Video Transformer Network

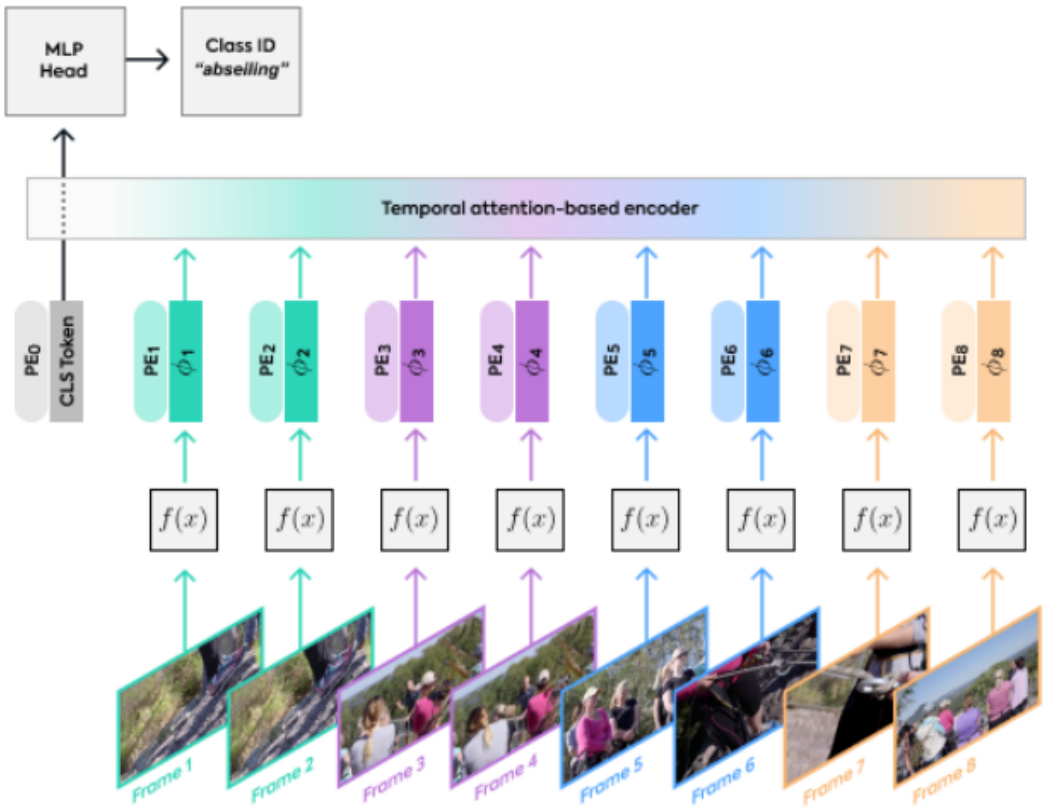

Figure 1. Video Transformer Network architecture. Connecting three modules: A 2D spatial backbone ($f(x)$), used for feature extraction. Followed by a temporal attention-based encoder (Longformer in this work), that uses the feature vectors (ϕ_i) combined with a position encoding. The $[CLS]$ token is processed by a classification MLP head to get the final class prediction.

Code and models are available at:
https://github.com/bomri/SlowFast/blob/ master/projects/vtn/README.md.

Архитектура VTN состоит из трех последовательных частей: 2D-модель извлечения пространственных признаков (spatial backbone), temporal attention-based encoder и MLP классификация.

Spatial backbone

[любая сеть для изображений]

The spatial backbone operates as a learned feature extraction module. It can be any network that works on 2D images, either deep or shallow, pre-trained or not,

convolutional- or transformers-based. And its weights can be fixed (pre-trained) or trained during the learning process.

Temporal attention-based encoder

We use a Transformer model architecture that applies attention mechanisms to make global dependencies in a sequence data. However, Transformers are limited by the number of tokens they can process at the same time. This limits their ability to process long inputs, such as videos, and incorporate connections between distant information.

In this work, we propose to process the entire video at once during inference. We use an efficient variant of self-attention, that is not all-pairwise, called Longformer [1]. Longformer operates using sliding window attention that enables a linear computation complexity. The sequence of feature vectors of dimension dbackbone (Sec. 3.1) is fed to the Longformer encoder. These vectors act as the 1D tokens embedding in the standard Transformer setup.

Like in BERT [8] we add a special classification token ([CLS]) in front of the features sequence. After propagating the sequence through the Longformer layers, we use the final state of the features related to this classification token as the final representation of the video and apply it to the given classification task head. Longformer also maintains global attention on that special [CLS] token.