Frame Attention Networks for Facial Expression Recognition in Videos

- Aim:

To perform video-based FER and classify a video into several basic emotions (Anger, Happiness, Sadness, Disgust, Fear, Neutral etc.).

- Methodology:

This paper proposes **Frame Attention Networks** (**FAN**). The FAN has two modules:

Feature Embedding – A deep CNN to embed each face image frame in the video into a feature vector (f_i); **Feature Attention** – To learn self-attention and relation-attention wts. to adaptively aggregate the feature vectors and obtain a single video representation vector.

Self-attention wts. – FC layer (q^0 parameters) and a sigmoid function applied to individual frame features and the obtained coarse attention wts. (a_i) are then aggregated to obtain a global representation (f'_v) as follows.

$$\alpha_i = \sigma(f_i^T \mathbf{q}^0)$$

$$f'_v = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i}.$$

- Methodology (Contd.):

Relation-attention wts. – Each individual frame feature is concatenated with the global representation (f'_v), passed through another FC layer (q^1 parameters) and then acted on by the sigmoid non-linearity. The i^{th} relation-attention wt. is given as β_i . Finally, the FAN aggregates all the frame features into a new compact feature.

$$\beta_i = \sigma([f_i : f'_v]^T \mathbf{q}^1) \qquad f_v = \frac{\sum_{i=0}^{n} \alpha_i \beta_i [f_i : f'_v]}{\sum_{i=0}^{n} \alpha_i \beta_i}.$$

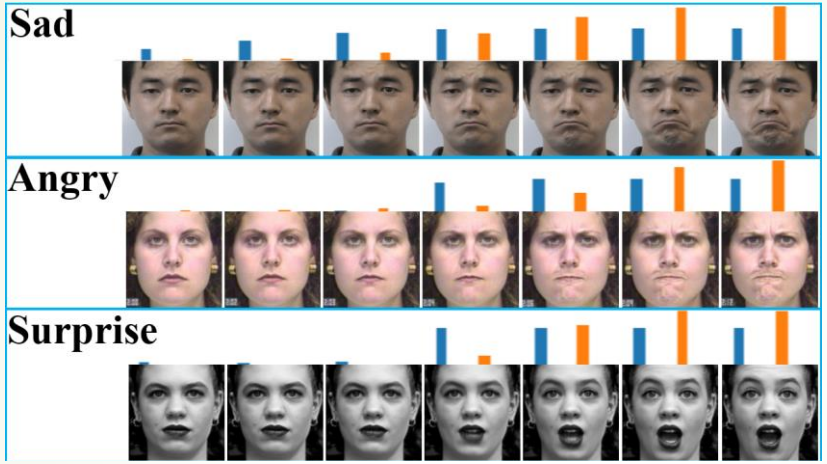
The final feature obtained above is used to then train a basic classifier. The FAN model was trained on the **CK+** and **AFEW 8.0** datasets. The video frames were pre-processed with face detection and alignment. RESNET18 was used for Feature Embedding followed by the Feature Attention Module.

- Results & Conclusion:

	CK+	**AFEW 8.0**
Score Fusion (Baseline)	94.80%	48.82%
FAN (w/o relation-attention)	99.08%	50.92%
FAN	99.69%	51.18%

- Results & Conclusion (Contd.):

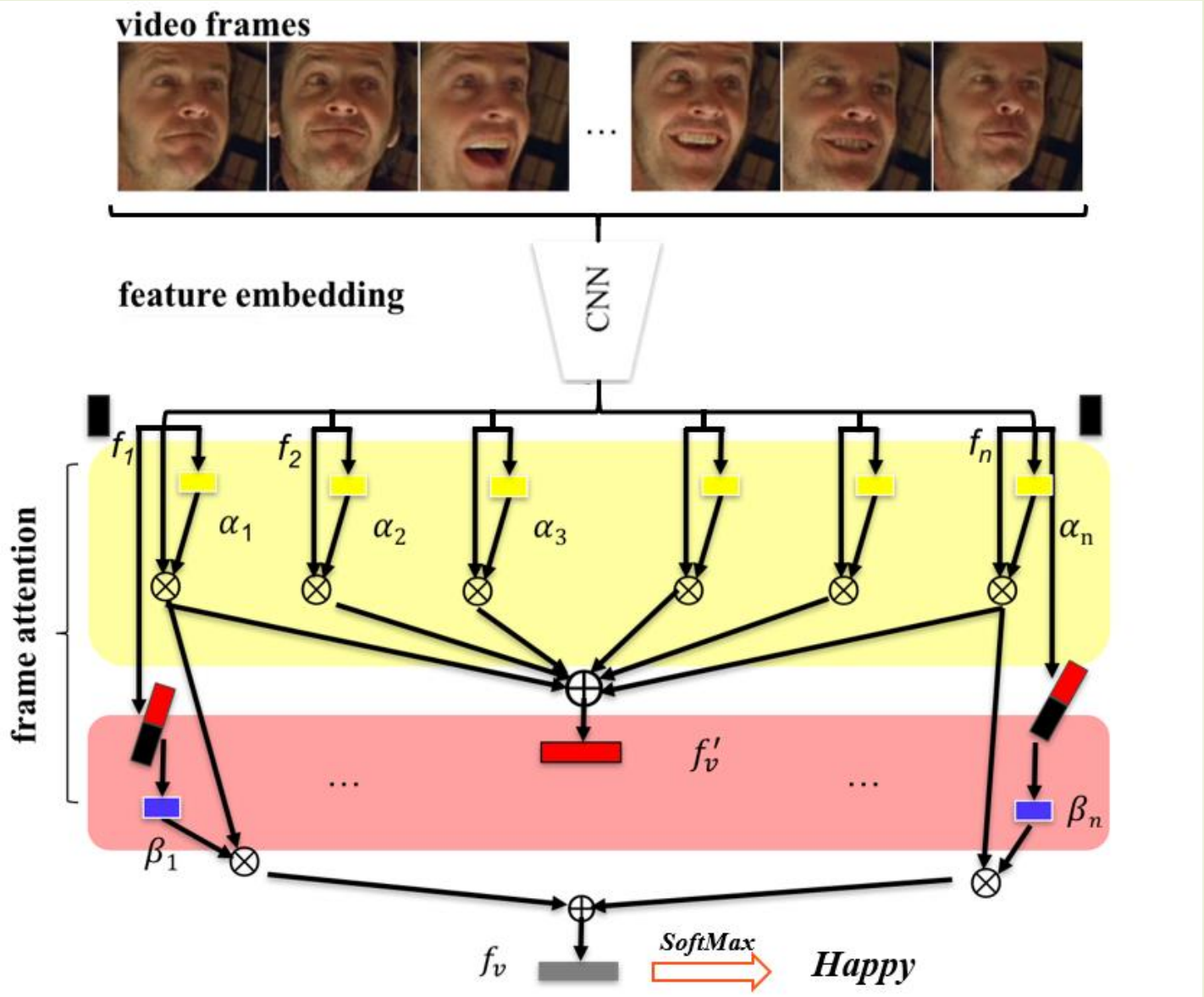
The attention wts. are visualized in the figure:



Blue (self-attention wts. only) & Orange (final wts. of the FAN model)

It is observed that the final wts. of our FAN can always assign higher wts. to the more obvious face frames, while self-attention modules could assign higher values on some obscure face frames.

Through this FAN model, state-of-the-art results were obtained on the CK+ (99.69%) and AFEW 8.0 (51.18%) datasets.



The proposed Frame Attention Network as a whole.