Discussion Questions

1. Can TimeSformer recognize actions that involve fast-moving objects?

2. Why does TimeSformer struggle with temporally-heavy datasets such as SSv2? How can we improve it?

3. What is the main reason that divided attention can outperform joint attention?

4. How would the performance change if we swapped the order of time and space attention in each block?

5. Why does the accuracy suddenly drop when the spatial crop side reaches 560 pixels?

6. Why does using the larger ImageNet-21K compared to the ImageNet-1K results in better performance on the K400 dataset but a similar performance on the SSv2 dataset?

7. What are the main advantages of video transformers over 3D CNNs (if any)?

8. Are the comparisons with 3D CNNs fair (given the varying parameter counts)?

9. What are the potential advantages of combining CNNs with Transformers for video recognition?

10. Will transformers replace convolution-based methods for video understanding? Why or why not?

11. How would this approach work for capturing longer range temporal dependencies (10min or more)?

Discussion Questions

1. Can TimeSformer recognize actions that involve fast-moving objects?

Discussion Questions

2. Why does TimeSformer struggle with temporally-heavy datasets such as SSv2? How can we improve it?

Discussion Questions

3. What is the main reason that divided attention can outperform joint attention?

Discussion Questions

4. How would the performance change if we swapped the order of time and space attention in each block?

Discussion Questions

5. Why does the accuracy suddenly drop when the spatial crop side reaches 560 pixels?

Discussion Questions

6. Why does using the larger ImageNet-21K compared to the ImageNet-1K results in better performance on the K400 dataset but a similar performance on the SSv2 dataset?

Discussion Questions

7. What are the main advantages of video transformers over 3D CNNs (if any)?

Discussion Questions

8. Are the comparisons with 3D CNNs fair (given the varying parameter counts)?

Discussion Questions

9. What are the potential advantages of combining CNNs with Transformers for video recognition?

Discussion Questions

10. Will transformers replace convolution-based methods for video understanding? Why or why not?

Discussion Questions

11. How would this approach work for capturing longer range temporal dependencies (10min or more)?