Attentive Hybrid Feature with Two-Step Fusion for Facial Expression Recognition

Jun Weng, Yang Yang, Zichang Tan, Zhen Lei

{jun.weng, yang.yang, zichang.tan, zlei}@nlpr.ia.ac.cn

Technically Co-Sponsored by IEEE COMPUTER SOCIETY

Motivation

- Facial expression recognition is inherently a challenging task, especially for the in-the-wild images with various occlusions and large pose variations, which may lead to the loss of some crucial information.
- Fusing different face regions features by using a simple sum or concat may hardly capture the latent correlations among those regions.

Contribution

- Attentive Hybrid Architecture (AHA) employs separate feature losses to encourage high attention weights for the most important regions and a large margin cosine loss for discriminative features in the whole network.
- We introduce a two-step fusion strategy to capture the hidden relations among different face regions.
- The new state-of-the-art performance on CK+, FER-2013, SFEW2.0 and RAF-DB datasets.

Attentive Hybrid Architecture (AHA)

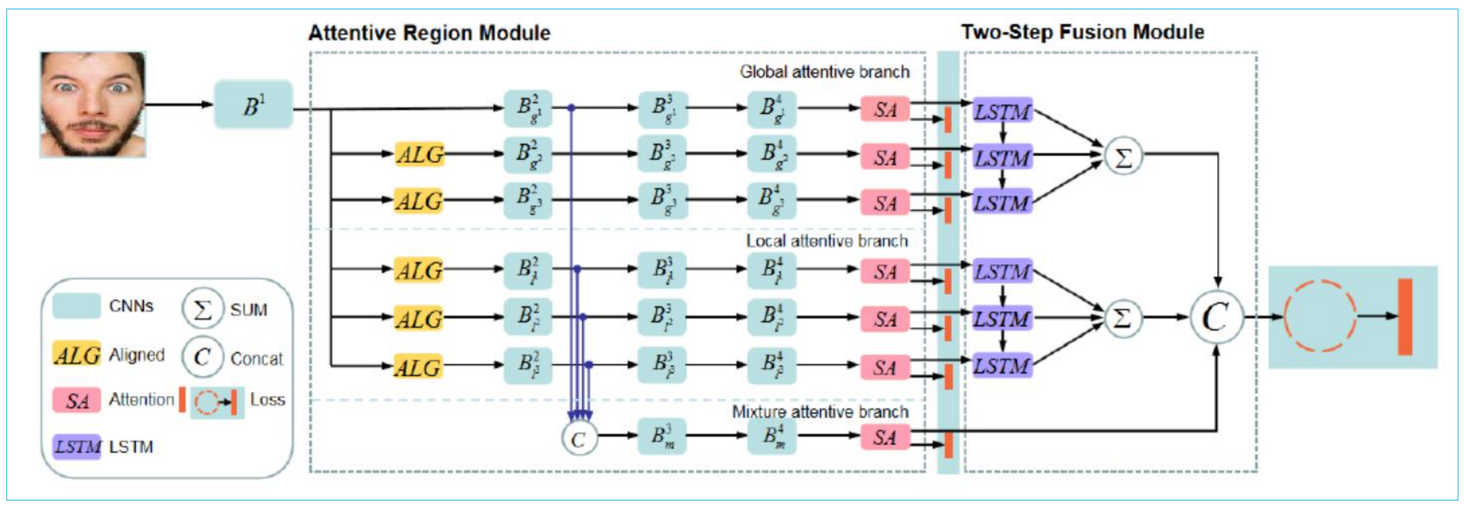
Attentive Region Module

➢ Aim to extract attentive hybrid features.

① A region align (ALG) component to generate aligned face regions , which helps the network to deal with large pose variations.

② Then, those face regions are fed into several parallel sub-networks to extract hybrid features.

③ After that, the spatial attention module is further employed to extract discriminative features.



The Loss Function

- The total loss of training the whole network:

$$L = -\frac{1}{n}\sum_{i=1}^{n} log(p_{y_i}(x_i)) - \frac{1}{n}\sum_{i=1}^{n}\sum_{\varepsilon\in\Phi}\sum_{k=1} log(p_{y_i}^{\varepsilon_k}(x_i))$$

Two-Step Fusion Module:

➢ Aim to capture the correlations among different regions.

① The first step isto explore the correlation among different region features by a recurrent fusion conducted on its own feature, which is called recurrent fusion step.

$$h_i^{l_k} = LSTM(z_i^{l_k}, h_i^{l_{k-1}})$$

$$h_i^{l} = \sum_{k=1}^{3} h_i^{l_k}$$

② The second step seeks to find the final discriminative features for facial expression recognition by concatenating these three features from the first step, namely sum operation step.

$$x_i = concat(h_i^g, h_i^l, x_i^m)$$

Experiments

State-of-the-art performance on RAF-DB, SFEW 2.0, FER-2013 and CK+ datasets

TABLE I
THE ANALYSIS OF ATTENTIVE REGION MODULE ON RAF-DB AND SFEW 2.0. GLOBAL(G), LOCAL(L), MIXTURE FEATURE(M), SPATIAL ATTENTION(A), SOFTMAX LOSS(S) AND COSFACE LOSS(C).

Branch	RAF-DB		SFEW 2.0	
	S	C	S	C
G	88.01	88.17	54.04	55.43
G+A	88.17	88.23	54.27	56.81
G+L	88.1	88.49	55.2	57.74
G+L+A	88.14	88.53	55.89	57.27
G+L+M	88.72	88.92	57.27	58.20
G+L+M+A	88.85	88.98	58.2	58.89

TABLE II
THE ANALYSIS OF TWO-STEP FUSION MODULES ON RAF-DB AND SFEW 2.0. SIMPLE FULL FEATURE CONCATENATION(CONCAT) AND TWO-STEP FUSION(TWO-STEP).

Network	RAF-DB		SFEW 2.0	
	concat	two-step	concat	two-step
G+C+A	87.84	88.23	56.35	56.81
G+L+C+A	87.91	88.53	56.58	57.27
G+L+M+C+A	88.27	88.98	56.81	58.89

TABLE III
THE COMPARISONS ON RAF-DB, SFEW 2.0, FER-2013 AND CK+.

Model	RAF	SFEW 2.0	FER-2013	CK+
DLP-CNN [24]	74.2	51.05	-	95.78
LTNET [28]	86.77	58.29	-	92.45
Conv. Pooling [3]	87.0	58.14	-	-
DAM-CNN [4]	-	42.3	66.2	95.88
Shao et al. [5]	-	-	71.14	95.29
Gan et al. [6]	86.31	55.73	73.73	-
ACNNs [10]	85.07	52.59	-	97.03
RAN [33]	86.90	56.4	-	-
Our Model	**88.98**	**58.89**	**73.84**	**97.86**

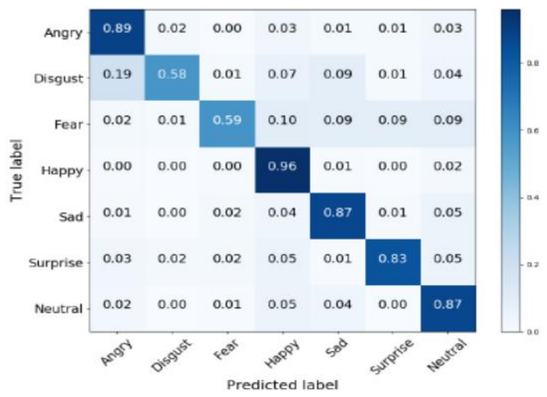


Fig. 4. The confusion matrix on RAF-DB.

Visualizations of the attention maps generated on RAF-DB

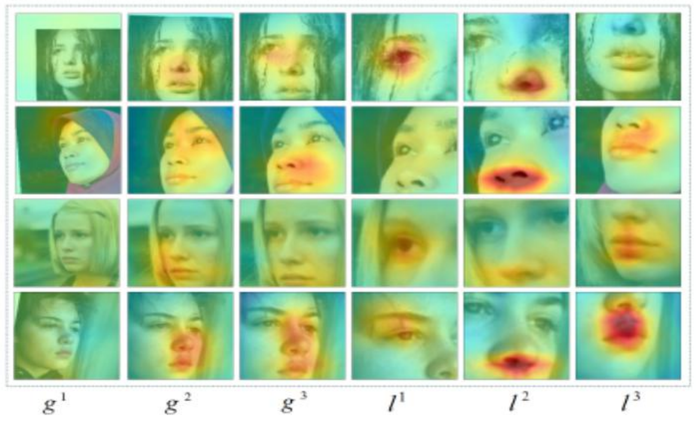


Fig. 5. Visualization of the attention maps generated on the RAF-DB.