ViViTT: A Video Vision Transformers Paper Review

Ashutosh - 200100037 Utsav - 200100054

Motivation

The field of NLP has recently bloomed because of self-attention transformer models and inspired approaches in Computer Vision to integrate transformers into CNN. ViT, an image transformer model outperformed state-of-art CNN models in image classification. ViViT aims to extend ViT for videos to capture the spatio-temporal correlation between frames & develop state-of-art transformer-based models for video classification.

Novelties & Major contributions

This paper introduced the concept of factorised spatial and temporal encoders. One major key difference between the 3D CNNs and the ViViT is that the 3D CNNs are designed to work for grid like data and learn the spatio-temporal dependencies from it, where as the ViViT can even work on the non-grid data and can learn the spatio-temporal as well as temporal dependencies from it. This means that ViViT can capture not only the motion and appearance of objects but also the relationships between events and their ordering within a video. In CNNs the receptive field grows linearly with the number of layers whereas each transformer layer models all pairwise interactions between all spatio-temporal tokens, and it thus models long-range interactions across the video from the first layer.
ViViT introduces tubulet encoding , a novel technique which fuses spatio-temporal information into tubelets (i.e.tokens), in contrast to simple uniform frame sampling, where the temporal change information is missing.
The paper proposes multiple transformers based architectures, of which the first is the Spatio-temporal attention which is a straightforward extension of ViT. Another advancement introduced is the Factorised Encoder. Here, initially all the spatial tokens of the same temporal index are passed through a Spatial Transformer Encoder and then the output of these Encoder are encoded together and send to a Temporal Transformer Encoder to model interactions between tokens from different temporal index. The third architecture, Factorized Self Attention, enhances computational efficiency by initially computing self-attention separately for tokens at the same temporal or spatial positions, allowing for more transformer layers despite reduced computational complexity. For factorised models overall computation is dominated by the initial Spatial Transformer. As a result, the total number of FLOPs for the number of temporal views required to achieve maximum accuracy is constant across the models

Critical analysis

A significant drawback of Multi-Headed Self Attention is its quadratic complexity concerning the number of tokens, and this becomes particularly pertinent in the context of video processing. As the number of tokens scales linearly with the number of input frames in a video, the computational demands can quickly become overwhelming. Another crucial observation is that the non-factorized model tends to overfit when trained on smaller datasets, necessitating the application of additional regularization techniques. Interestingly, this overfitting issue is mitigated in the factorized models, highlighting their practical advantages. Furthermore, it's worth noting that this paper heavily leans on the utilization of pre-trained image classifiers.