ViViT: A Video Vision Transformer

Paper Review

Dhananjay Agnihotri **Ninad Chaphekar**
200100055 **200010016**

Motivation

Looking at the success of self-attention based transformer models in NLP and image classification, the authors try to use the advantages of such attention-based models for videos. Inspired by ViT, which is a pure transformer model, the authors extended the vision transformer for videos so that they could capture the temporal correlation between frames most effectively.

Novelties and Major Contribution

- **Pure-transformer based model for video classification** - Leveraging the latest success of self attention models, the author introduces a pure transformer based model for video classification.
- **Introduction of spatio-temporal tokens:** The model introduced the concept of extracting spatio-temporal tokens from the input video, which are then encoded by a series of transformer layers.
- **Proposal of efficient variants:** The authors proposed several efficient variants of their model which factorize the spatial- and temporal-dimensions of the input. This is a significant contribution as it addresses the challenge of handling long sequences of tokens encountered in video.
- **Efficient variants:** In order to handle the long sequences of tokens encountered in video, the authors propose several efficient variants of their model which factorize the spatial- and temporal-dimensions of the input.
- **Effective regularization:** Although transformer-based models are known to only be effective when large training datasets are available, the authors show how they can effectively regularize the model during training and leverage pretrained image models to be able to train on comparatively small datasets.
- **State-of-the-art results:** The paper achieves state-of-the-art results on multiple video classification benchmarks including Kinetics 400 and 600, Epic Kitchens, Something-Something v2 and Moments in Time, outperforming prior methods based on deep 3D convolutional networks.

Critical Analysis

- **Spatio-Temporal Attention** model forwards all the tokens extracted from the video to the encoder. Each layer models all pairwise interactions between all spatio-temporal tokens and thus the Multi Headed Self Attention has quadratic complexity. Which makes it computationally expensive.
- In the **Factorized self-attention** model, Instead of having MSA compute across all pairs of tokens, it factorizes it and first only computes self-attention spatially and then computes temporally. The Dividing of Spatial and Temporal Attention might result in loss of certain key information
- **Small Dataset size -** Even though they have used pretrain models, the big transformer architecture model might overfit to the small dataset used.