Discussion Questions

1. Compared to BERT in NLP and ImageMAE that work best with 15% and 75% masking ratios respectively, why does VideoMAE work well with very high mask ratios (e.g., 90%)?

2. Conceptually, why should tube masking work better than random masking?

3. Why does the proposed framework work well on small amount of videos?

4. Why is it beneficial to do MAE pretraining first, and then fine-tune the resulting model on the same dataset?

5. Does the learned representation generalize when pretraining is done on dataset A and finetuning on dataset B? Why or why not?

6. Would the proposed approach work on longer, more temporally-heavy videos (e.g., movies)?

7. In terms of impact, is MAE-based visual pretraining comparable to its BERT counterpart in NLP? Why or why not?

8. Does self-supervised learning have a future in computer vision?

Discussion Questions

1. Compared to BERT in NLP and ImageMAE that work best with 15% and 75% masking ratios respectively, why does VideoMAE work well with very high mask ratios (e.g., 90%)?

Discussion Questions

2. Conceptually, why should tube masking work better than random masking?

Discussion Questions

3. Why does the proposed framework work well on small amount of videos?

Discussion Questions

4. Why is it beneficial to do MAE pretraining first, and then fine-tune the resulting model on the same dataset?

Discussion Questions

5. Does the learned representation generalize when pretraining is done on dataset A and finetuning on dataset B? Why or why not?

Discussion Questions

6. Would the proposed approach work on longer, more temporally-heavy videos (e.g., movies)?

Discussion Questions

7. In terms of impact, is MAE-based visual pretraining comparable to its BERT counterpart in NLP? Why or why not?

Discussion Questions

8. Does self-supervised learning have a future in computer vision?