Image Classification by ResNet50 and ViT

Caiwu Chen
Columbia University
cc4786@columbia.edu

Abstract

We primarily evaluate two image classification methodologies, Deep Residual Network-50 and Vision Transformer, by retraining the model and comparing accuracy and loss. We also introduce transfer learning and apply fine tune techniques, data augmentation and dropout and batch normalization, attempting to achieve satisfying results. We surprisingly find out that ResNet50 have better performance than ViT in our small low resolution images.

1 Introduction

Multi-label classification has been a pivotal task in deep learning and machine learning. In classification, a number of classified training examples will be provided for training. Models will able to predict the labels of test data. This research lies in the meticulous examination of two prominent models: ResNet50 and Vision Transformer (ViT). Our primary focus is on evaluating the accuracy of predictions within super-classes and sub-classes. With the overarching goal of enhancing the understanding of multi-label classification capabilities, this study explores the intricate patterns learned by ResNet50 and the unique self-attention mechanisms employed by ViT. By utilizing a curated dataset tailored for multi-label image classification, we employ data processing techniques and embark on a comparative analysis.

2 Related Work

2.1 Deep Residual Networks

Residual Networks, short for ResNet, introduced by Kaiming He et al. in 2015, employs a residual learning framework, allowing the training of extremely deep networks with hundreds of layers. The key innovation is the introduction of shortcut connections or skip connections, enabling the flow of information directly across layers. ResNet50 is a specific variant of ResNet that consists of 50 layers, including residual blocks. It has become widely adopted in computer vision tasks, demonstrating superior performance and ease of training compared to earlier architectures.

2.2 Vision Transformer

Vision transformer short for ViT, proposed by Dosovitskiy et al. in 2020. ViT employs the transformer architecture, originally designed for natural language processing. The model divides an image into non-overlapping fixed-size patches, embeds them, treating the image as a sequence of tokens, and leverages self-attention mechanisms to capture complex relationships between patches, allowing for excellent performance in various computer vision tasks.

3 Method

3.1 ResNet50

ResNet50 is used to process the image data and firstly predict the label of image (He et al. 2015). Differing from the traditional Convolution Neural Network, ResNet introduces residual block into the the architecture. Residual learning is the activation of a layer is fast-forwarded to a deeper layer in the neural network.

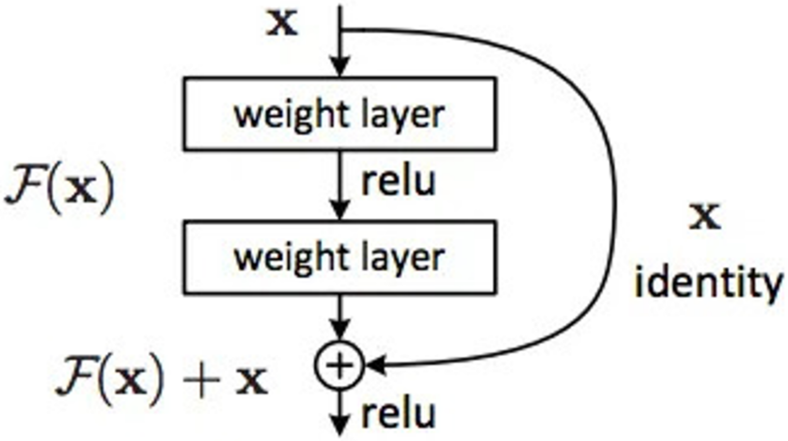
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 1: Residual Block. Source [1]

The ResNet50 uses a bottleneck design for the building block. A bottleneck residual block uses 1×1 convolutions, known as a "bottleneck", which reduces the number of parameters and matrix multiplications. This enables much faster training of each layer. It uses a stack of three layers rather than two layers. ResNet50, just as its name, has up to this point the network has 50 layers.

Considering multi-label categorical situation this problem have, ResNet50 use Sparse Categorical Cross Entropy (SCCE) loss funtion, defines with the following formula:

$$L_{SCCE} = -\sum_{i=1}^{N} y_i \cdot \log \hat{y}_i$$

The final predictions after the dense layer using softmax instead of previous relu activation function to correspond to the number of classes in classification task:

$$\hat{y}_i = \frac{e^{(W_i \cdot x + b_i)}}{\sum_{j=1}^{NumberofClasses} e^{(W_j \cdot x + b_j)}}$$

After that we are able to determine the image label associated to the greatest possibility of the predictions.

3.2 Vision Transformer (ViT)

Compared to ResNet50, ViT splits input images into a sequence of flattened 2D patches and maps them to a fixed latent vector size using a linear projection (Dosovitskiy et al. 2020). A learnable embedding is added to the sequence of embedded patches and Position embeddings are incorporated to retain positional information.

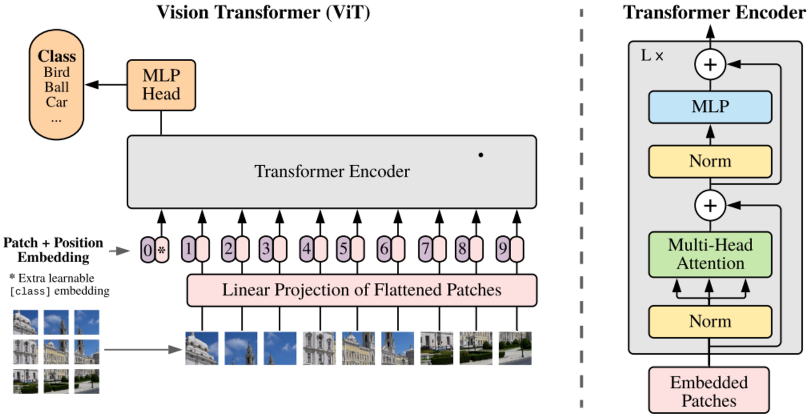

Figure 2: Model overview. Source [2]

Images Split When patch embedding, the number of patches are tunable.

$$z_0 = \left[x_{\text{class}}; x_1^{PE}; x_2^{PE}; \ldots; x_N^{PE}\right] + E_{\text{pos}}, \quad E \in \mathbb{R}^{(P^2 \cdot C) \times D}, \quad E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$$

Multi-Head Self Attention (MSA) The Transformer encoder (Vaswani et al., 2017) consists of alternating layers of multiheaded selfattention (MSA)

$$z_0' = \text{MSA}(\text{LN}(z_0^{‘-1})) + z_0^{‘-1}, \quad ‘ = 1, \ldots, L$$

MLP with GELU The Transformer encoder (Vaswani et al., 2017) also consists of alternating layers of MLP blocks with a GELU non-linearity.

$$z_0^{‘} = \text{GELU}(\text{LN}(z_0^{‘})) + z_0^{‘}, \quad ‘ = 1, \ldots, L$$

$$\text{GELU}(x) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right)$$

Output Layer Layernorm (LN) is applied before every block, and residual connections after every block (Wang et al., 2019; Baevski & Auli, 2019).

$$y = \text{LN}(z_0^L)$$

Similarly, after that we are able to determine the image label associated to the greatest possibility of the predictions.

4 Experiments and Discussion

4.1 Transfer Learning

For image classifications, Transfer learning has advantages of saving resources, improving efficiency, saving time and other benefits. We self define a ResNet50 model and compare with the pre-defined ResNet50 model from Keras. We set dropout rate to 0.2 at this experiment.

	Self-defined Model	Pre-defined Model
Superclass Accuracy	0.4651	**0.5007**
Superclass Loss	3.6300	**1.6871**
Subclass Accuracy	0.0246	**0.0455**
Subclass Loss	6.4955	**5.5932**

To address potential overfitting, the superclass ran for 10 epochs, while the subclass ran for 15 epochs. The table reveals that the pre-trained ResNet50 model consistently outperforms the self-defined model, boasting higher test accuracy and lower losses under the same conditions. Subsequently, in the ensuing experiments, we leverage the superior performance of the pre-trained model for further validation and analysis.

4.2 Data Augmentation

In our dataset, the test dataset is twice larger than train dataset (6323 training images and 12733 testing images). Data Augmentation technique was expected to benefit the classification tasks lacking sufficient data (Perez and Wang 2017). In our ResNet50 model, we preprocess training database by fine-tune images, applying rotation, shift, sheer, zoom and flip.

	Without Data Augmentation	With Data Augmentation
Superclass Accuracy	**0.5025**	0.5007
Superclass Loss	4.4438	**1.6871**
Subclass Accuracy	0.0407	**0.0455**
Subclass Loss	9.0367	**5.5932**

Surprisingly, Data Augmentation yielded minimal improvement in accuracy, potentially due to the limited diversity in the training data. The homogeneity of our training images may have hindered the extension of our models' applications. However, there is a notable decrease in loss.

Additionally, we experimented by switching the training dataset and test dataset to enrich the variety and diversity of the training set. The results indicate that when testing on the training super class, accuracy reaches 0.6966, and when testing on the training sub class, accuracy reaches 0.0742. While not comparable to the original train-test accuracy, this outcome demonstrates the potential for enhancing model performance through dataset variety and diversity.

4.3 Dropout

We also introduce Dropout technique in our models to reduce overfitting and increase the robustness (Hinton 2014). This experiment is more focus on ResNet50. We fine tune dropout rate to testify the performance of ResNet50 under three settings: 0.2, 0.5 and 0.8.

Dropout Probability	0.2	0.5	0.8
Superclass Accuracy	**0.5025**	0.4991	0.4184
Subclass Accuracy	0.0407	**0.0410**	0.0123

The results indicate that there is no significant difference in accuracy between dropout probabilities at 0.2 and 0.5. However, as the rate increases to 0.8, there is a substantial decrease in accuracy. This suggests that a higher dropout rate negatively impacts the model's ability to generalize and perform well on both super- and sub- classification tasks.

4.4 Compared To ViT

Besides those fine tune techniques on ResNet50, we also explore the disparities on model. ViT also train based on augmented training dataset and the dropout rate at 0.5.

Model	ViT	ResNet50
Superclass Accuracy	0.4979	**0.4991**
Superclass Loss	**1.4830**	1.6871
Subclass Accuracy	0.0027	**0.0410**
Subclass Loss	9.4922	**5.5932**

Surprisingly, the results reveal that ViT does not outperform ResNet50 in our database, a relatively small and low resolution images. This could be attributed to ViT requiring more extensive training to capture intricate patterns in high-resolution images. The limited examples in our dataset may not fully reveal the potential gap between ResNet50 and ViT. Additionally, although using bicubic interpolation (R. Keys 1981) to increase ViT's resolution enhances accuracy and reduces loss, it substantially increases training time.

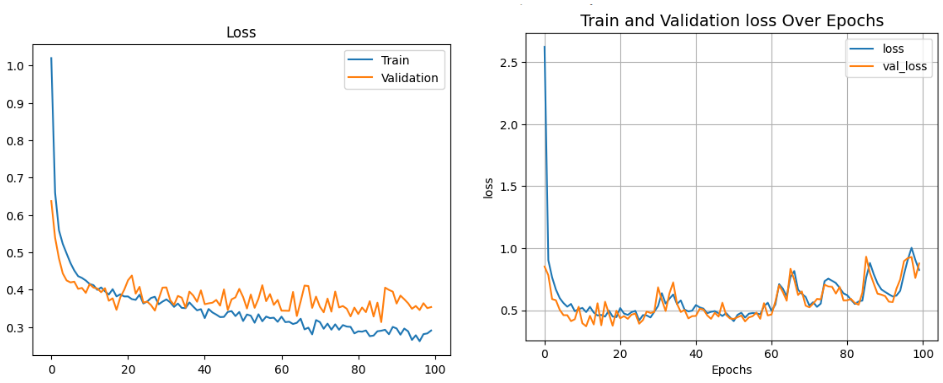

Figure 3: Loss Plot of ResNet50 and ViT

The loss plot further differentiates ResNet50 and ViT. Over 100 epochs, ResNet50 exhibits a consistent decreasing trend on both the training and validation datasets. In contrast, ViT's loss plot reaches its global minimum around the twentieth epoch, experiencing a sudden spike, and then an increasing loss trend after 50 epochs, displaying a sign of overfitting (Figure 3).

5 Conclusion

This project investigates the structures and workings of ResNet50 and Vision Transformer (ViT) in image classification. Our primary focus is on ResNet50, comparing pre-trained and self-defined models while applying diverse data processing techniques. Results are then compared with ViT, indicating both models achieving a 50% accuracy in super label classification. However, in sub-label classification, ResNet50 reaching a substantial 6% accuracy, surpassing ViT's lower than 1%. Despite their competency, the overall accuracy prompts contemplation on their suitability for small, low-resolution image classification tasks. In conclusion, this examination emphasizes ResNet50's functionality, providing practical insights and consideration for nuanced image classification scenarios.

References

[1] 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 (CVPR 2016), pp. 770-778.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), pp. 200-215.

[3] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In ACL, 2019.

[4] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In ICLR, 2019.

[5] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, Qing He. "A Comprehensive Survey on Transfer Learning." In Proceedings of the NeurIPS, 2019, pp. 43-76. URL: https://api.semanticscholar.org/CorpusID:207847753.

[6] Perez, L., & Wang, J. (2019). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), (pp. 123-145).

[7] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2014). Improving neural networks by preventing co-adaptation of feature detectors. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), (Vol. 28, No. 3, pp. 1058-1066).

[8] Keys, R. (1981). Cubic convolution interpolation for digital image processing. In: Advances in Neural Information Processing Systems (NeurIPS), 29 (6), 1153–1160. doi:10.1109/TASSP.1981.1163711.