Paper Critique: VideoMAE

Dan Peng
Department of Computer Science
Sep 24, 2024
danpeng@unc.edu

1. Research Problem

1.1. What research problem does the paper address?

This paper addressed the problem of extracting more effective video representations during the self-supervised video pre-training process.

1.2. What is the motivation of the research work?

VideoMAE is motivated by the recent success of MAE models in NLP and images and the need for a more data-efficient video learning method.

2. Technical Novelty

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

Key technical challenges are: handling temporal redundancy in video frames, eliminating temporal correlation that could lead to information leakage during the video reconstruction, and traing in a data-efficient way without large-scale pretraining.

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

This paper offers a novel method to video pre-training using masked autoencoders and can be seen as incremental compared to Image MAE. Unlike existing methods like TimeSformer or ViViT, VideoMAE introduces a high masking ratio and a tube masking strategy that makes the reconstruction task more challenging.

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

- Introduces tube masking with an extremely high masking ratio (90%-95%)
- Becomes SOTA by using plain ViT backbone without any complex adjustments
- Achieves data-efficient learning on small video datasets without requiring large-scale pre-training

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

- The tube masking may not perform well with low motion or static content
- Not scalable to large multi-modal datasets

3. Empirical Results

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

- The model shows that 90%-95% masking ratios achieve optimal performance in video pre-training tasks, which is higher than masking strategies in images (e.g., 50%-75%). This indicates the temporal redundancy makes it different to the extremely high masking ratios in video data.
- VideoMAE achieves 87.4% top-1 accuracy on the Kinetics-400 dataset, 75.4% on SSV2, achieving SOTA in video classification tasks.

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

No, I think this experimental section is almost perfect. The only comparison that I could not understand very well is the column of Extra data in Table 6 and Table 7 where there are few similar datasets and it's hard to compare between these models based on different extra dataset.

4. Summary

This paper presents a self-supervised video pre-training method using MAE with extremely high masking ratios (90%-95%) to handle temporal redundancy. By applying an asymmetric encoder-decoder architecture, the model achieves SOTA performance in video understanding tasks

without large-scale labeled datasets. I like this paper because it rethinks the need for extensive labeled datasets and proposes a self-supervised pretraining. This method could help reduce the computational and data costs that typically limit smaller research groups.

5. QA Prompt for a Paper Discussion

5.1. Discussion Question

Why doesn't the comparison with the state-of-the-art methods on Something-Something V2 use the same extra data, such as K400?

5.2. Your Answer

To emphasize VideoMAE's strength in data-efficient learning without extra pretraining dataset. (I'm not pretty sure about the correctness of this answer and I'm open to discuss later)