Explain to Train (ET): Leveraging explanations to enhance the training of a Multimodal Transformer

Meghna P Ayyar, Jenny Benois-Pineau, Akka Zemmari,

Overview

- **Introduction**
- **Feature Explanation Method (FEM)**
- **Rollout-FEM for Transformers**
- **ET Framework**
  - Video Transformer
  - Signal Transformer
  - Multi-modal Training
- **Results**
  - Validation on UCF50 dataset
  - Multimodal dataset
- **Conclusion**

Introduction

- Explainable AI (XAI) is vital for improving transparency and reliability of neural network decisions.
- Transformers have emerged as SOTA for various tasks for single modality like image, language, ... and multimodal approaches.
- The potential of XAI methods for training transformers remains underexplored.

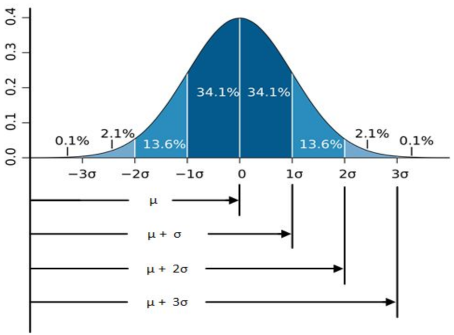
A young lady wearing blue and black is running past an orange cone.

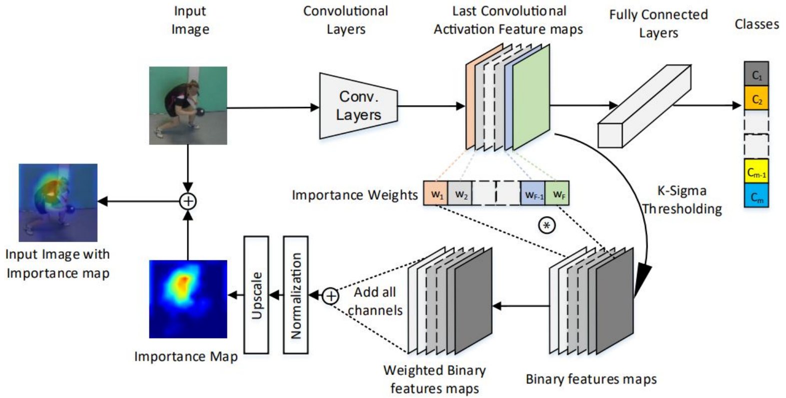
Explanation-guided Training (ET): adapts an XAI method (FEM) [2] for transformers and identifies important input regions to guide the model to focus on the salient regions during fine-tuning

[1] Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X. and Sclaroff, S., 2018. Top-down neural attention by excitation backprop. *IJCV*, *126*(10), pp.1084-1102.
[2] Fuad, K.A.A., Martin, P.E., Giot, R., Bourqui, R., Benois-Pineau, J. and Zemmari, A., 2020, November. Features Understanding in 3D CNNs for Actions Recognition in Video. In 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6). IEEE.

FEM: Feature Explanation Method [1]

The core of the method relies in the back-tracing of “strong” features from the last feature-layer (conv layer). It “explains” the Network decisions at the generalization step.





K-Sigma Thresholding: Convolutional follows normal distribution. So we can apply ***μ* ± *kσ*** threshold rule to extract rare important features. Values higher than the threshold is kept.

$$B_k(a_{i,j,k}) = \begin{cases} 1 & \text{if } a_{i,j,k} \geq \mu_k + K * \sigma_k \\ 0 & \text{otherwise} \end{cases}$$

Publicly Available at:
https://github.com/labribkb/fem/blob/main/FEM.ipynb

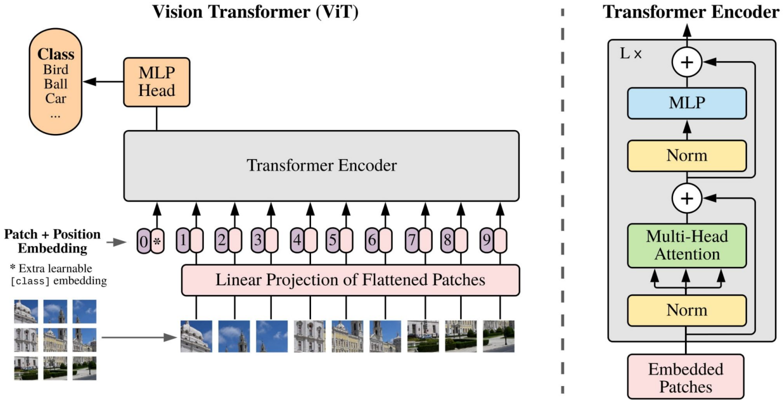
Step 1: Generate Binary Map of the last conv layer activations with K-Sigma thresholding

Step 2: Weighted Average of the binary maps using the mean activations as weights

Step 3: Normalize and Upscale to input dimension

[1] Fuad, K.A.A., Martin, P.E., Giot, R., Bourqui, R., Benois-Pineau, J. and Zemmari, A., 2020, November. Features Understanding in 3D CNNs for Actions Recognition in Video. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.

Rollout- FEM for Transformers



Typical ViT[1] Model

- Self-attention A for each encoder block is computed as

$$A = Q\dot{K}^T$$

- Attention Rollout [2] is used to visualize these attentions across the layers

$$A'^{l} = I + \sum_{h=1}^{H} A_h^l \qquad A_{roll} = \prod_{l=1}^{L} A'^{l}$$

- Rollout weights attentions of different heads equally.

- **We propose:** Use FEM to assign importance per head and choose only the strong attentions for visualization

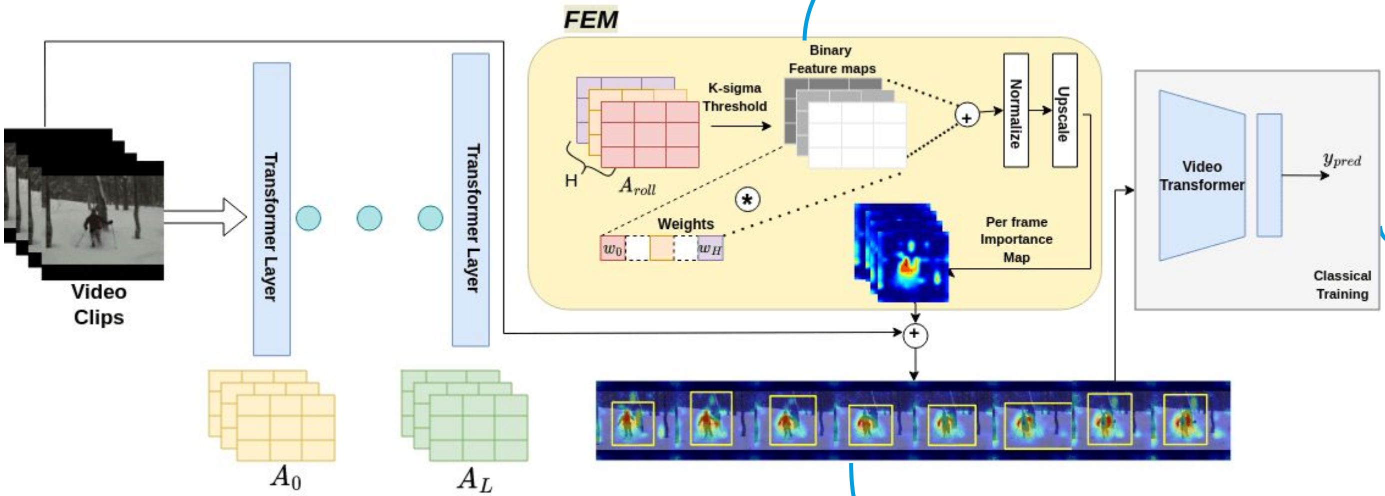
$$A_h'^{l} = I + A_h^l \forall h = 1, \ldots H \qquad A_{h,roll} = \prod_{l=1}^{L} A_h'^{l}$$

$$b_h(A_{h,roll}) = \begin{cases} 1 & \text{if } a_{i,h} \geq \mu_h + K * \sigma_h \\ 0 & \text{otherwise} \end{cases}$$

[1]Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021
[2] Samira Abnar, Willem H. Zuidema, Quantifying Attention Flow in Transformers. ACL 2020: 4190-4197

ET Framework: Training Video Transformer

2. FEM: K-sigma thresholding and combine with channel importance



$$A^h_{roll} = \prod_{l=1}^{L} A^h_l$$

4. Classical training of the transformer with non-salient regions masked out

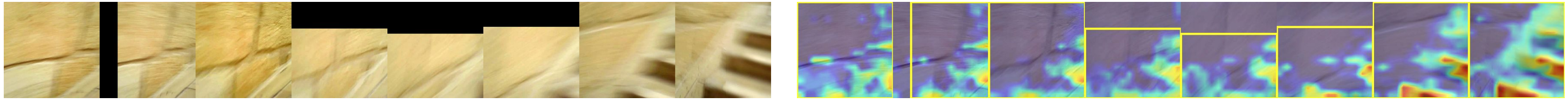
1. Extract attention Maps: A

3. Dynamic size cropping of the input frames to retain salient region with highest area

Visualizations

UCF50[1] - Horse Riding

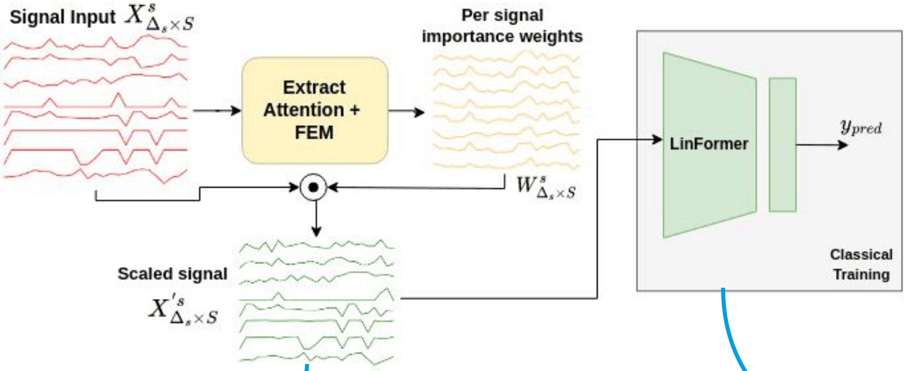
UCF50[1] - Golf Swing

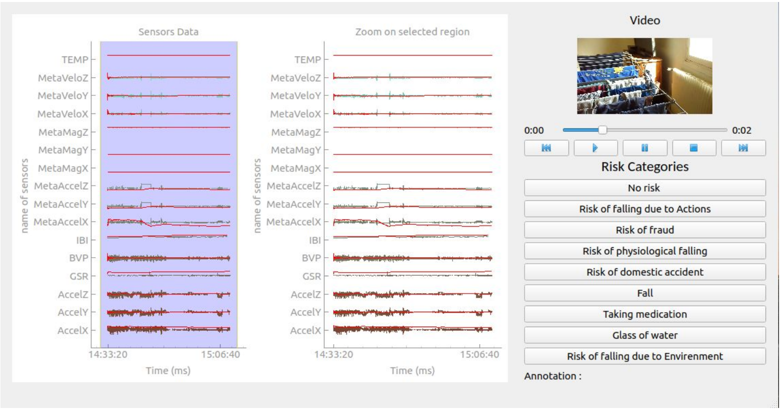


BIRDS[2]: Risk of Fall

[1] Reddy, K.K. and Shah, M., 2013. Recognizing 50 human action categories of web videos. Machine vision and applications, 24(5), pp.971-981.
[2] Mallick, R., Yebda, T., Benois-Pineau, J., Zemmari, A., Pech, M. and Amieva, H., 2022. Detection of risky situations for frail adults with hybrid neural networks on multimodal health data. IEEE MultiMedia, 29(1), pp.7-17.

ET Framework: Training Sensor Transformer





BIRDS[1] : 16 sensors, sampled with window size of 25

Scale the signals by the importance weights

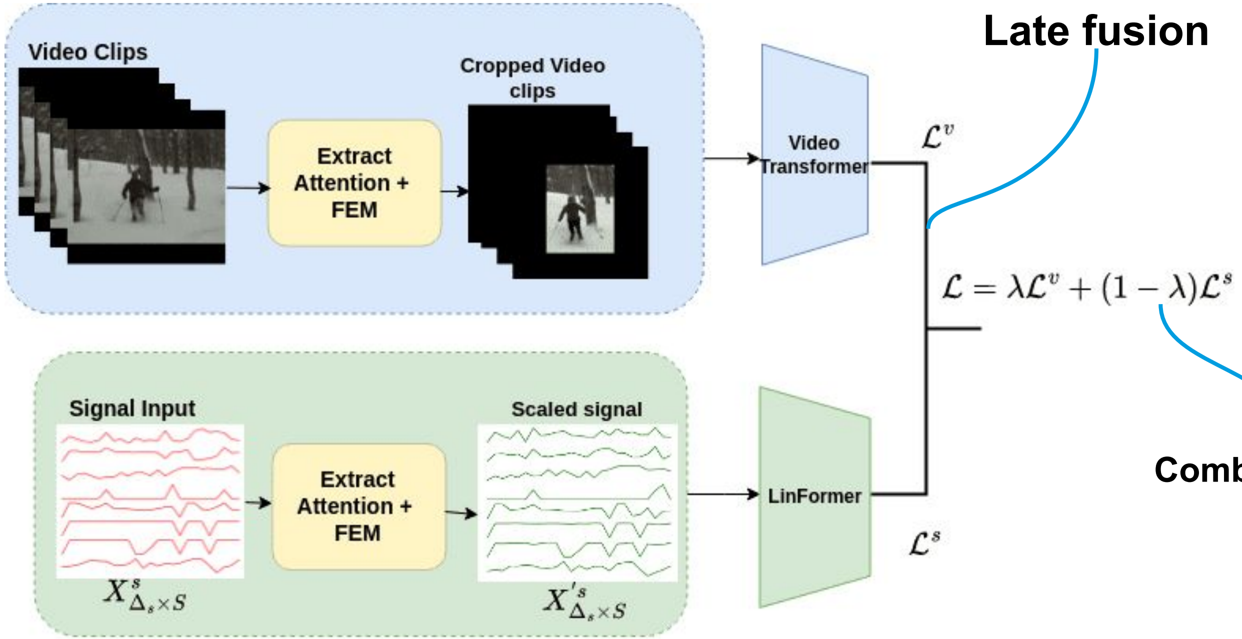
Classical training of the transformer with scaled signals

[1] Mallick, R., Yebda, T., Benois-Pineau, J., Zemmari, A., Pech, M. and Amieva, H., 2022. Detection of risky situations for frail adults with hybrid neural networks on multimodal health data. IEEE MultiMedia, 29(1), pp.7-17.

ET Framework: Multimodal Training

Video Branch

Sensor Branch



Late fusion

Combined Loss

Results: Validation on UCF50 dataset

Model	Top-1 Acc
TimeSFormer [11]	92.27%
Swin Transformer (Swin-T) [12]	91.01%
Video Swin-T-In (IFI) [25]	93.04%
TimeSFormer + ET (Ours)	**94.14%**

Top-1 test accuracy on the UCF50 dataset for videos

- UCF50 activity recognition dataset of 50 action classes.
- Interpreting For Improving (IFI) [1]: combines class-specific attention gradients with the attention weights, to provide extra supervision during training

Our method improves on both the vanilla TimesFormer and training with IFI

[1] Rupayan Mallick, Jenny Benois-Pineau, Akka Zemmar, IFI: Interpreting for Improving: A Multimodal Transformer with an Interpretability Technique for Recognition of Risk Events. MMM (4) 2024: 117-131

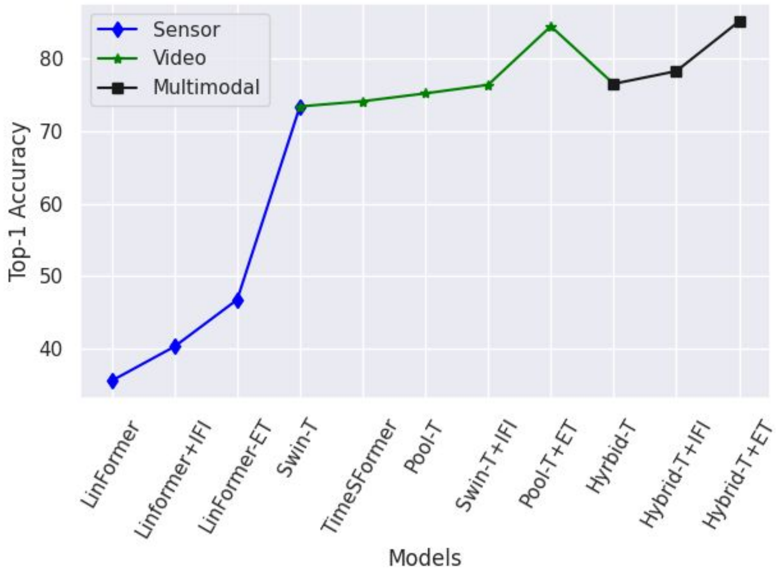
Results: Multimodal Dataset

Model	Top-1 Acc
TimeSFormer [6]	74.11%
Swin Transformer (Swin-T) [7]	73.39%
Pooling Transformer (Pool-T) [8]	75.19%
Video Swin-T-In (IFI) [2]	76.37%
Pool-T + ET (Ours)	**84.45%**

Table 1. Top-1 test accuracy on the BIRDS dataset for videos

Model	Top-1 Acc
LinFormer [9]	35.55%
LinFormer-In (IFI) [2]	40.26%
LinFormer-ET (Ours)	**49.09%**

Table 2. Top-1 test accuracy on the BIRDS dataset for signal modality



Comparison with the different modalities and models. **Hybrid: Multimodal training**

- Multimodal Training has 75.41%, with IFI 78.26%, and with ET has an accuracy of **85.12%** which is an increase of ~ **8.6%** and ~ **7%**
- **ET** thus improves for the video, signal and the multimodal training

Conclusion

- The ET framework that we proposed is able to improve the performance of training by guiding the network to focus only on the salient regions in the input
- ET can be combined with other XAI methods but we used it with our method Rollout-FEM and trained Transformer based models for an image, video and signal dataset
- ET shows promise with both the single modality and multi modality.
- Input pruning, by setting certain features to zero during frame cropping in videos, could reduce computation, training time and improve generalization when fine-tuning on different datasets.

"Scan this QR code to access our code for the Explain to Train (ET) framework."