Vision Transformer
Group 6 Changhong Zhang

In the previous work, we have used basic CNN model and Resnet model to do this classification work. However, a brand-new framework name "Transformer" is created and developed overwhelmingly among NLP area. After that, some researchers are trying to use the model to do some CV task with as less changes as possible. The model we used in the project is named "Vision Transformer", which is really similar to the Bert model (one of the latest Transformer models in NLP area).

1. Main idea of Vision Transformer model – Attention Mechanism
   In fact, the Vit model use attention mechanism to capture the characteristic of image.

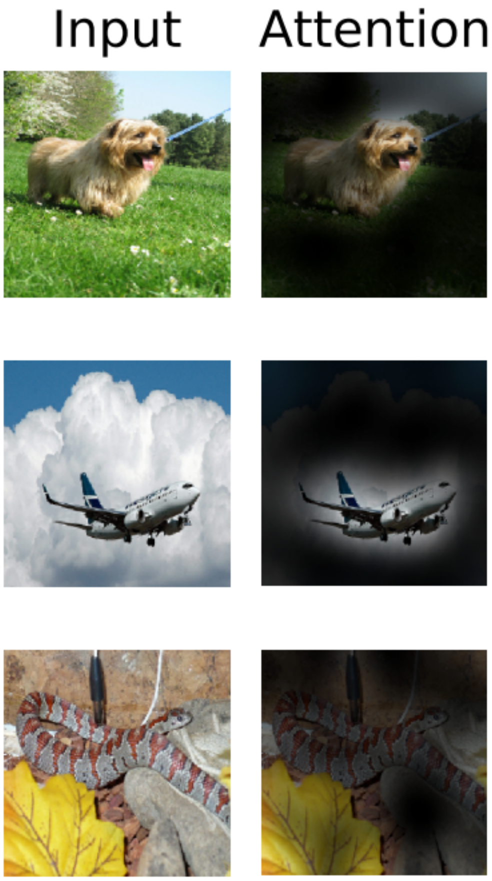


Figure 1: Representative examples of attention from the output token to the input space

Basically, attention is weights for the pixel of image in CV area, which is similar the attention in NLP area. But the difficulty for using attention is that we can't directly change the image from a high dimension matrix into a vector for computing, because there is a limit for the memory (RAM) of Computer now. So, some researchers divided the image into patches, and this patches can be seen as words in NLP, which overcome the computing limits problems. [1]

2. Structure/architecture of transformer

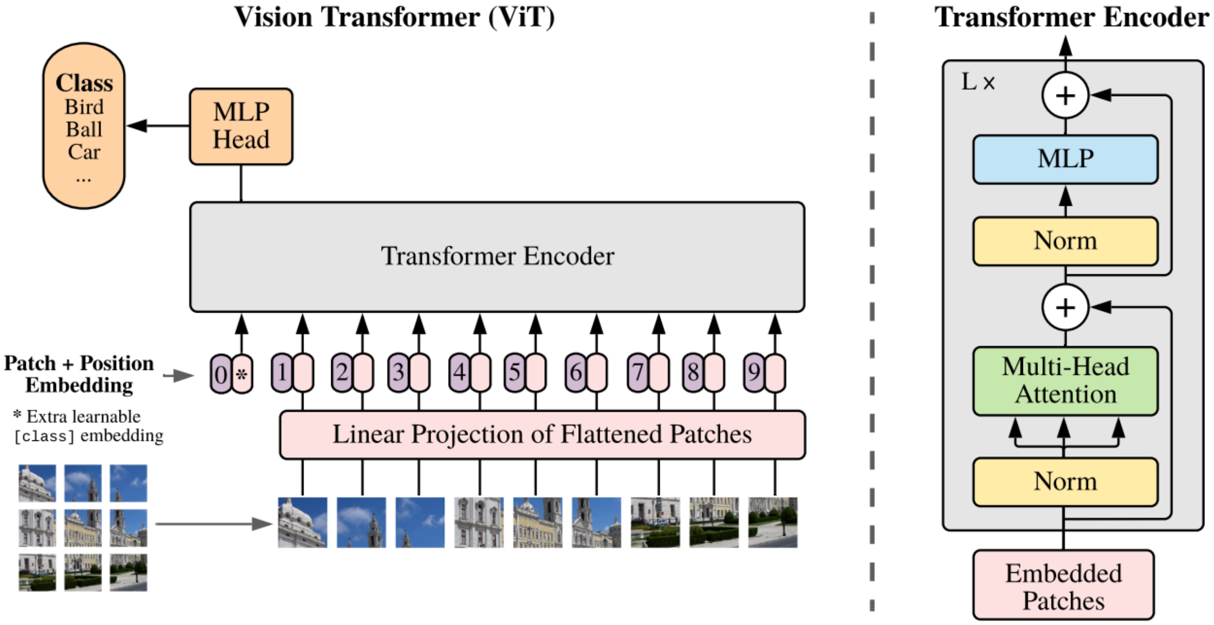


Figure 2: Model overview.

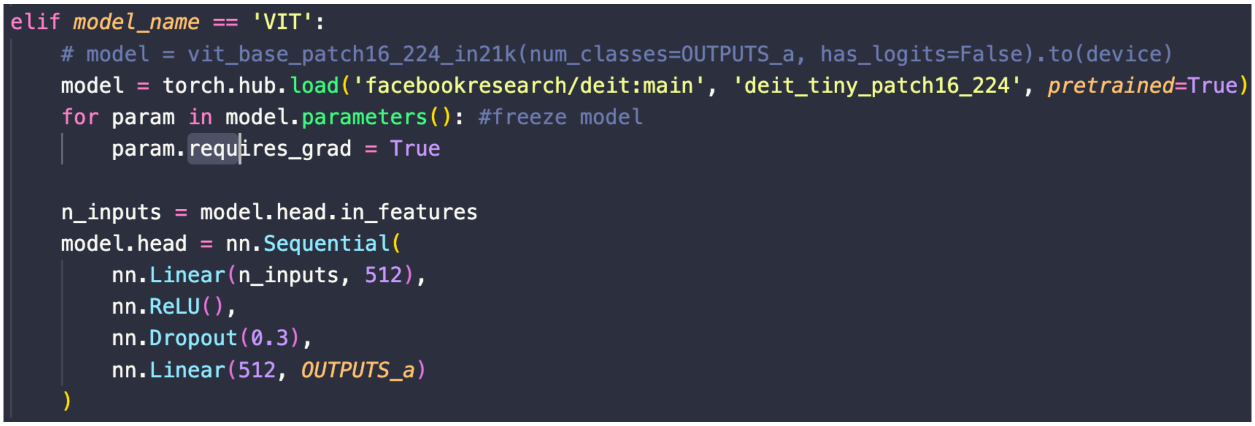
$$
\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, && \mathbf{E}\in\mathbb{R}^{(P^2\cdot C)\times D},\ \mathbf{E}_{pos}\in\mathbb{R}^{(N+1)\times D} && (1)\\
\mathbf{z'}_\ell &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1\ldots L && (2)\\
\mathbf{z}_\ell &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, && \ell = 1\ldots L && (3)\\
\mathbf{y} &= \mathrm{LN}(\mathbf{z}_L^0) && && (4)
\end{aligned}
$$

We split an image into fixed-size patches (for this model is 16x16 size), linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. (Which makes the model as much as similar to the transformer model we used in NLP area)

3. How we train the model

a. Input size: We use the 224x224 size as the input size, which is also the original size of our data. Since the patches size is fixed to 16x16, when we implement position embedding process, the image must keep the same input size as the pretrained model, otherwise the position information of the pretrained model will be affected.

b. Training based on pretrained model
Like using pretrained model to train Resnet model, we only unfreeze the final layer of the VIT model (MLP Head layers). In this part, we passed the final outputs of Transformer Encoder layer to a fully connected layer with 512 neural, and use ReLu function as the activation function. After that, we did a dropout process at 0.3 level to improve the model's robustness. Finally, we passed the vectors to a fully connected layers, which neural number is the number of the classification categories of our data set (176 kinds of tree leaves).



```
elif model_name == 'VIT':
    # model = vit_base_patch16_224_in21k(num_classes=OUTPUTS_a, has_logits=False).to(device)
    model = torch.hub.load('facebookresearch/deit:main', 'deit_tiny_patch16_224', pretrained=True)
    for param in model.parameters(): #freeze model
        param.requires_grad = True

    n_inputs = model.head.in_features
    model.head = nn.Sequential(
        nn.Linear(n_inputs, 512),
        nn.ReLU(),
        nn.Dropout(0.3),
        nn.Linear(512, OUTPUTS_a)
    )
```

Figure 3: The Vision Transformer definition part of our code

4. Why we didn't perform better than Resnet model

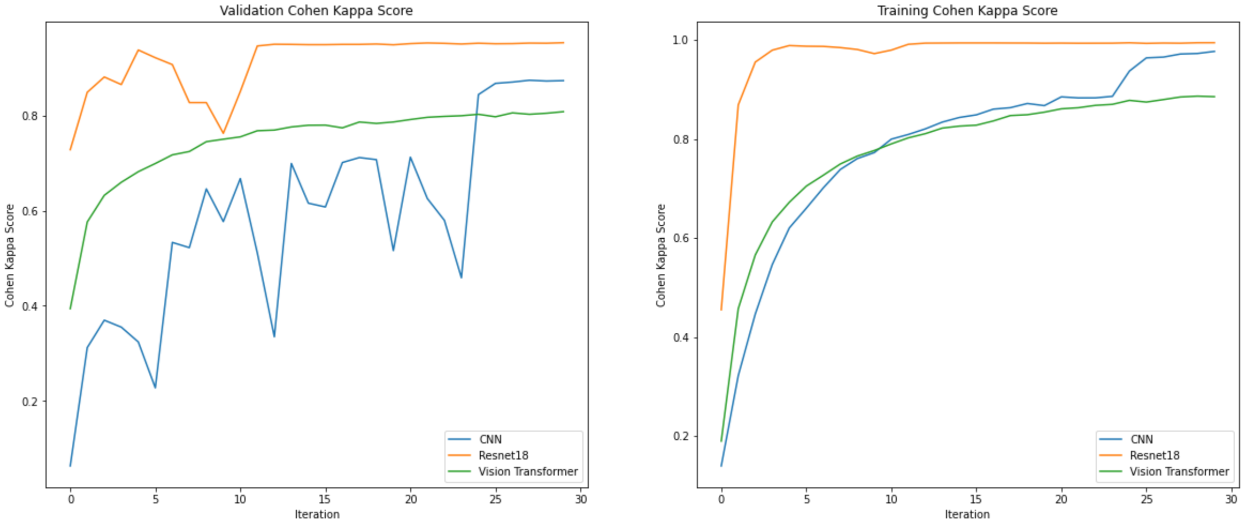


Figure 5: The accuracy of models in the tree leaves data set classification task

From figure 5, we can easily find that Vision transformer model didn't perform better than Resnet, even worse than CNN model. However, the latest ranking of models in image classification area shows that 8 of the top 10 models is using transformer architecture. So why this happened?

If we ignore the hyperparameter tuning temporally (our model results had converged), the main reason Vit model didn't achieve a better score is that, the data set size is too small to offer enough data to train the model. As the paper of Vit model said, Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data. If we can get more data, we may train a better Vit model.

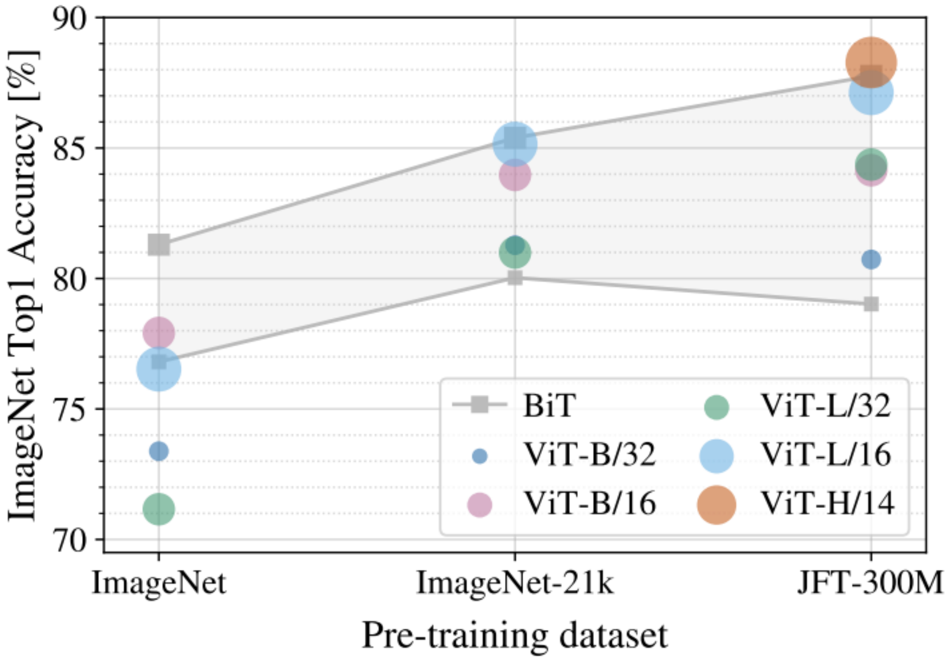


Figure 5: Transfer to ImageNet.

While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.