Paper Summary

Paper: Siamese Masked Autoencoders

Relevant previous works:
1. Masked Autoencoders are Scalable Vision Learners
2. A Simple Framework for Contrastive Learning of Visual Representations
3. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
4. Masked Autoencoders as Spatialtemporal Learners
5. Emerging Properties in Self-Supervised Vision Transformers

Novelty:
1. Asymmetric masking
2. Boundary-object detection.

Contributions:
1. Asymmetric masking

Key ideas:
1. Temporal vs. spatial correlation.
2. Contrastive vs. predictive modeling as a self-supervised task.
3. Redundancy in the temporal dimension.

Key concepts to understand better:
1. Isotropic.
2. Representation collapse.
3. Affinity matrix.
4. Linear projections [31]
5. Effective epoch
6. k-nearest neighbor inference.
7. Emergent abilities
8. Zero-shot usability
9. CLS token

Questions:
1. How to do more with less?
2. What weights are shared and why?
3. What is the effect of the patch size on the method?
4. What is the effect of the occlusion strategy on the method? In terms of % as well as occlusion placement.
5. Why do they do minimal data augmentation, why is that a positive thing and how does it affect training?
6. What is happening with the numbers of Table 20 for 0.9 vs 0.95 mask ratio.
7. What is the impact of the frame gap and is that a good measurement or should the model train on uniform variance in pixel correlation?
8. Is the location of the non occluded part of the image important or the sampling strategy?

Future work:
1. Extend to multiframe prediction
2. Scalability in terms of data and model.
3. Data type: egocentric videos vs. "in-the-wild" internet videos.

Method

Step 1: Random sample two frames: f_1 & f_2 with $t_{f_1} < t_{f_2}$.

Step 2: Patchify like in ViT.
Add positional embedding + linear projections + CLS token.

Step 3: Masking asymmetrically
No masking for f_1
Masking of f_2.

Step 4: Encoding
⇒ 1. Joint
2. Siamese

Step 5: Decoding
⇒ 1. Joint
2. Cross-self
3. Cross