Project Proposal - ECE 176

Jayla Cho
Computer Science & Engineering
A18058562

Keyi Chen
Mathematics
A16870360

Abstract

This project aims to implement and improve the approach presented in "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"[1]. Transformer architecture was traditionally used in NLP, but is now leveraging for large-scale image recognition tasks. Our project targets three critical areas for enhancement: Patch Size and Structure Optimization, Transformer Architecture Improvements, and Regularization and Training Techniques. Adapting Patch Size and Structure beyond 16x16 pixels could improve model sensitivity to crucial features. In addition, our project will refine the Transformer architecture itself to better capture the intricacies of image data. Finally, we expect Regularization and advancing training techniques to help improve model robustness.

1 Problem Definition

1.1 About the paper

The core problem addressed by the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" is to explore and demonstrate the applicability of Transformer models, which was used traditionally in NLP, to the field of image recognition.

The Vision Transformer starts by converting 2D images to 2D flattened patches which is treated as a token similar to words in NLP. Each flattened patch is then linearly projected into D-dimensional e,bedding space to create patch embeddings. Then a learnable embedding, referred to as the class token, is prepended to the sequence of patch embeddings. The Transformer encoder consists of alternating layers of Multiheaded Self-Attention (MSA) and Multilayer Perceptrons (MLP), with Layer Normalization (LN) applied before each block and residual connections after each block. The Vision Transformer incorporates less image-specific inductive bias compared to CNNs. As an alternative to using raw image patches, ViT uses feature maps from a CNN as the input sequence. ViT models are typically pre-trained on large datasets and fine-tuned on smaller, task-specific datasets.

1.2 Improving the paper

Applying different patch size and structure to ViT is expected to make the model more adaptive and potentially more effective, allowing it to pay more attention to important details in the image. Smaller patches could be used for areas with high detail to capture finer features while larger patches could be used for homogeneous areas to reduce computation and focus on broader patterns.

Improving the Transformer architecture within the context of ViT can be done by various ways. For instance, mplementing more efficient variants of the self-attention mechanisms such as Sparse Attention can reduce the computational complexity to linear with respect to the input sequence length.

Regularization and Training Strategy enhancements can help the model learn more robust and generalized features. Reducing overfitting, performing consistently are expected for the outcome.

2 Tentative Method

Detailed Structure:

Our tentative method involves the use of Vision Transformers (ViT) for image classification.

The process begins with resizing the input image to a fixed resolution, followed by dividing it into a grid of small, non-overlapping patches. These patches are then linearly embedded, and positional embeddings are added to retain spatial information. This step transforms the 2D image into a sequence of 1D tokens, akin to words in a sentence, which are then ready to be processed by the Transformer encoder.

(NB: For a hybrid architecture, we can take the input sequence can be formed from feature maps of a CNN as an alternative to raw image patches.)

The transformer encoder consists of alternating layers of multi-headed self-attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Layernorm (LN) is applied before every block, and residual connections after every block :

- Layer Normalization (LN)
- Multi-Headed Self-Attention (MSA)
- Residual connections
- Layer Normalization (LN)
- Multi-Layer Perceptron (MLP)
- Residual connections
- Feed-Forward Network: consists of two linear transformations with a non-linearity in between.

After passing through the Transformer encoder, the state of the [CLS] token is used as the representation of the entire image. This representation is passed through a simple feed-forward network (often just a linear layer) to produce the final class predictions.

Reason:

We chose ViT due to its proven effectiveness in achieving state-of-the-art results on various image recognition benchmarks with less computational resources compared to CNNs. The Vision Transformer has much less image-specific inductive bias than CNNs.

Its ability to handle sequences of image patches as input makes it a promising approach for complex image classification tasks.

Strengths:

The main strength of ViT is that 1) the scalable NLP Transformer architectures and their efficient implementations can be easily used. 2) it has much less image-specific inductive bias than CNNs, enabling it to learn more flexible representations and potentially leading to improved performance on diverse image classification tasks.

3 Experiments

Datasets:

- *Stanford Cars*: A dataset consists of 196 classes of cars with a total of 16,185 images, taken from the rear.

  Data format: The data is divided into almost a 50-50 train/test split with 8,144 training images and 8,041 testing images. The images are 360×240.

  Purpose: To evaluate the performance of ViT on a widely recognized image classification benchmark.

Experiments: We plan to conduct a series of experiments to compare the performance of our re-implemented ViT and the improved ViT with some new techniques. These experiments will include:

- Training ViT and fine-tuning the pre-trained models on Stanford Cars Dataset.
- Evaluating classification accuracy, model efficiency, and robustness to image variations.
- Analyzing the impact of different image patch sizes, Transformer configurations and regularization on performance.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.