideas

In this section, we provide an overview of some fundamental concepts, such as transfer learning, and highlight several influential works that have significantly informed our experiments. These include the cutout data augmentation technique, some interesting characteristics of Vision Transformer (ViT), and ways to effectively fine-tune a model and optimize the training process [5].

1.1 Transfer Learning

The goal of Transfer learning is basically to mimic human cognition, where individuals can easily establish an understanding of new concepts with just one or a few examples. It enables the reuse of knowledge learned from previous tasks, saving significant time and resources. Instead of training a model from scratch, transfer learning allows models to start with pre-existing knowledge and adapt it to new tasks. This is particularly valuable when labeled data for the target task is limited or expensive to obtain.

1.2 Cutout Augmentation

Cutout augmentation [1] is a popular image data augmentation technique used to improve widely used technique in image data augmentation, specifically designed to enhance the robustness and generalization of deep learning models. It involves randomly masksing out square-shaped patches within from training images by setting their pixels to zero. By introducing local occlusions, cutout encourages models to rely on more robust features, enhancing their generalization and reducing over-fitting. This regularization technique increases resilience to noise and promotes spatially invariant representations. With adjustable patch sizes, cutout proves effective across computer vision tasks like classification, object detection, and segmentation. Its simplicity and ability to improve model performance make cutout augmentation a valuable tool, particularly in scenarios with limited training data.

In practice, the size and shape of the cutout patches can be adjusted based on the specific requirements of the task. Generally, larger cutout patches tend to provide more regularization but may also remove important information, while smaller cutouts offer more fine-grained control over occlusions.

1.3 Interesting facts about Vision Transformer

According to [4], there are three important facts about Vision Transformers. Firstly, the residual layers of Vision Transformers can be processed efficiently in parallel without significantly affecting accuracy. Secondly, fine-tuning the weights of the attention layers is sufficient to adapt Vision Transformers to a higher resolution and to other classification tasks. Last but not least, adding MLP-based patch pre-processing layers improves Bert-like self-supervised training based on patch masking.

These findings have been extremely enlightening and our model design has been heavily informed by these three insights. For example, inspired by the fact of fin-tuning only the attention layer can achieve good performance, we design different learning rate for different part of our model, which not only reduced training time, but also promoted our result.

2 Implementation

2.1 Model Selection and Classifier Design

Our implementation is direct and simple. One pre-trained model as feature extractor, mapping images to features; and one simple classifier using features to decide the class. As mentioned above, our work is always based on [2]. We tried various models in early stage of our implementation, performance of similar size model like Deit [3] or Resnet [5] is always worse than [2]. Another reason is that, as a classical model, we can easily find information to leverage our design. No more mention its balance of performance, parameter numbers and training cost. In a word, we think

it is the best choice of our task. Inspired by the fact that small vit [4] perform better on small data-set, we tried to decrease our parameters. This change improved our performance. Our design of classifier is easy. Only a Layer-Norm layer and a Liner layer included. Design of easy classifier can also be supported by Vit [4] for keeping a small MLP as attention layers increased.

2.2 Data Augmentation

We employed multiple data augmentation methods, including rotation, flip, affine transformation, color jitters, Gaussian blur, random crop, and cutout. The detailed algorithm and procedure can be found in Algorithm 1.

算法 1 Data Augmentation Procedure

Require: is_train, args, img_size, mean, std

Initialize a list of pre-defined transformations Aug = [
Resize(256),
RandomCrop(224, padding=16),
Cutout(0.5, scale=(0.1, 0.2)),
ColorJitter(0.2, 0.2, 0.2),
GaussianBlur(1, 2.0),
RandomHorizontalFlip(p=0.3),
RandomVerticalFlip(p=0.3),
RandomRotation((-10, 10))]
if *is_train* is true **then**
    Create a list TL for training transformations
    Append Aug to TL
    Append ToTensor and Normalize transformations to TL
    return Composed transformations (TL)
else
    Create a list TL for non-training transformations
    Append Resize, CenterCrop, ToTensor, Normalize transformations to TL
    return Composed transformations (TL)
end if

2.3 Model Parameters

• **Batch Size:** Set to 16. In the early stages of the experiment, considering limited training resources and the eager to try more combination to determine effects of certain parameter, we initially tested the model using a larger batch size. Once other parameters were determined, we made a decision based on a comprehensive evaluation and experimentation, ultimately selecting a smaller batch size of 16.

• **Epochs:** Set to 30. A small number of epochs may result in under-fitting, while a large number may lead to over-fitting. Still, considering the small sample size in few shot learning makes it more prone to over-fitting, we opted for a smaller value of 30 for the number of epochs to mitigate this risk.

• **Learning Rate:** Set to 1e-4. Since this experiment utilizes transfer learning, which was built upon a pre-trained model, it is crucial to avoid setting the learning rate too high, as doing so could lead to oscillations and thereby hinder convergence.

• **Weight Decay:** Set to 0.01. The weight decay parameter is critical for regularization. It reduces the magnitude of adjustments to avoid excessive fluctuations. It also helps prevent the occurrence of oscillations by constraining the weights' values and promoting smoother updates.

3 Result and Evaluation

3.1 Validation Set

This is performance of vit-small with 20M parameters, 16 patch, 30 epochs on val set.

	cifar	country	food	oxford	stanford
ACC1	61.7	35.5	62.0	81.8	43.0
ACC5	77.3	39.3	76.3	97.0	55.1
LOSS	2.32	4.27	2.26	1.01	3.32

3.2 Test set

Similar result in test set.

	cifar	country	food	oxford	stanford
ACC1	0.04	0.48	0.78	0.16	0.35

参考文献

1 T. Devries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

2 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

3 H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021.

4 H. Touvron, M. Cord, A. El-Nouby, J. Verbeek, and H. Jégou. Three things everyone should know about vision transformers. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, volume 13684 of *Lecture Notes in Computer Science*, pages 497–515. Springer, 2022.

5 R. Wightman, H. Touvron, and H. Jégou. Resnet strikes back: An improved training procedure in timm. *CoRR*, abs/2110.00476, 2021.