Paper review
ViViT: A Video Vision Transformer

Motivation

The Video Vision Transformer architecture is inspired from the vision transformer model (with a pure transformer based architecture), which outperformed its CNN counterparts in image classification. Various attention based architectures have been proposed in the paper as attention mechanism allows the modeling of long-range contextual relationships in video, whereas in CNNs the focus is mainly on local context and the receptive field grows linearly with the number of layers.

Novelties

The authors consider two embedding methods to map a video to a sequence of tokens.

1) Uniform frame sampling: sample frames from video at uniform time stamps and embed each 2D frame independently using methods as in ViT, and then concatenate these tokens together.
2) Tubelet embedding: It fuses temporal-spatial information together by extracting tokens from temporal, height and width dimensions, and then projecting them linearly.

The authors propose multiple transformer based architectures:

1) Spatio Temporal Attention: Forwards the spatio-temporal tokens through the transformer encoder; each layer models all pairwise interactions between spatio-temporal tokens and so models long-range interactions. However, complexity is $O((n_t.n_h.n_w)^2)$.
2) Factorized encoder: Consists of two separate encoders, spatial and temporal. Corresponds to the late fusion of temporal information and is analogous to CNN architectures. Complexity is $O((n_h.n_w)^2 + n_t^2)$.
3) Factorised self attention: Computes self-attention spatially first and then temporally; Has same number of transformer layers as model 1.
4) Factorized dot-product attention: For half of the attention heads, only tokens from spatial dimensions are considered and for the rest, temporal dimension. The outputs of multiple heads are combined by concatenating them and using a linear projection.

To counter the requirement of large amounts of data for efficient training of transformers, the ViViT model initializes its parameters from the pretrained image models.

Major contributions

The results obtained from the paper are compared with the state of the art results on multiple datasets including Kinetics, Moments in Time and more. The ViViT has outperformed all the other models on most of these datasets. The paper has provided 4 different transformer models for video classification with some differences which all can be analyzed in future work to build models with more complex tasks than video classification.

Critical analysis

Using the pretrained image models raise some questions because some image models might not be compatible with the ViViT model. Another problem with the architectures is that they require a very large amount of data for efficient training and a large amount of TFLOPS i.e. the amount of floating point computations.