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

Dan Peng
Department of Computer Science
Aug 20, 2024
danpeng@unc.edu

1. Research Problem

1.1. What research problem does the paper address?

This paper addressed the problem of video classification by adapting the image model Vision Transformer architecture to video with ”divided attention”: temporal attention and spatial attention.

1.2. What is the motivation of the research work?

The motivation came from inherent limitations of CNNs in imposing less inductive biases and modeling long-range spatiotemporal dependencies for video analysis. Also, the self-attention in standard transformer is computationally costly due to the large number of patches in the video.

2. Technical Novelty

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

Key technical challenges are: scaling self-attention mechanisms for large video dataset; efficiently capturing long-range spatiotemporal dependencies; reducing computational cost as well as achieving model efficiency.

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

Compared to the established paradigm of convolution-based video architecture, TimeSformer follows a radically different design and achieves better accuracy compared with the SOTA in this field. It can also be used for long-range modeling of many-minute videos.

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

- Achieved SOTA results on current video classification benchmarks

- Reduced training times and computational costs compared to previous models like I3D amd SlowFast

- Adapted to long-range video for many minutes

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

- Achieved lower accuracy compared to the SOTA methods on SSv2 which requires learning more complex temporal patterns

- Failed to test the model on clips longer than 96 frames due to GPU memory limits

3. Empirical Results

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

- TimeSformer-L achieves the highest accuracy on the Kinetics-400 (80.7%) and Kinetics-600 (82.2%) datasets, and TimeSformer achieves the lowest TFLOPs (0.59), indicating the method is very effective for action recognition tasks in videos.

- The model can process much longer video clips compared to traditional CNNs, indicating its ability for long-range video understanding

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

Yes, there are two weaknesses in the experimental design. First, when comparing accuracy on Diving-48, the table 7 didn’t show scores of previous methods though it’s due to the issue of Diving-48 labels. It’s unclear to compare how well the model achieves while missing information of other methods. Secondly, the ablation study missed the specific parameter settings of space and time self-attention mechanisms. I also wonder how the results look like when there’s only spatial attention or time attention.

4. Summary

Given that the paper presents an interesting method to apply divided self-attention mechanisms from image to video and achieves many SOTA results on current video classification benchmarks, I 80% like this model and its results while 20% dislike the ablation study. Because it doesn't fully investigate the impact of each component of the TimeSformer model. The paper fails to offer enough comparative analysis and show how this specific design improves performance over alternative methods.

5. QA Prompt for a Paper Discussion

5.1. Discussion Question

Why does the TimeSformer model work better on long-range, many-minute videos compared to other traditional methods like CNNs?

5.2. Your Answer

In many-minute videos, key actions spread out over time, so the ability to link these actions is important for accurate video classification. Unlike traditional CNNs, which typically rely on local convolution layers failing to integrate context over such long sequences, the TimeSformer can capture long-range spatiotemporal dependencies by "divided attention" mechanism.