Discussion Questions

1. Does the lack of inductive biases in Transformers make them better suited for large-scale multimodal self-supervised training than CNNs? Is the proposed framework applicable to CNNs?

2. Is Transformer the ultimate tool for multimodal learning? Will it still be dominant in 5 years?

3. Why do we need separate video features for contrastive video-audio and video-text matching?

4. Is the proposed approach limited by the need for datasets with all three modalities (video, audio, text)?

5. Is modality-agnostic architecture that treats all modalities the same way a good idea?

6. What future work could this paper inspire? It seems the model is very expensive to train.

7. Is there a point to train on all 3 modalities if are only going to test on one modality? Should we use a more specialized pretraining designed for our downstream tasks (e.g., action recognition)?

8. Could the self-supervised training framework of VATT completely replace fully supervised training of transformers for video tasks? What are the weaknesses of the proposed approach?

9. Why don’t the authors use MAE-based objective? Is contrastive pretraining better than generative pretraining (i.e., masked autoencoding, autoregressive generation)?

Discussion Questions

1. Does the lack of inductive biases in Transformers make them better suited for large-scale multimodal self-supervised training than CNNs? Is the proposed framework applicable to CNNs?

Discussion Questions

2. Is Transformer the ultimate tool for multimodal learning? Will it still be dominant in 5 years?

Discussion Questions

3. Why do we need separate video features for contrastive video-audio and video-text matching?

Discussion Questions

4. Is the proposed approach limited by the need for datasets with all three modalities (video, audio, text)?

Discussion Questions

5. Is modality-agnostic architecture that treats all modalities the same way a good idea?

Discussion Questions

6. What future work could this paper inspire? It seems the model is very expensive to train.

Discussion Questions

7. Is there a point to train on all 3 modalities if are only going to test on one modality? Should we use a more specialized pretraining designed for our downstream tasks?

Discussion Questions

8. Could the self-supervised training framework of VATT completely replace fully supervised training of transformers for video tasks? What are the weaknesses of the proposed approach?

Discussion Questions

9. Why don’t the authors use MAE-based objective? Is contrastive pretraining better than generative pretraining (i.e., masked autoencoding, autoregressive generation)?