ViViT: A Video Vision Transformer

Sarvesh Gharat and Raj Gothi

Motivation

Inspired from success of attention-based models in NLP, a good number of successful attempts were made in integrating CNNs with transformer and replacing them completely as well. Motivated from these successful results and the core advantage of transformers to model the long term dependencies, in this study, the authors make an attempt to develop a transformer based model for video classification. As self-attention is primarily computed using spatio-temporal tokens that maybe present in a large number in a video, the authors also present several methods for factorising the model along spatial and temporal dimensions to increase scalibility and efficiency of the model

Novelties

This is one of the first attempts in solving a video classification problem using a pure transformer-based architecture. The authors also provide multiple methods of factorising the model along spatial and temporal dimensions along with showing how to regularise the model during training for smaller datasets. As this is one of the fist attempts, the authors determine the best design choices for these architectures. To do this, the authors conduct a thorough study on regularisation methods, tokenisation strategies and model architecture.

Major Contributions

There are three major contributions of this study. The former includes the proposal of a couple of algorithms for mapping a video to a sequence of tokens. Next, the authors discuss multiple transformer-based architectures starting from a basic extension of ViT to developing more efficient models that factorises the temporal and spatial dimensions of the input video at various levels of the architecture. Finally last but not the least, the authors comment on how to leverage pretrained models to effectively use transformers on a small scale dataset.

Critical Analysis

The authors provide a novel transformer based architecture for video classification. Though the results are satisfying, due to lack of literature it's really a question as to whether this is a vanilla model or really a state of the art algorithm. Besides that, there also happens to be some questions on capability of the architecture to domain adaptation. It would have been great and would have helped generalise a video classifier incase the authors would have compared a similar results. The another question that remains unanswered is the reliability of the encoder for various tasks such as video segmentation and hence again keeping an existing question on generalisation to different tasks. Overall the work done in manuscript is great and does open a field of exploration through transformers in video regime, however it still remains a question about reliability of architecture in real word scenario and for other video based problems.