Multimodal Vision Transformers with Forced Attention for Behavior Analysis (Supplementary Material)

Tanay Agrawal
INRIA
Valbonne, France

Michal Balazia
INRIA
Valbonne, France

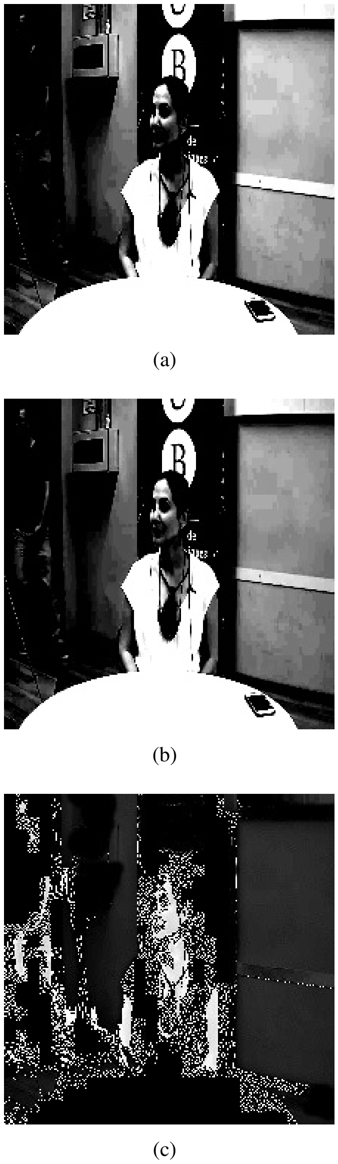
Philipp Müller
DFKI
Saarbrücken, Germany

François Brémond
INRIA
Valbonne, France

tanay.agrawal@inria.fr https://github.com/Parapompadoo/FAt-Transformers

A. Effect of segmentation map on attention:

Attention in the full frame sequence is visualised in Figure 1, with and without segmentation map. The first FAt transformer block is used to generate the figure. We see apparent change in attention in the figure when segmentation map is provided as an additional input to the model. Figure 1 (b) has a brighter face and a more dull background which shows more attention to the important parts even among the bright parts of the segmentation map as face is known to have the most information. An interesting phenomenon is that it also has a brighter person in the background which is not provided as input with the segmentation map as there is only one person in the segmentation mask. Figure 1 (c) highlights the attended parts in the image. So the model not only learns to attend to the foreground provided to it in the input but also find salient parts of the background which relate to the foreground and performs better as discussed in the ablation study above.



(a)

(b)

(c)

Figure 1. Example of attention (a) without segmentation map, (b) with segmentation map and (c) their difference.

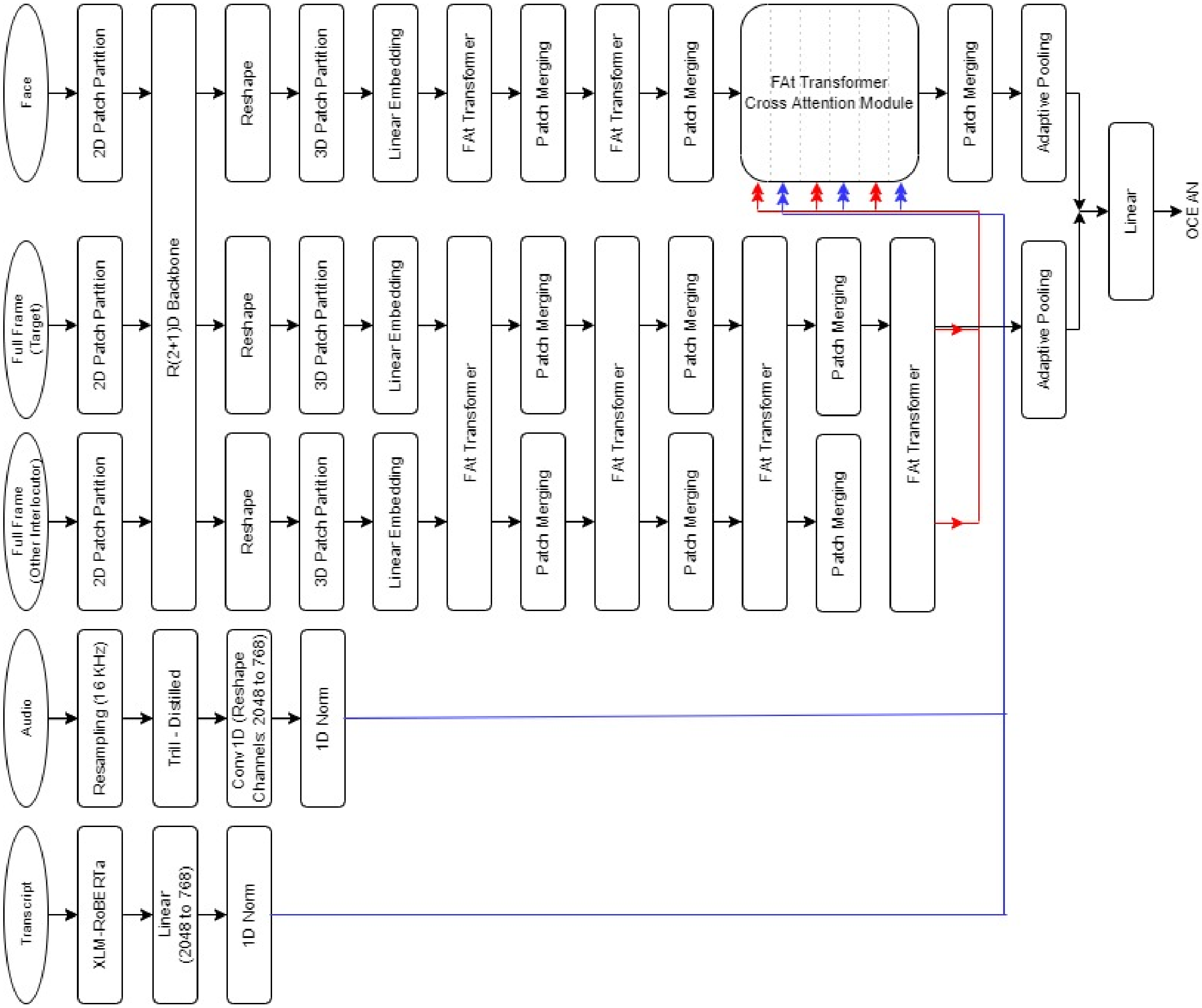


Figure 2. Model architecture for Udiva v2. Segmentation map is input into each FAt transformer module and is not shown here to reduce complexity.

B. Architecture

The architecture of the entire pipeline is displayed in Figure 2. The inputs to the model are: face crop (128 frames of 224 × 224 resolution), full frame (128 frames of 896 × 896 resolution) of the target and the other interlocutor and the audio and transcript cropped to align with the input video modality. Details about the shape after each layer for the face branch are given in Figure 1 of the main text.

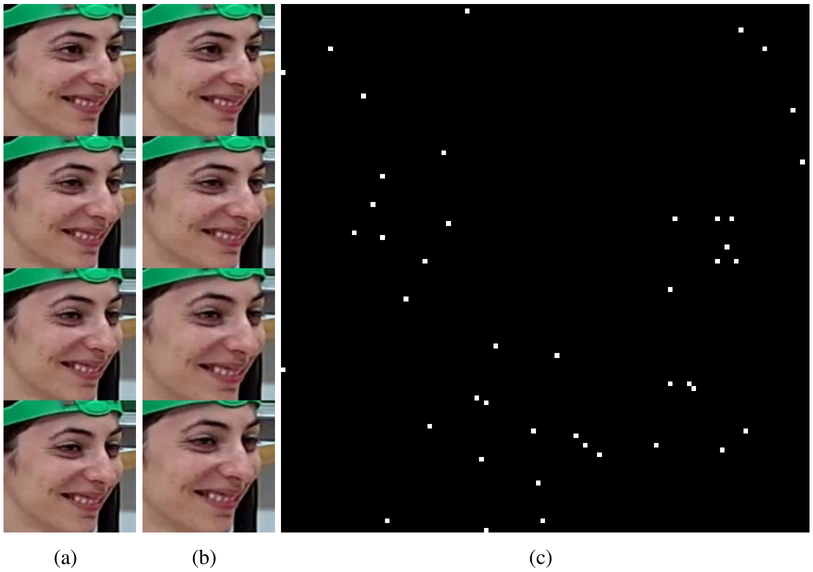


Figure 3. (a) Input without noise (frame 1 on top, frame 4 at the bottom); (b) Input with transformation noise (frame 1 is stretched vertically, frame 2 is stretched horizontally, frame 3 and 4 are cropped and resized); (c) Visualisation of the difference between the activation after the CNN backbone with and without noise.

C. Noise Handling by CNNs

There exists transformation noise in our input, specifically in the face branch. Figure 3 is a visualisation for the noise reduction capability of the CNN backbone for processing input for the transformer part of the model. The size of embedding after the CNN backbone for one time chunk (4 frames) is $256 \times 112 \times 112$ ($C \times H \times W$). We average over the channel dimension and visualise the difference of activation with and without noise (Figure 3 (b) and (a) respectively). It can be seen that very few spatial locations have a difference even though there is variation in the input. We can see that most of the change is around the boundaries and salient parts of the image which are relevant to our task, for example, lips.