Maastricht University

Investigating the Explainability Power of Transformers in the Video Domain.

Research Internship

Christopher du Toit

Introduction

Introduction

- Explainable AI (XAI) is very important.
- Transformers new in CV.
- First time emotion detection is done.
- Adapt visual techniques from images to videos.
- Quantitatively compare visual techniques.
- Emotion trends inline with affective computing?

Table of contents/structure

- Methodology
- Experiments and results
- Discussion and further research
- Conclusion

Methodology

Problem Statement

How does the self-attention mechanism identify emotions?

Research Questions

1. What is the performance of the TimeSformer model on detecting emotions when trained and tested on the CREMA-D [3] dataset?
2. How can visual techniques such as Grad-CAM be extended to the video domain?
3. Can the reasoning behind the prediction outputs of the TimeSformer model be visualized using the extended visual techniques?
4. Can occlusion testing be used to quantitatively evaluate the visual technique performances?
5. Based on the performed analysis, can we obtain a set of visual patterns for each emotion?

TimeSformer

- Multi-Head Self-Attention (MHSP) mechanism success in NLP (BERT) [7].
- Extended version of the vision transformer (ViT) [8].
- Pure transformer model for videos [2], [14], [15].

TimeSformer

Patches = N
Frames = F

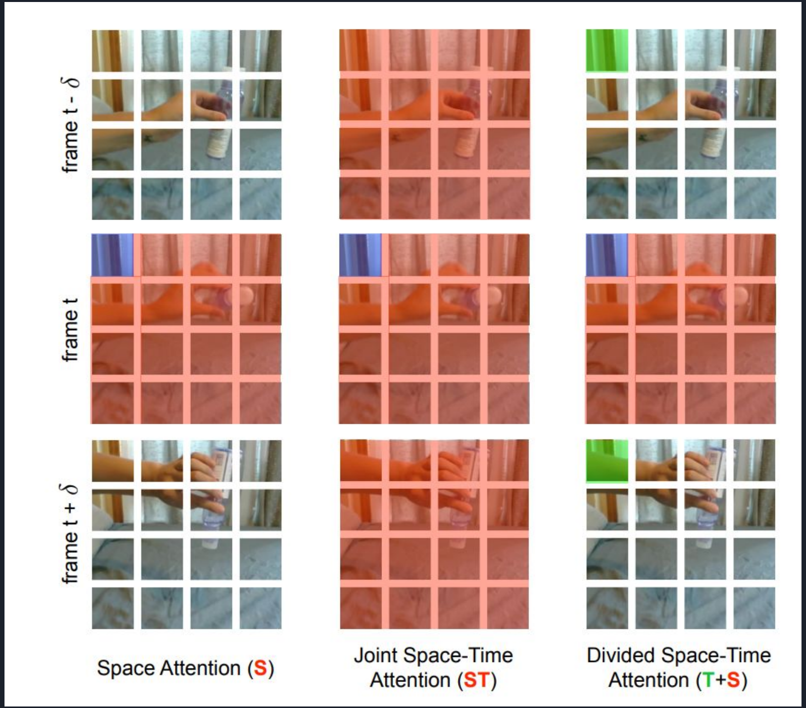

frame t - δ
frame t
frame t + δ
Space Attention (S)
Joint Space-Time Attention (ST)
Divided Space-Time Attention (T+S)

Image from [2].

N NF N + F

TimeSformer

- Multi-Head Self-Attention (MHSP) mechanism success in NLP (BERT) [7].
- Extended version of the vision transformer (ViT) [8].
- Pure transformer model for videos [2], [14], [15].

TimeSformer

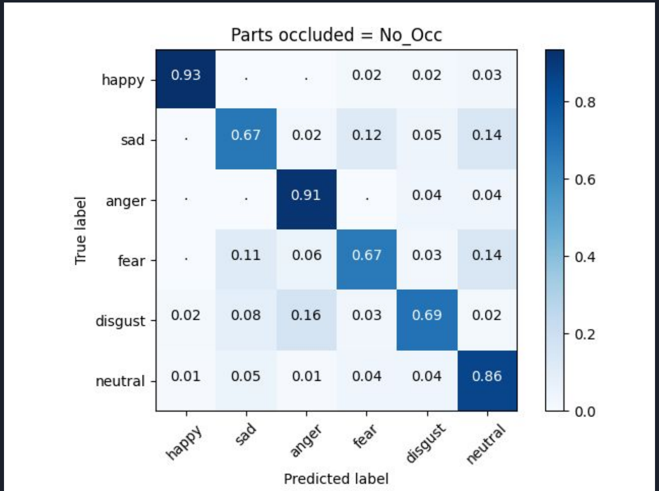
- Multi-Head Self-Attention (MHSP) success in NLP (BERT) [7].
- Extended version of the vision transformer (ViT) [8].
- Pure transformer model for videos [2], [14], [15].
- Divided Space-Time attention.
- In practice: temporal attention then spacial attention.
- Transformers are ideal for transfer learning.

CREMA-D

- 6 emotions: Happy, Anger, Sad, Disgust, Fear and Neutral.
- 7442 videos.
- 91 actors.
- Crowd-sourced labelling.
- ‘Non-spontaneous’ emotions.
- Facial expressions.

Model training and performance

- 8 evenly spaced out frames.
- Cropping with FaceNet.
- Dataset split 90% training, 5% validation, 5% testing.
- 81% accuracy produced.
- Additional data augmentation not needed.



Visual Techniques

- Grad-CAM
- Rollout
- Transformer Interpretability Beyond Attention Visualization (TIBAV)

Grad-CAM

- Grad-CAM [13]:
  - Uses gradients of last layer.
  - Class-specific.
- Grad-CAM++ [4]:
  - Second order gradients of last layer.
  - Class specific.
- Eigen-CAM [9]:
  - First principle component of the 2D activations.
  - Not class specific.

Rollout

- Rollout [1]:
  - Designed for NLP.
  - Recursively multiples attention maps together from each block.
  - $A_L = \min_h(A_L)$
  - $R = (I + A_L)R$
- Rollout-Grad:
  - Inspired by TIBAV.
  - $A_L = \mathbb{E}_h(A_L * \Delta A)$
  - $R = (I + A_L)R$
  - Class-specific

TIBAV

- Class-specific [5], [6].
- Similar to Rollout-Grad.
- Takes positive attention.
- $A_L = \mathbb{E}_h((A_L * \Delta A)^+)$
- $R = R + A_L \bullet R$

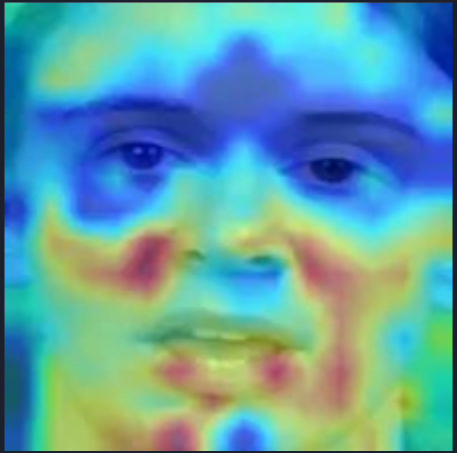
Experiments and results

Adapting and testing visual techniques

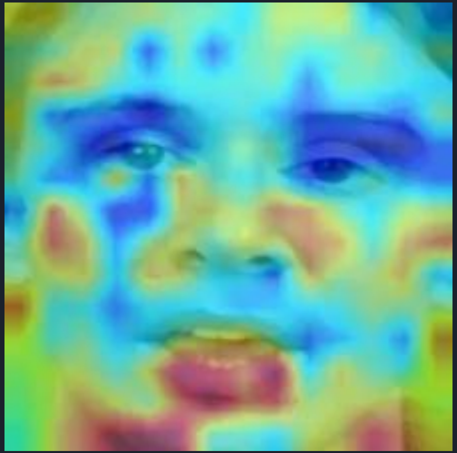
- Adaptation relatively simple.
- Divided space-time attention.
- Spacial attention encapsulates the temporal attention.

Adapting and testing visual techniques

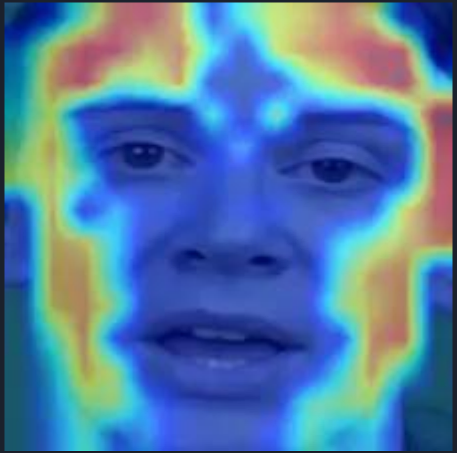
Anger



Grad-CAM



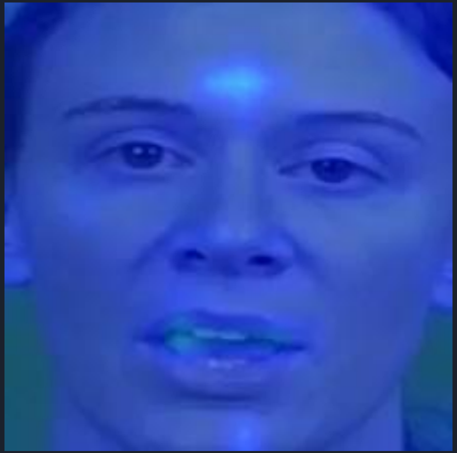
Grad-CAM ++



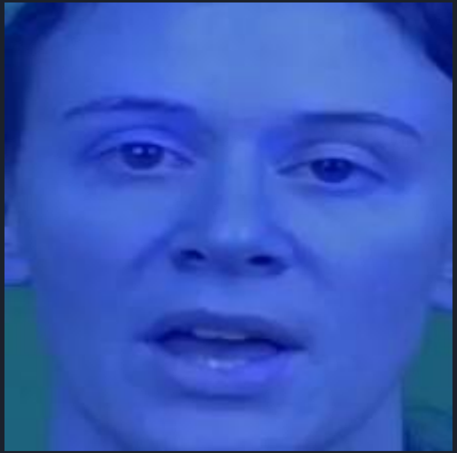
Eigen-CAM

Adapting and testing visual techniques

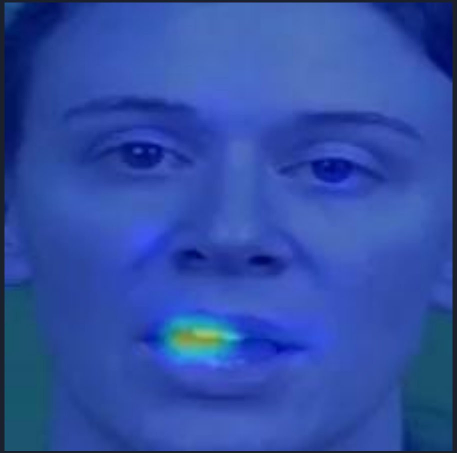
Anger



Rollout



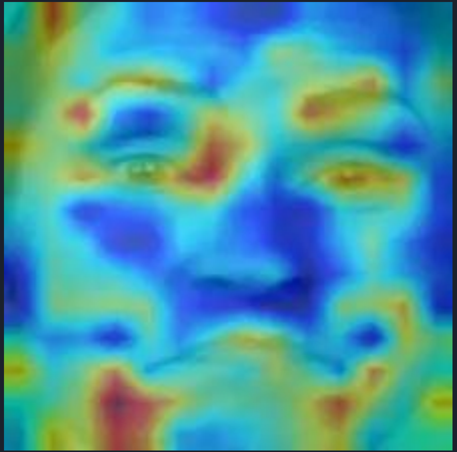
Rollout-Grad



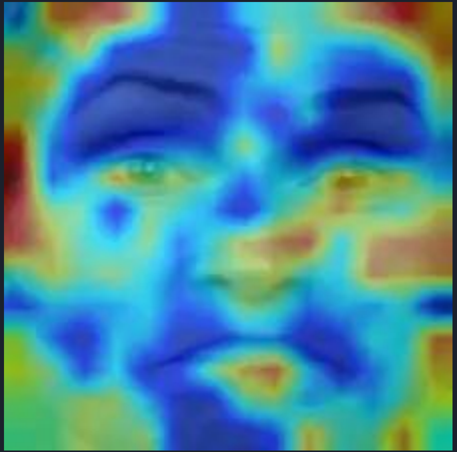
TIBAV

Adapting and testing visual techniques

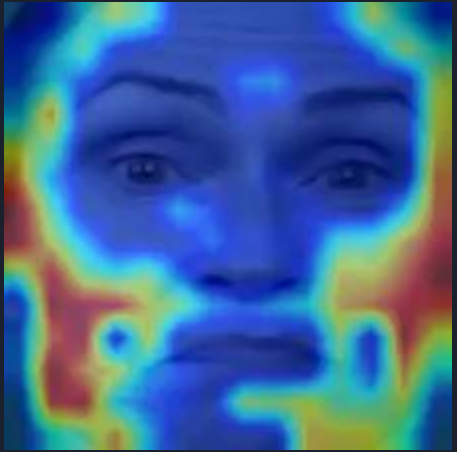
Happy



Grad-CAM



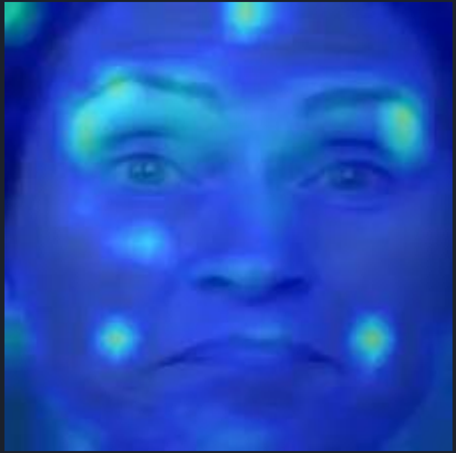
Grad-CAM ++



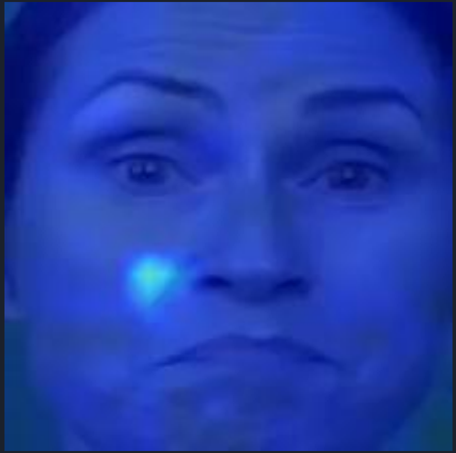
Eigen-CAM

Adapting and testing visual techniques

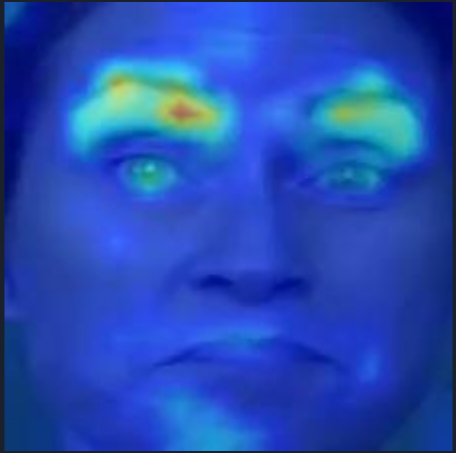
Happy



Rollout



Rollout-Grad

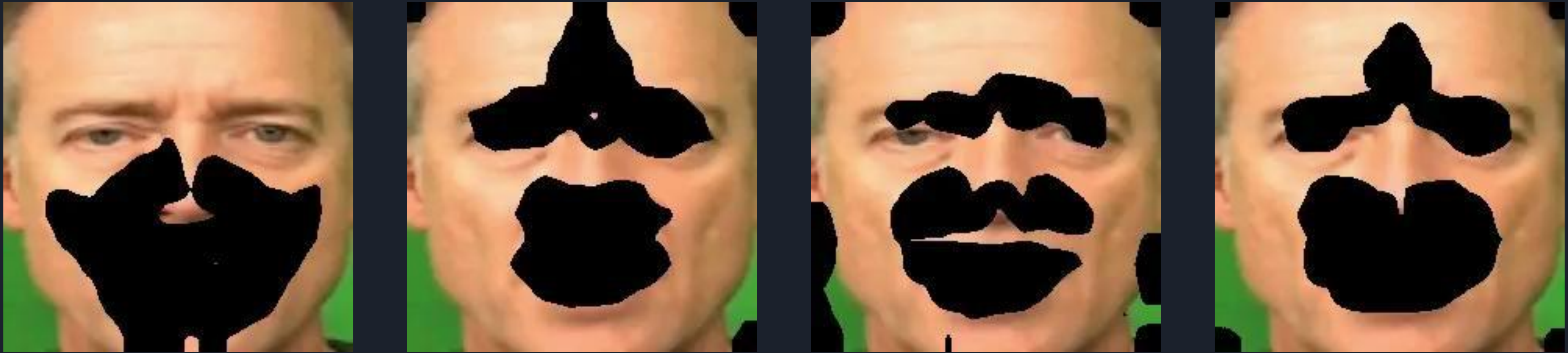


TIBAV

Visual technique occlusion testing

- Quantitative comparison.
- Occluded most important patches.
- Check testing accuracy.
- Repeat for least important patches.
- Grad-CAM acts as baseline - should perform badly.

Visual technique occlusion testing



Grad-CAM

Rollout

Rollout-Grad

TIBAV

30% spacial and 2 frames occluded.

Visual technique occlusion testing

Grad-CAM

Rollout

Rollout-Grad

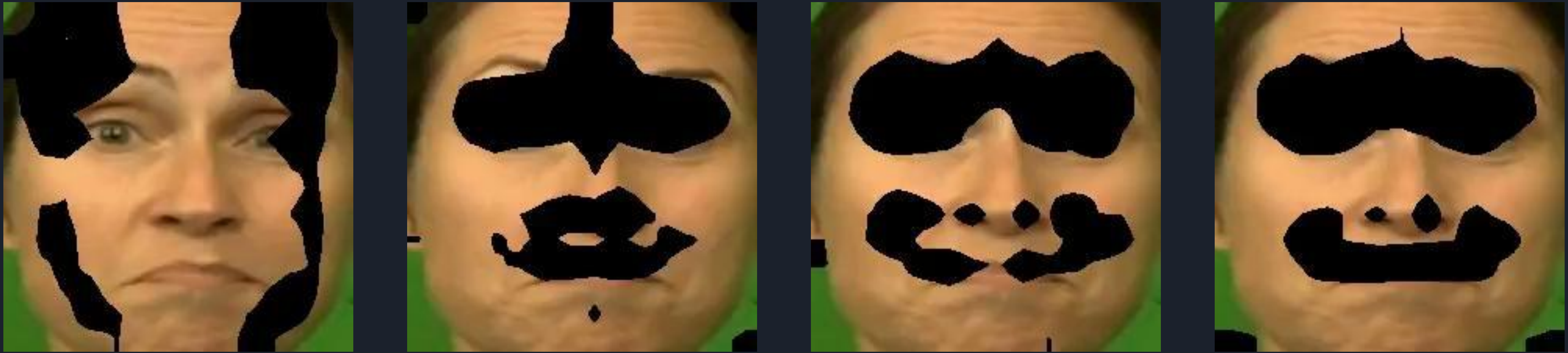
TIBAV

30% spacial and 2 frames occluded.

Visual technique occlusion testing

- Quantitative comparison.
- Occluded most important patches.
- Check testing accuracy.
- Repeat for least important patches.
- Grad-CAM acts as baseline - should perform badly.

Visual technique occlusion testing

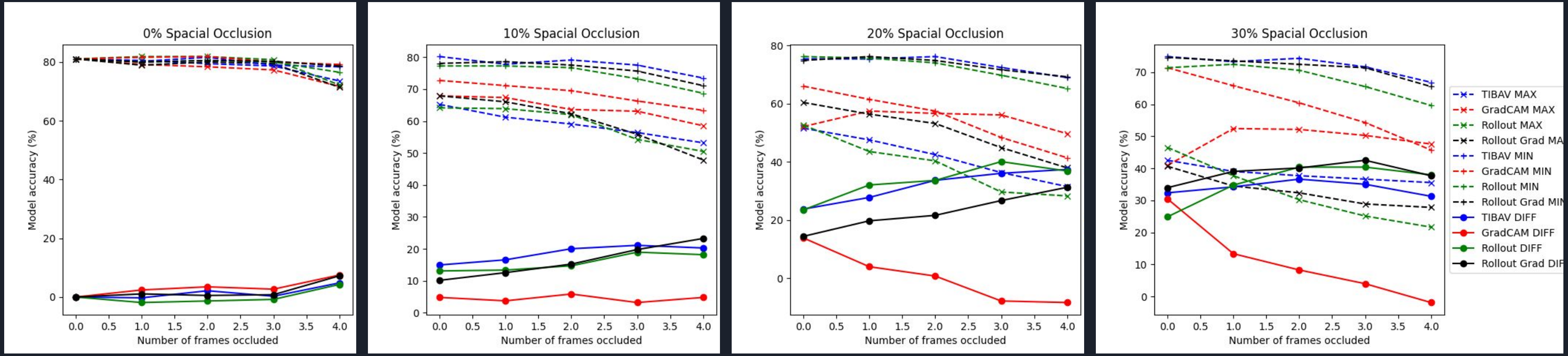


30% spacial and 2 frames occluded.

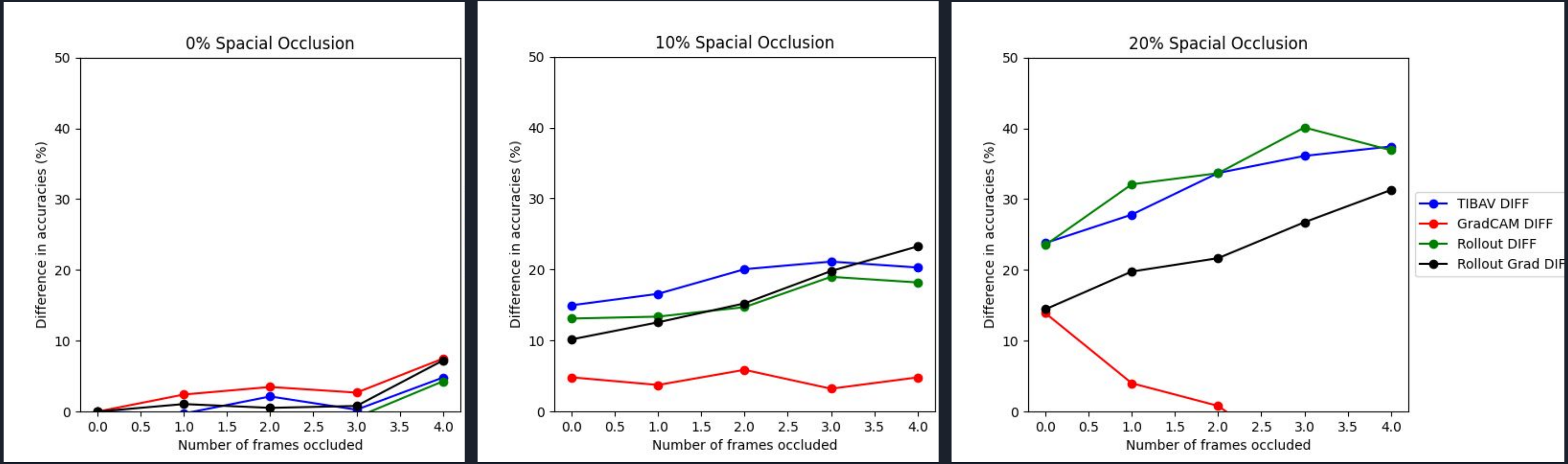
Visual technique occlusion testing

30% spacial and 2 frames occluded.

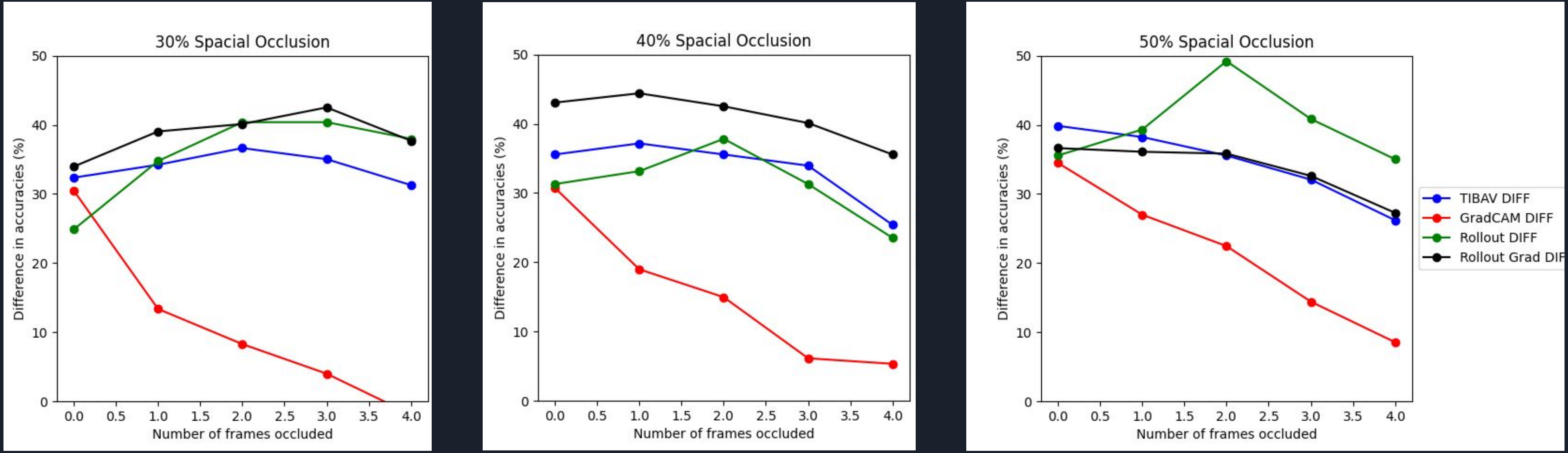
Visual technique occlusion testing



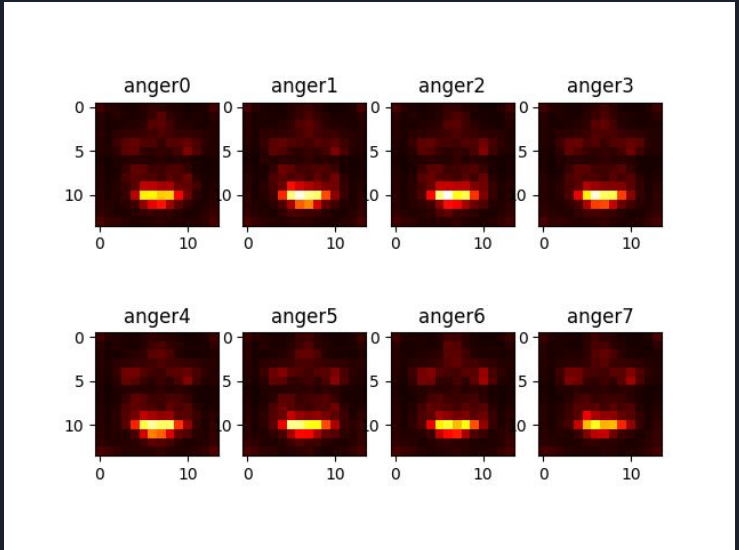
Visual technique occlusion testing

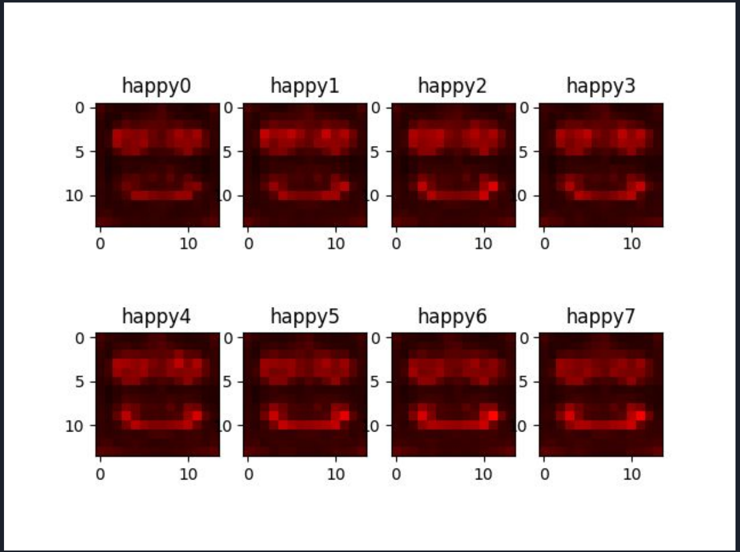


Visual technique occlusion testing



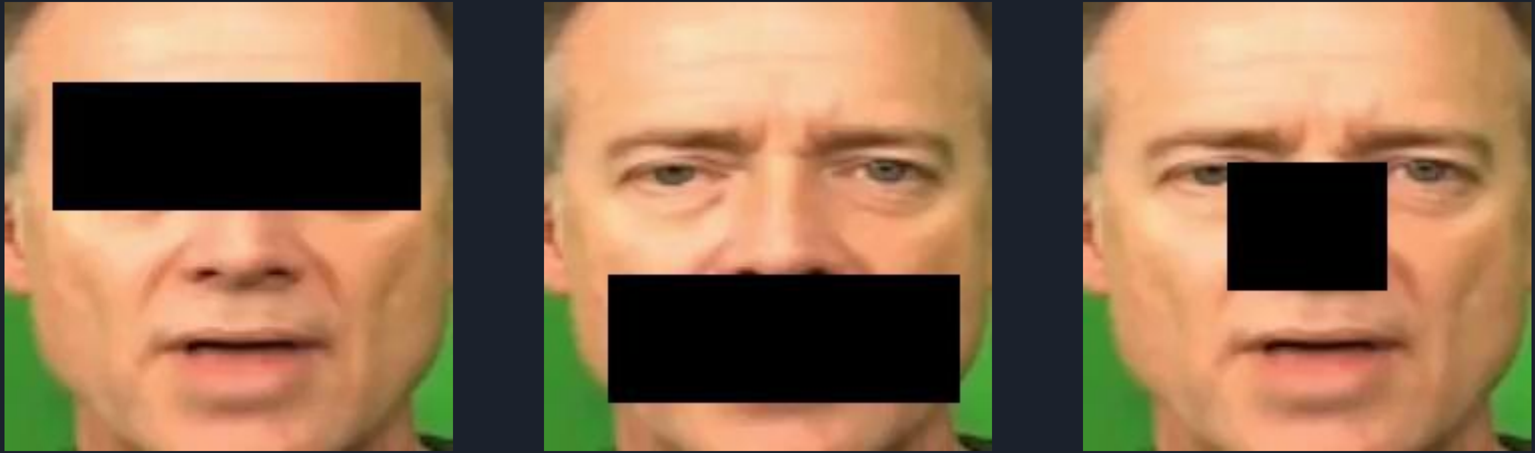
Emotion trends



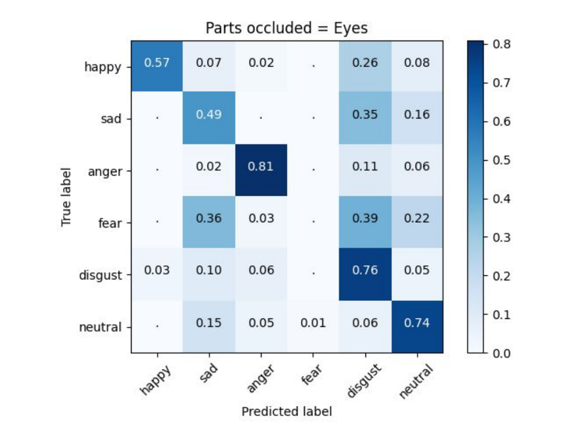


Plots are to scale

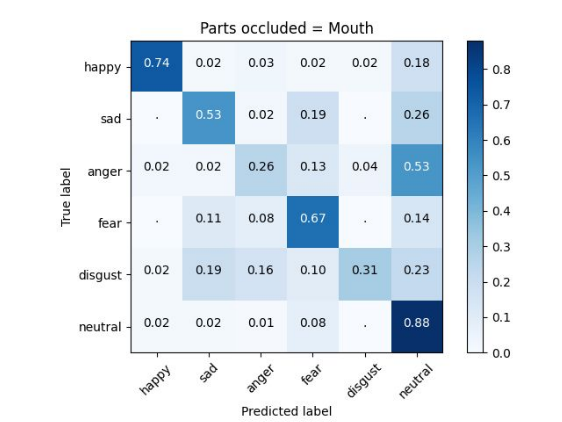
Facial part occlusion testing



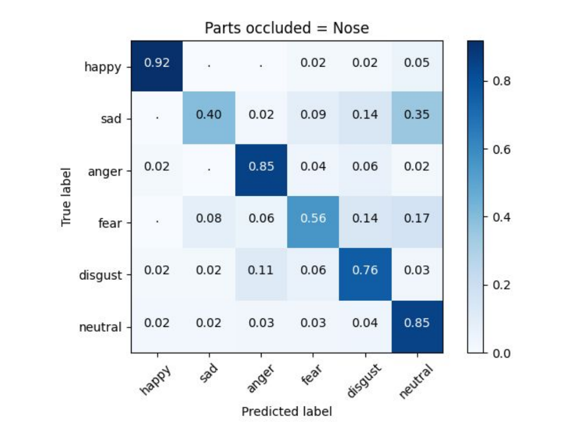
Facial part occlusion testing



Parts occluded = Eyes

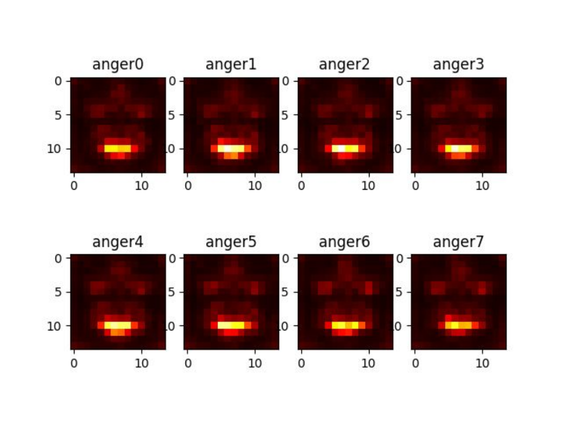


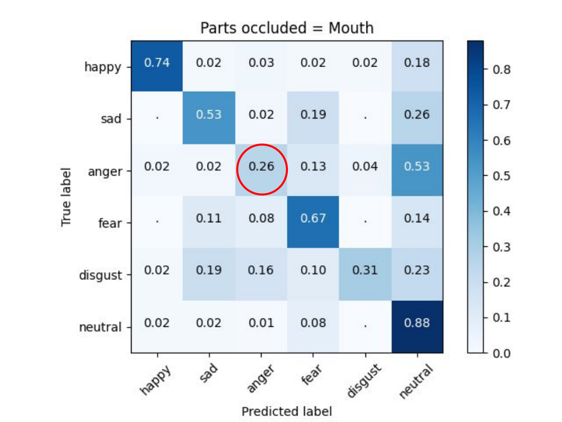
Parts occluded = Mouth



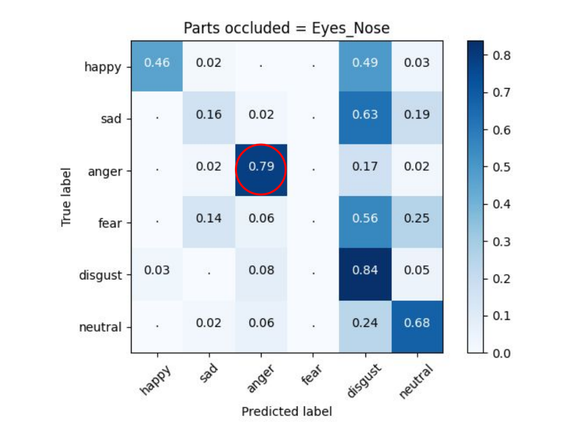
Parts occluded = Nose

Facial part occlusion testing



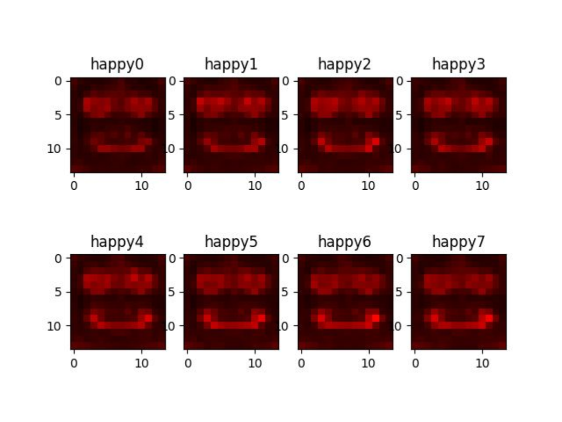


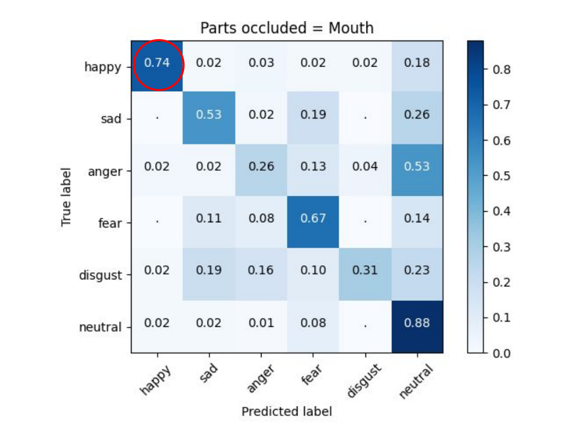
	happy	sad	anger	fear	disgust	neutral
happy	0.74	0.02	0.03	0.02	0.02	0.18
sad	.	0.53	0.02	0.19	.	0.26
anger	0.02	0.02	0.26	0.13	0.04	0.53
fear	.	0.11	0.08	0.67	.	0.14
disgust	0.02	0.19	0.16	0.10	0.31	0.23
neutral	0.02	0.02	0.01	0.08	.	0.88

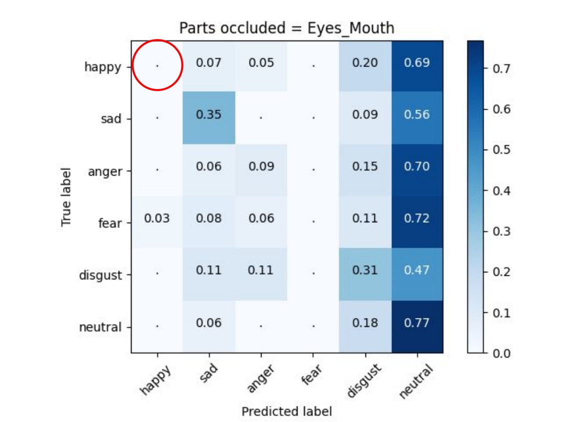


	happy	sad	anger	fear	disgust	neutral
happy	0.46	0.02	.	.	0.49	0.03
sad	.	0.16	0.02	.	0.63	0.19
anger	.	0.02	0.79	.	0.17	0.02
fear	.	0.14	0.06	.	0.56	0.25
disgust	0.03	.	0.08	.	0.84	0.05
neutral	.	0.02	0.06	.	0.24	0.68

Facial part occlusion testing







Discussion and further research

Discussion

- Divided attention makes it simple.
- Temporal attention had little influence.
- Trends:

Happy	Eyes and mouth
Anger	Mouth
Sad	Eyes and nose
Fear	Eyes
Disgust	Mouth
Neutral	Whole face - not well defined

- Trends inline with affective computing [11].

Further research

- Fine tune model on spontaneous emotion dataset.
- Repeat and compare results.
- Come up with an improved visual technique.
- Multimodal visualization with sound.

Conclusion

Conclusion

- TimeSformer 81% accuracy on CREMA-D dataset.
- Grad-CAM, Rollout and TIBAV adapted and analyzed.
- TIBAV performs best overall with occlusion testing.
- TIBAV visuals agree with facial occlusion trends.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021.

[3] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset, 2014.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Gradcam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar 2018.

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, 2021.

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

References

[9] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. 2020 International Joint Conference on Neural Networks (IJCNN), Jul 2020.

[10] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[11] Rosalind W. Picard. Affective computing. ISBN 978-0-262-16170-1, 1997.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2015.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2):336–359, Oct 2019.

[14] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021.

[15] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers, 2021.