VideoMAE For Emotion Recognition

This Project is funded by the European Union HORIZON EUROPE - WIDERA-2021-ACCESS-03 (Twinning) programme under the grant agreement No. 101078950

Natia Kukhilava
Tatia Tsmindashvili
Rapael Kalandadze
Laura Ferrari
Valeryia Strizhkova

Joy

Trust

Fear

Surprise

Sadness

Disgust

Anger

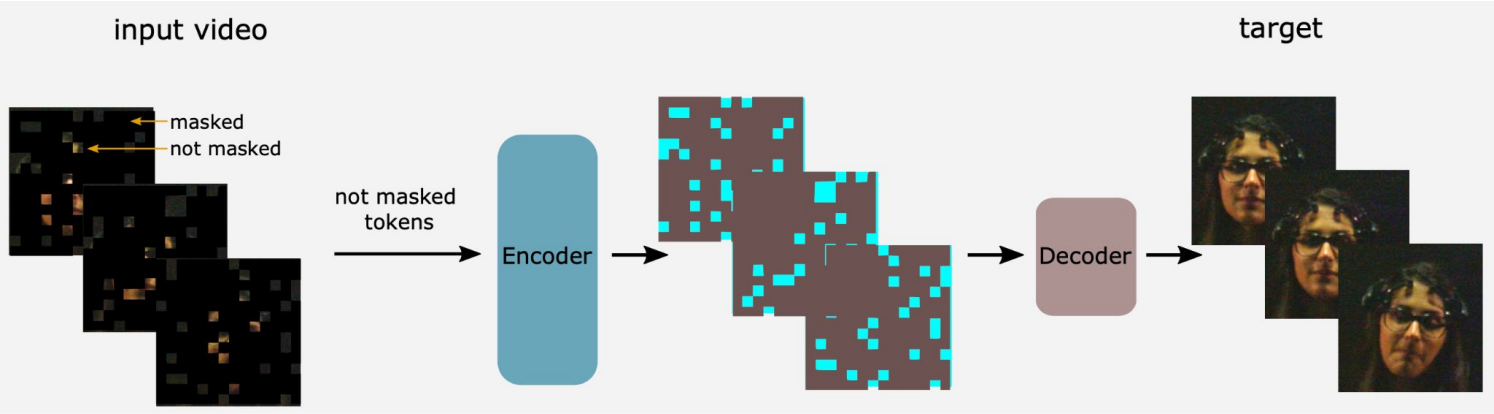
Anticipation

Research Goal

- Try VideoMAE on larger videos
- Scaling VideoMAE
- Compare SSL and SL pretraining strategies

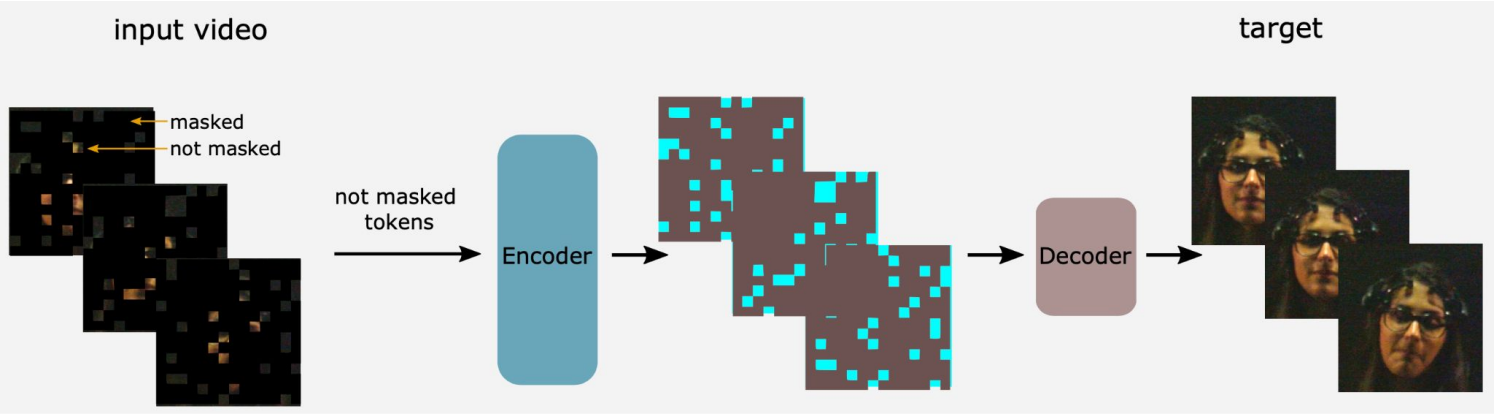
VideoMAE

- Self Supervised Video pre-training Method
- Inspiration
- Adaptation to Video - masks random “cubes”



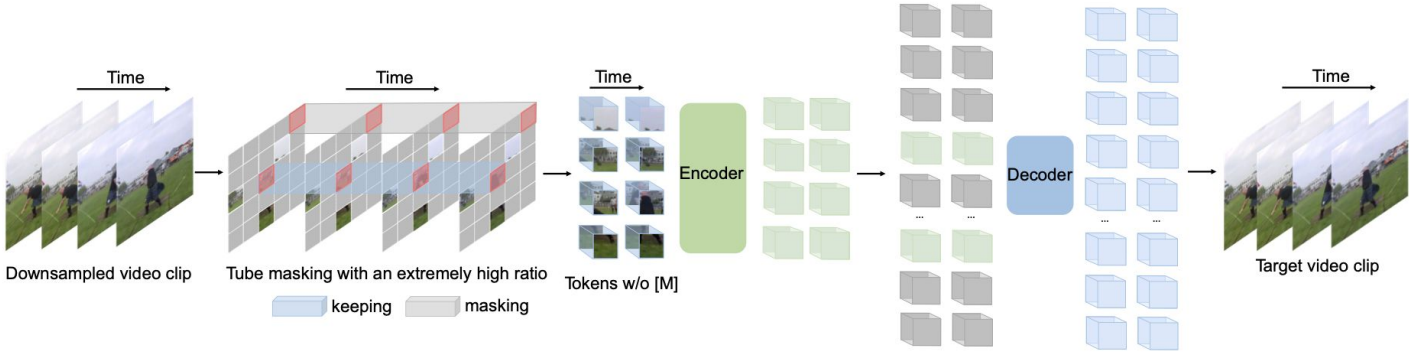
VideoMAE - Challenges

- Temporal Redundancy - easy to guess missing part without understanding video's content
- Temporal Correlation - masked in one frame, visible in the next



VideoMAE - Strategic Masking

- Temporal Redundancy - high masking ratio
- Temporal Correlation - tube masking strategy

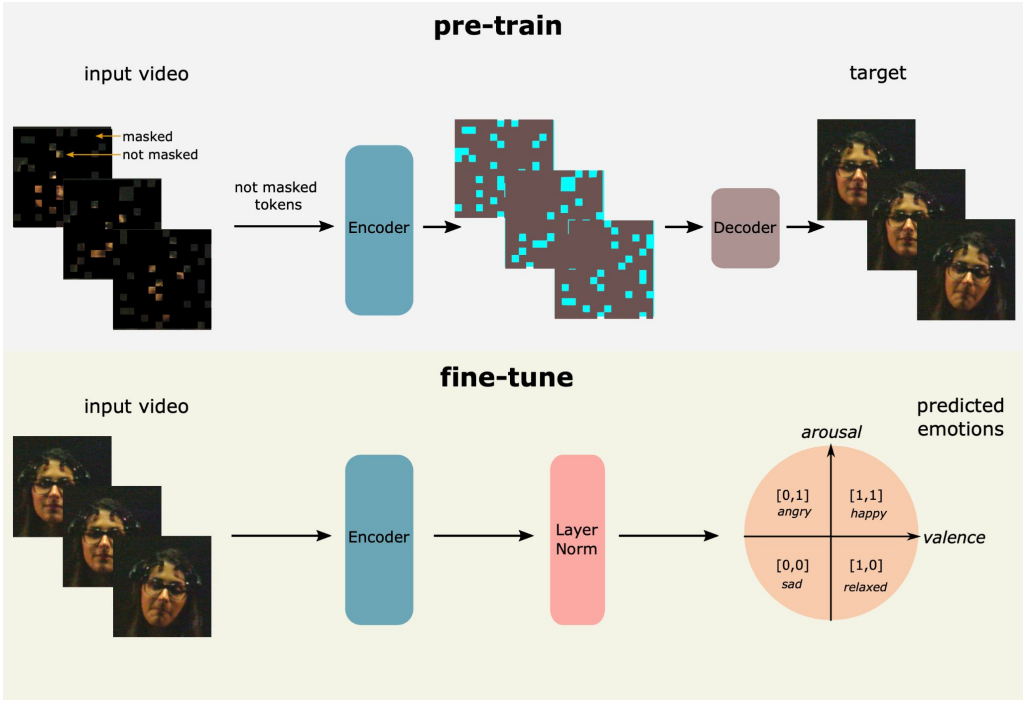


VideoMAE - Strategic Masking

Comparisons with the results of previous self-supervised pre-training methods on different datasets:

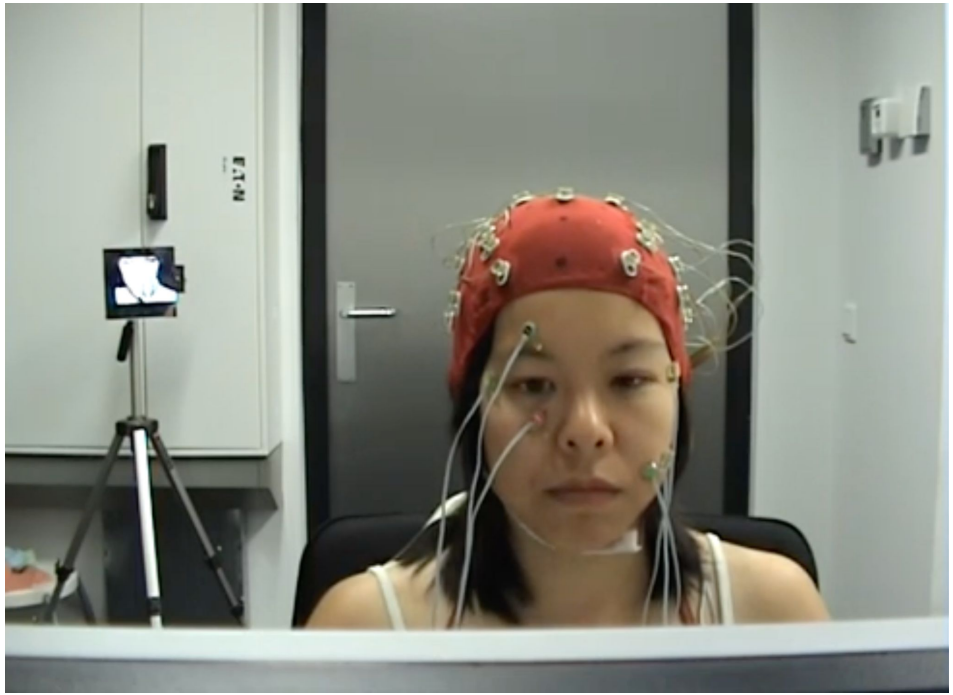
dataset	training data	*from scratch*	MoCo v3	VideoMAE
K400	240k	68.8	74.2	**80.0**
Sth-Sth V2	169k	32.6	54.2	**69.6**
UCF101	9.5k	51.4	81.7	**91.3**
HMDB51	3.5k	18.0	39.2	**62.6**

Pipeline


pre-train
input video
masked
not masked
not masked tokens
Encoder
Decoder
target
fine-tune
input video
Encoder
Layer Norm
arousal
valence
[0,1] angry
[1,1] happy
[0,0] sad
[1,0] relaxed
predicted emotions

Datasets - DEAP

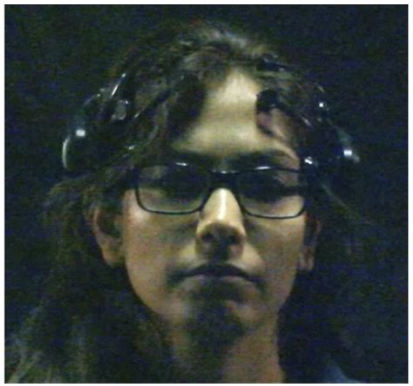
- 22 Participants
- 40 Videos
- Video length - 1 minute

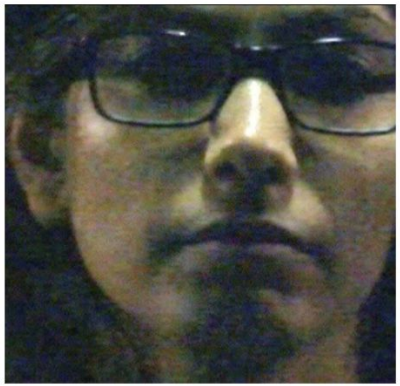


Datasets - AMIGOS

- 40 Participants
- 16 Videos
- Video length - 2-3 minute

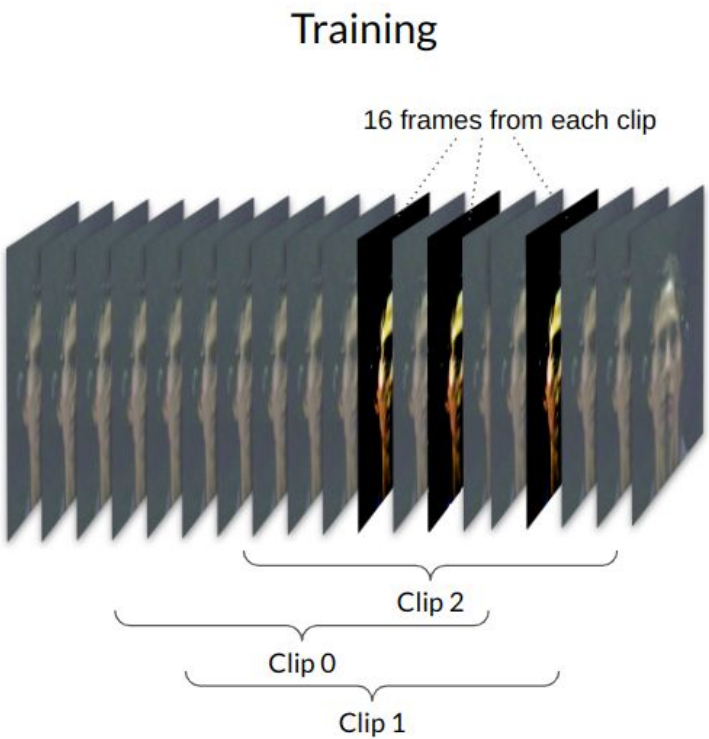
Cropping Strategy

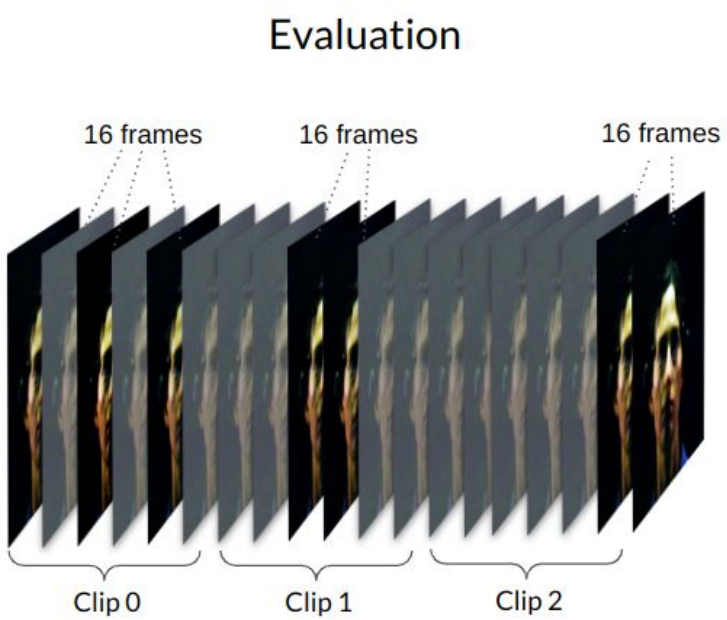




Larger crop produces 7-9% higher accuracy and weighted F1 score

Data Input


Training
16 frames from each clip
Clip 2
Clip 0
Clip 1


Evaluation
16 frames
16 frames
16 frames
Clip 0
Clip 1
Clip 2

Experiments - DEAP

SSL pre-training data	Supervised pre-training data	Backbone	GFLOPs	Params	F1-score
Kinetics-400	-	ViT-B	180	87M	52.51
Kinetics-400	Kinetics-400	ViT-B	180	87M	53.35
Kinetics-400	Kinetics-400 & AMIGOS	ViT-B	180	87M	**60.69**
KAD	-	ViT-B	180	87M	55.64

Framework: https://github.com/EmotionLab/EmotionVMAE

Future Work

- Try ViT-L Backbone for model scaling
- Try without weights
- Finalize and Publish Framework

Questions?