Paper Critique – Is Space-Time Attention All You Need for Video Understanding?

Alessandro Folloni
Computer Science
afolloni@unc.edu

1. Research Problem

1.1. What research problem does the paper address?

The paper proposes a substantial new model to video classification task through a convolution-free approach based mainly on self-attention over space and time.

1.2. What is the motivation of the research work?

The main motivation is studying the effect of spatial and temporal attention in the domain of videos and finding a new model that could effectively work on long clips too.

2. Technical Novelty

2.1. What are the key technical challenges identified by the authors?

Classic video classification techniques used convolutional models although they are not explicitly suited for capturing local and global dependencies through a sequence of images. The new TimeSformer model aims at overtake these limitations by using the self-attention mechanism to obtain higher expressivity. Moreover, training deep CNNs remains very costly, especially for high-resolution and long videos even with new tech improvements in the computer architecture, resulting in a bad scalability.

2.2. How significant is the technical contribution of the paper? If you think that the paper is incremental, please provide references to the most similar work

The paper is influenced by other works that studied self-attention in combination or in complete substitution of convolutional models, but the new model gains its efficiency from the decomposition of the video into a sequence of frame-level patches with a subsequent feeding of a Transformer. Nevertheless, we can report some similar works that exploited self-attention as a substitute for convolutions: Parmar et al., 2018; Ramachandran et al., 2019; Cordennier et al., 2020; Zhao et al., 2020.

2.3. Identify 1-5 main strengths of the proposed approach.

- As we are trying to inference information from a video clip, it is essential to consider the temporal dimension. Attention mechanism effectively succeeds on this aspect.
- Transformers enjoy faster training and inference compared to the past CNNs models.

2.4. Identify 1-5 main weaknesses of the proposed approach.

- There is no sufficient previous knowledge about the task so it can be harder to make proper comparisons.

3. Empirical Results

3.1. Identify 1-5 key experimental results, and explain what they signify.

- Divided space-time attention can effectively work on higher spatial resolution and longer videos thanks to a sensibly inferior computational cost; this is even better considering its large learning capacity. Overall, this is a huge result for further analysis.
- The model can work on clips of 96 frames that highly surpasses the previous result obtained through the convolutional approach.
- Strong experimentation between the different variants, leading to top results for accuracy using the TimeSformer-L (long) version, state-of-the-art inference cost for the base version with a decent accuracy.
- TimeSformer-L doesn't require many clips to achieve optimal results, compared to this issue present for previous models.

3.2. Are there any weaknesses in the experimental section (i.e., unfair comparisons, missing ablations, etc)?

We can just highlight that pre-training is essential for this model, the training from scratch would be hard.

4. Summary

I really appreciated the paper because it provides a sensibly new approach to an existing problem leading to better results that could be farther investigated in the future. The paper discusses several variations on the approach but using the same notions. The architecture exploits Transformers and divided space-time attention for its results; in particular, we can notice optimal results regarding the fact that it can work on long clips and at a low inference cost. For these reasons, even if the actual results of the models are not state-of-the-art for every aspect, I strongly believe they can be the base for further analysis.

5. QA Prompt for a Paper Discussion

5.1. Discussion Question

The paper mentions that TimeSformer can work efficiently on long video clips. How does TimeSformer achieve this efficiency, and what are the benefits of being able to process long video sequences?

5.2. Your Answer

TimeSformer achieves efficiency by decomposing video clips into frame-level patches, which are then processed by a Transformer model. This decomposition reduces computational cost compared to traditional CNNs. Processing long video sequences is beneficial because it allows for capturing more context and temporal information, which is crucial for tasks like action recognition and video classification.