Video Transformer Networks

GxLabs

Video Transformer Network

Abstract

- VTN, a transformer-based framework
- Classifies actions by attending to the entire video sequence information
- Whole video analysis, via a single end-to-end pass

Video Transformer Network

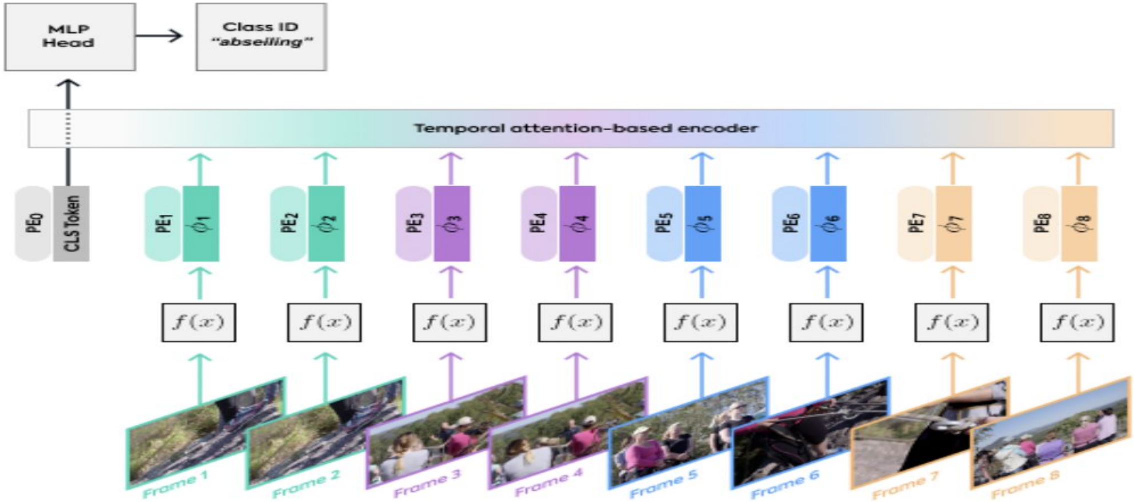

Figure 1. Video Transformer Network architecture. Connecting three modules: A 2D spatial backbone ($f(x)$), used for feature extraction. Followed by a temporal attention-based encoder (Longformer in this work), that uses the feature vectors (ϕ_i) combined with a position encoding. The $[CLS]$ token is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

Video Transformer Network

Spatial backbone

- Spatial backbone operates as a learned feature extraction module
- Any network on 2D images
  - either deep or shallow, pre-trained or not, convolutional- or transformers-based
  - weights can be fixed(pre-trained) or trained during the learning process

ViT-Base

- pre-trained on ImageNet-21K

ResNet50/101

- pre-trained on ImageNet

DeiT-B/BD/Ti

- pre-trained on ImageNet

Video Transformer Network

Temporal attention-based encoder

- Transformers are limited by the number of tokens they can process at same time
  - Limits their ability to process long inputs such as videos

- Longformer, process the entire video at once during inference
  - Operates using sliding window attention that enables a linear computation complexity
  - local-context self-attention + task-specific global attention

- Adding a special classification token [CLS]

Video Transformer Network

Classification MLP head

- Contains two linear layers
    - GELU activation function and Dropout between them

- Input token representation is first processed with a Layer normalization

Video Transformer Network

Inference methods

- Due to memory limitation, suggest several types of inference methods

1) Processing the entire video in an end-to-end manner

2) Processing the video frames in chunks, extracting features first, and then applying them to the temporal attention-based encoder

3) Extracting all frames’ features in advance and then feed them to the temporal encoder

Video Transformer Network

Conclusion

- a modular transformer-based framework for video recognition tasks
- Efficient way to evaluate videos at scale, both in terms of computational resources and wall runtime
- Current video classification benchmarks are not ideal for testing long-term video processing ability
  - When such Datasets become available, models like VTN will show even larger improvements compared to 3D ConvNets