RecVis Assignment 3: ViT for Bird Classification

Rajae Sebai

rajae.sebai@ens-paris-saclay.fr

Abstract

In this paper, we make use of Vision Transformers to perform classification on the provided Bird dataset. We finetune a pretrained ViT model on our augmented dataset and compare the result to a simplified AlexNet baseline. We also attempt to visualize and interpret the self-attention on the last layer of the model.

1. Introduction

In this work, we are going to make use of the Vision Transformers (ViT) encoder architecture introduced in [1] to perform image classification. Since the Transformers [2] are known to have a very large number of parameters compared to more inductive biased models such as CNNs, we will use transfer learning from a pre-trained model to perform classification on our dataset. Note that, if we compare it to the novel large scale datasets on which Transformers based models are trained, our dataset turns out to be extremely small. It is very unlikely to get fair results with Transformers by training the model from scratch. In fact, the Vision Transformer (ViT) model was pre-trained on the large ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 384x384. We will fine-tune this model on our data by adding a classification head to the 12 ViT encoders.

2. The ViT model

We will briefly recall how the ViT model works. Just like in natural language models (with BERT for example), an image is considered as a sentence and presented to the model as a sequence of tokens or patches. These patches are fixed-size (16x16) parts of the original image (resized to 224x224). We then end up with 14x14 patches entries for each image. The authors in ViT suggest adding a class [CLS] token to the beginning of the sequence. After training the encoders to "understand" the image representations, the [CLS] token can be fed to a classification head to perform the classification task.

3. Data Pre-Processing

The images in our training (1082 images) and validation (103 images) datasets are resized to the same resolution (224x224) and normalized across the RGB channels with mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5). In addition, we eventually use random resized crop, rotation (±20 degrees) and horizontal flip for the training data and a center crop for the validation data.

4. Results

We get the following results when fine-tuning the ViT model on our Bird dataset using eventually data augmentation and comparing to a baseline trained from scratch model.

Model	Test Accuracy
Simplified Alexnet	0.26
Fine-tuned ViT	**0.82**
Fine-tuned ViT + Aug	0.78

At the moment, the fine-tuned model using only the resized+normalized data has the best performance. Further research is being done to include more bird images to retrain the encoder part of the model in an unsupervised way.

5. Visualizing the attention

In this section, we are interested in visualizing the attention matrix row belonging to the classification token [CLS], i.e, the dot product of the [CLS] with each of the other 14x14 tokens. This way, we would get information about the part of the image (patch) the network is

focusing on. The following Figure illustrates this visualization on two images. We show the mean across the 12 self-attention heads. We can see that, on average, the parts with relatively high values correspond to parts of interest in the images.

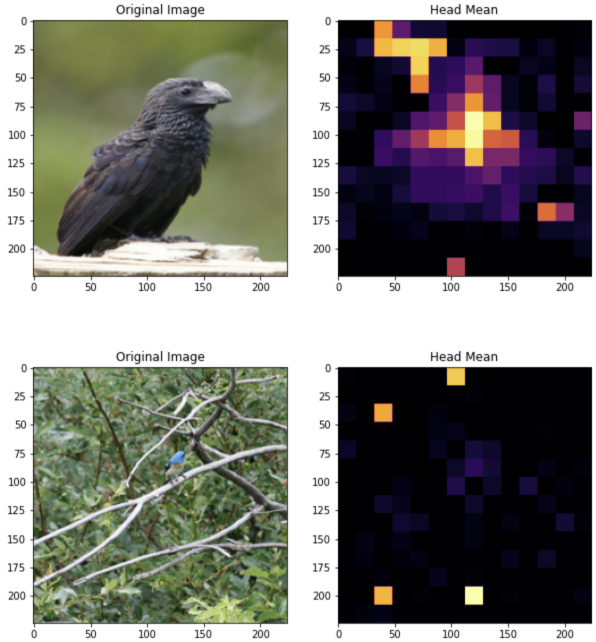

Figure 1 : The dot product of the [CLS] self-attention with each of the other 14x14 tokens', averaged over the 12 heads of the last layer.

6. References

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2002.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.