

Transformer-Based Fine-Tuning and Zero-Shot Learning for Image Classification

Haoyang Fei^a

School of Software, South China University of Technology, Guangdong, China

Keywords: Image Classification, Vision Transformer (ViT), Contrastive Language-Image Pretraining (CLIP), Zero-Shot Learning.


Abstract: An automated, efficient, and accurate image classification approach is essential across various domains. This research compares different state-of-the-art image classification approaches, including the fine-tuned Vision Transformer (ViT), base Contrastive Language-Image Pretraining (CLIP) model, and fine-tuned CLIP model, on specialized image classification tasks. The research evaluates classification accuracy, zero-shot classification ability for unseen categories, and deployment costs. The findings indicate that while the fine-tuned ViT model excels in test accuracy, the base CLIP model demonstrates remarkable zero-shot learning capabilities, making it highly efficient for unseen categories. However, fine-tuning the CLIP model results in a significant loss of its zero-shot ability without a proportional increase in performance, with the fine-tuning cost far exceeding that of the ViT model. The author suggests that the fine-tuned ViT model is more suitable for tasks requiring high accuracy, while the base CLIP model is preferable for applications valuing versatility and lower deployment costs. Fine-tuning the CLIP model is suitable only if the dataset is sufficiently large and deployment cost is not a concern. These insights provide a nuanced understanding of the trade-offs involved in selecting an image classification model for specialized tasks, emphasizing the importance of considering both the task's nature and available resources.

1 INTRODUCTION

Image recognition and classification techniques are crucial for many fields including medical fields, media creation field, design field and data science field. However, interpreting images manually is labor intensive and sometimes difficult. Like in the medical field, the interpretation of medical images requires specialized radiologists who are scarce globally. As the volume of images all over the world increases, the time and labor intensity required is too enormous to be applied in actual production environment (He, 2015). These obstacles are hard to overcome with human efforts alone. In machine learning, Image classification task is a complex but mature task. In recent years, image classification models like LeNet, Residual Network (ResNet) (He, 2016), and Vision Transformer (ViT) (Dosovitskiy, 2020) are widely used on image classification. LeNet and Residual Network is based on Convolutional Neural Network (CNN), while Vision Transformer (ViT) is an

emerging model that applies the Transformer architecture to image recognition tasks, and it have a best performance in average in all the Computer vision (CV) models above (Rawat, 2017).

However, most of the out-of-the-box image classification model only have a great performance for common classes like 1000 classes in ImageNet. For a more detailed categories, like classify different medical images, clothes styles, food types, etc., a fine-tune is commonly required. Fine-tuning a model requires an annotated dataset and computing power, which is costs that need to be considered while applying image classification model into production. A model called Contrastive Language-Image Pretraining (CLIP) introduced by OpenAI aiming to provide an approach to make a zero-shot image classification, like the “zero-shot” capabilities of GPT and other Large Language Models (Radford, 2021). CLIP uses contract learning to train an image encoder and a text encoder separately, then align these encoders with cosine similarity. This feature enables

^a <https://orcid.org/0009-0005-2652-1343>

CLIP model to use natural language to perform classification and allows CLIP to handle different categories way more than the 1000 in ImageNet datasets. However, since CLIP is trained with images and natural-language descriptions from variant aspects, so it might not have a great performance on a specific aspect without extra fine-tuning. In actual production cases, choosing a correct method to perform the image classification task might be challenging. Developers must consider a series of metrics including the accuracy of classification, deployment cost, and other aspects that might impact the result of deployment. The main purpose of this study is to compare the performance of popular transformer-based image classification models in a specific case, then provide a conclusion on how to choose a model on specific classification task. Specifically, first, the training set of food-101 dataset is used to finetune the model in this research (if finetune is necessary). Second, the valid set of Food-101 dataset is used for evaluating each model's performance based on several metrics.

This research compares three different approaches for specialized image classification: classification with a fine-tuned ViT model (ViT-16/B used in this study), classification with the CLIP model (without fine-tuning, base), and classification with a fine-tuned CLIP model (base). The study evaluates the performance of these approaches across three metrics: valid accuracy, which measures accuracy when inferring on the validate dataset during deployment (including fine-tuning); cost, representing the computer power required for fine-tuning; and extended accuracy (for CLIP-based models only), measuring accuracy when classifying categories not seen during fine-tuning. The primary contribution of this study lies in its comprehensive comparison and analysis of different methods, providing empirical data to support specialized image classification and offering insights for further research in the field. The findings will aid in optimizing the selection and deployment of image classification models, improving classification accuracy, reducing computational costs, and driving technological advancements and applications in related domains.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

In this research, the author leverages the Food101

dataset, a comprehensive real-world food dataset created from diverse food photos collected from internet media, with a manually annotation. Unlike previous datasets, such as pfid, which primarily consist of standardized fast-food photos collected under specific, unified conditions, Food-101 offers a diverse collection of images representing 101 different named dishes around the globe. The dataset comprises 101,000 images from real-world. Each category consists of 750 images for training and 250 images for testing. Notably, the training images intentionally retain some level of noise, including intense colors and occasional mislabeling, to better simulate real-world scenarios and challenge computer vision algorithms. With a maximum side length of 512 pixels, Food101 encompasses a wide variety of food classes, ranging from Apple pie to Bibimbap, facilitating research into scalable recognition algorithms (Bossard, 2014). Examples of the dataset are illustrated in Figure 1.

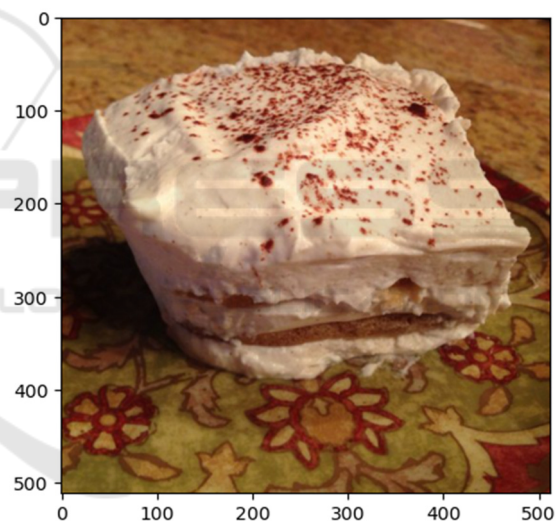


Figure 1: Sample Image of food-101 dataset (Photo/Picture credit: Original).

2.2 Proposed Approach

In this research, the dataset is sliced into 2 parts: The first part $C_1..C_{10}$ will be used for extended accuracy evaluation for CLIP-based models. Then the remaining 91 categories $C_{11}..C_{101}$ will be used for model training, and accuracy evaluation. The dataset is partitioned into training and validation set at a ratio of 0.8:0.2, respectively. The training dataset is used to finetune both ViT and CLIP models, while the validation dataset is used to evaluate these models' performance. While fine-tuning, the author uses the AdamW optimizer to increase the model's

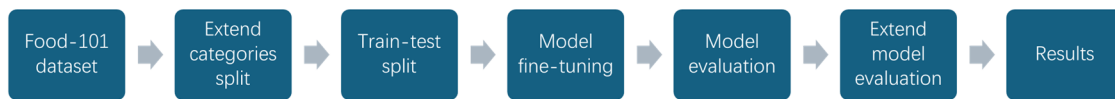


Figure 2: The pipeline of the research (Photo/Picture credit: Original).

generalization performance and reduce the training cost (Loshchilov, 2017). Figure 2 illustrates the comprehensive pipeline of the research.

2.2.1 Transformer

Both ViT and CLIP uses transformer architecture as its encoder. The transformer model is an architecture of deep learning model that adopts a self-attention mechanism to process sequential data, such as text and images (Vaswani, 2017). Transformer generally contains two parts, which are an encoder and a decoder, both composed of numerous layers of self-attention and feedforward neural networks. With the self-attention mechanism, transformer-based models are able to focus on the most important parts of the source sequence when making predictions. This enables the model to capture long-range dependencies and improve the performance when performing sequential tasks.

2.2.2 Vision Transformer (ViT)

In this study, the author introduces the ViT model for image classification tasks, inspired by the success of the Transformer architecture in natural language processing. Unlike traditional approaches that rely on CNN for computer vision tasks, ViT directly applies the Transformer architecture to sequences of image patches. This departure from CNN-based methods demonstrates promising performance across a spectrum of image classification benchmarks, including ImageNet, CIFAR-100, and VTAB (Dosovitskiy, 2020).

The ViT model operates by initially reshaping images into a sequence of flattened 2D patches, which are then processed through the Transformer architecture. Each patch undergoes a trainable linear projection to generate a fixed-dimensional embedding. Similar to the BERT model in NLP, ViT incorporates a learnable embedding at the start of the patch sequence, serving as input representation for the encoder of ViT. The encoder consists of several alternating layers of multi-head self-attention and MLP layers, with layer normalization and residual connections at each layer. Position embeddings added to the patch embeddings are used for retaining positional information. Notably, ViT displays less image-specific bias compared to CNNs. The reason is

that only the MLP layers exhibit local and translational equivariance, whereas the self-attention layers are global in nature. Furthermore, ViT supports a hybrid architecture where the input sequence can be generated from CNN feature maps, offering flexibility in model design. In this study, the author fine-tuned the `vit_base_patch16_224` model on 91 different categories of images from the food-101 training dataset and evaluated its performance on the validation dataset to showcase its effectiveness compared to the CLIP model. Detailed descriptions of the experimental setup, including training procedures and hyperparameter settings, are provided in subsequent sections.

2.2.3 Contrastive Language-Image Pretraining (CLIP)

CLIP is a multimodal model that learns to associate images and text through a contrastive objective. CLIP is trained on a large-scale dataset of images, together with their associated text, such as image captions, to learn a joint embedding space in which semantically similar image and text pairs have less distance to each other. This joint embedding space enables CLIP to perform a diverse array of vision-language tasks, which includes image classification, image segmentation, and detection of objects (Radford, 2021). CLIP can be seized as a composition of two separate components: a vision encoder and a text encoder. The images are first resized into a fixed size (224 by 224 in CLIP ViT/B-32) and normalized into standard pixel values. The image encoder takes the normalized images as input, passes the image through a CNN backbone, such as ResNet or ViT, to extract image features. The extracted image features are then projected to a fixed-dimensional embedding using a learnable linear projection.

In this research, the author fine-tuned the CLIP model on 91 different categories of images from food-101 training dataset which consists of the image and it correspond text label. All the text labels are extracted from the class names of the image, which are then tokenized and processed by the text encoder to acquire text embeddings. These embeddings, together with the image embeddings processed by the image encoder are then compared using cosine similarity to determine the semantic similarity between the image and labels from the dataset. The

result of the classification is determined by the label with the highest similarity. Since fine-tuning CLIP is challenging and the choose of hyperparameters significantly affect the evaluation result, the author leverage the fine-tuning approach from Xiaoyi Dong to provide a best result possible (Xiaoyi, 2022). The author provides detailed descriptions of experimental setup in the following sections.

2.2.4 Cosine Similarity

In CLIP-based models, Cosine Similarity is used to compare the similarity between the image and text through their embeddings and determine the best-match category. Cosine similarity provides a metric to measure the similarity of two non-zero vector's direction. The definition of cosine similarity is the cosine of the angle between two input vectors. For instance, if there are two input vectors (v_1 and v_2), the cosine similarity of them can be represent as:

$$\cos(\theta) = \frac{v_1 \cdot v_2}{|v_1||v_2|} \quad (1)$$

The greater the cosine similarity means the more similar the two vectors are. In the research, Cosine Similarity is used to determine the corresponding categories of image while evaluating CLIP-based models.

2.3 Implementation Details

This research used Python 3.10 and Pytorch 2.1.1 for implementing all the models above. ViT model and CLIP model is based on hugging face transformers (Wolf, 2020). Data visualization is provided by Matplotlib. All the training and testing is completed on Ubuntu 22.04 over 2x Nvidia RTX4090 24G with cuda version of 12.1 and 64GB of System Memory. The training hyperparameters for the Google ViT model are detailed in Table 1, while those for the CLIP model are provided in Table 2.

Table 1: Hyperparameters of ViT model fine-tuning.

Hyperparameter	Value
Base Model	vit-base-patch16-224
Learning Rate	10^{-4}
Batch Size	256
Weight Decay	0.05
Epochs	10
Optimizer	AdamW
β_1, β_2	0.9, 0.99
\mathcal{E}	$1e^{-6}$

Table 2: Hyperparameters of CLIP model fine-tuning.

Hyperparameter	Value
Base Model	CLIP ViT-B/32
Learning Rate	10^{-4}
Batch Size	2048
Weight Decay	0.05
Epochs	10
Optimizer	AdamW
β_1, β_2	0.9, 0.999
\mathcal{E}	$1e^{-6}$

3 RESULTS AND DISCUSSION

The result of the research shows a significant better performance on the fine-tuned model based on Google ViT model then other models. The base CLIP model also has a great performance in the research, especially considering that it works out-of-the-box and no extra training steps are required and have a great zero-shot ability to unseen categories. However, CLIP model is much harder to fine-tune compared to base ViT model, the training cost of fine-tuning clip model is larger than ViT while the performance is worse. Besides, fine-tuned CLIP model loses most of its zero-shot ability. The overall result is as shown in Figure 3.

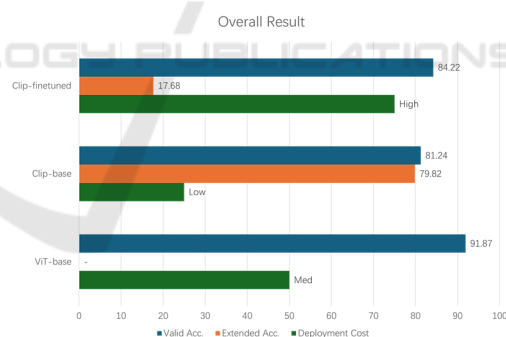


Figure 3: Overall experiment results (Photo/Picture credit: Original).

3.1 Test Accuracy

For test accuracy, the fine-tuned ViT model gained the best result, at 91.87%. Fine-tuned CLIP model is about 3% better than clip base, but still worse than fine-tuned ViT model. The base CLIP model gained an accuracy of 81.24%, which is a promising result considering it does not require any training and work out-of-the-box. The detailed test result is shown in Table 3.

Table 3: Test accuracy.

Model	Test Accuracy (%)
ViT-base (fine-tuned)	91.87
CLIP-base	81.24
CLIP-base (fine-tuned)	84.22

3.2 Extended Accuracy

Fine-tuned CLIP model has a much lower accuracy on unseen categories, of 17.68%. As a contrast, base CLIP model's accuracy is 79.82%, basically equal to the test accuracy of CLIP model. This shows that clip model lost most of its generalize and zero-shot ability while fine-tuning and have a much worse ability when facing unseen categories. Fine-tuned ViT model doesn't have any zero-shot ability, show it isn't included in this test. The result of extended accuracy is as shown in Table 4.

Table 4: Extended accuracy.

Model	Extended Accuracy (%)
ViT-base (fine-tuned)	-
CLIP-base	79.82
CLIP-base (fine-tuned)	17.68

3.3 Deployment Cost

The base CLIP model doesn't require any extra training or fine-tuning, so its training cost is the lowest. Both ViT and fine-tuned CLIP require extra training, however, CLIP model is much harder to be trained, compared to ViT. CLIP is more difficult to converge during training than ViT, as shown in Figure 4. Choosing a correct hyperparameter for CLIP requires testing and an incorrect fine-tune might result in worse accuracy than the base model. The computing cost of fine-tuning CLIP is also higher than fine-tuning ViT. So, the deployment cost of Fine-tuned CLIP is higher than fine-tuned ViT. The result of deployment cost is shown as in Table 5.

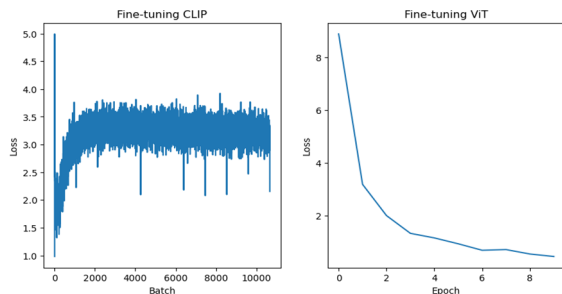


Figure 4: Training loss while fine-tuning two models (Photo/Picture credit: Original).

Table 5: Deployment cost.

Model	Deployment Cost
ViT-base (fine-tuned)	Medium
CLIP-base	Low
CLIP-base (fine-tuned)	High

4 CONCLUSIONS

This study proposes a new object detection based on transformer modelling. In addition, this paper sets up a bidirectional matching loss for prediction. The model contains a Resnet-101 model as a backbone, an encoder part with an attention mechanism, a decoder with an object query input, and a feedforward network. The loss function is a two-step set prediction loss carefully designed for object detection. In addition, migration learning techniques are invoked to demonstrate the effectiveness of improving model performance through two baseline object detection datasets. The paper then conducts various experiments to analyse the performance of the model on these two datasets. The authors implement Faster R-CNN model for comparison. On both datasets, the transformer model outperforms the Faster R-CNN and has a higher AP by 3.0. Meanwhile, the transformer trained on the PASCLA VOC maintains AP of 68.8, which is significantly higher than that of COCO 2007. The effectiveness of transfer learning is well demonstrated. This redesigned approach to the detection system presents a number of challenges, particularly in the areas of training, optimization, and small-object performance. Previous detection models have been improved over the years to address similar problems. In the future, semantic segmentation tasks for transformers will be considered as the next phase of research.

REFERENCES

- Bossard, L., Matthieu, G., and Luc, Van, G., (2014). Food-101—mining discriminative components with random forests. *Computer Vision—ECCV 2014: 13th European Conference*.
- Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.
- He., Kaiming., et al. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*.

- He., Kaiming., et al. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.
- Loshchilov, I., and Frank, H., (2017). Decoupled weight decay regularization. arXiv:1711.05101.
- Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. International conference on machine learning.
- Rawat, W., and Zenghui W., (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, vol.29(9), pp: 2352-2449.
- Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems.
- Wolf, T., et al. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations.
- Xiaoyi, D., et al. (2022). Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. arXiv:2212.06138.

