Weakly Supervised Local-Global Relation Network for Facial Expression Recognition

The basic block diagram of our paper is as follows:

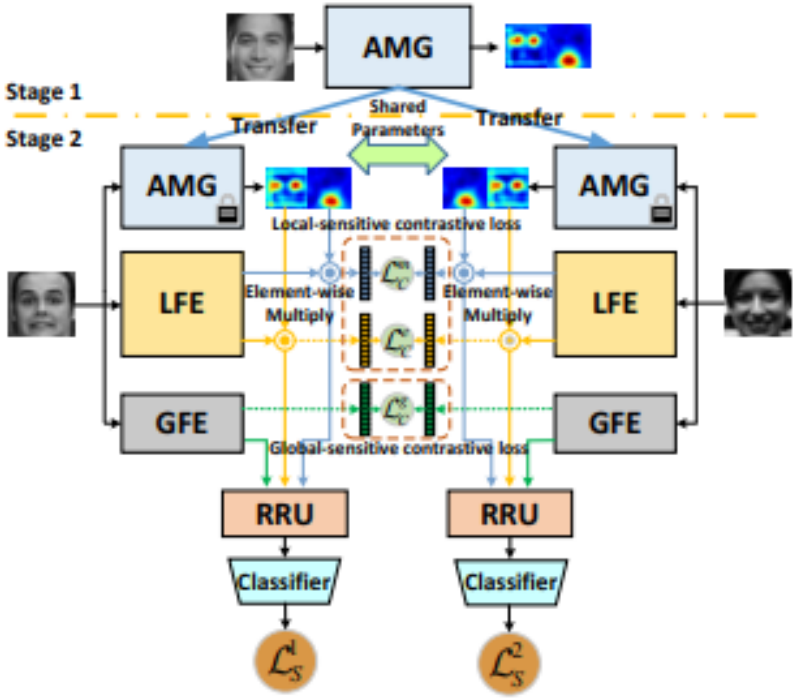


Figure 2: Overview of the proposed framework.

The framework described in the paper has two stages; stage 1 consists of the AMG(Attention Map Generator), as shown in the figure. The existing Facial Expression Datasets do not have eyes and mouth regions highlighted in the dataset's images, so making a dataset with these annotations is also a very time-consuming task. To overcome this issue, the paper proposes an **AMG**(Attention map generator) which consists of a densenet. The AMG acts on the Celeb-A dataset, which has a csv file comprised of the attributes of the images. We use a densenet to identify the eyes and the mouth regions in a Weakly Supervised learning algorithm proposed in the paper. Once the model learns , we then freeze the weights and transfer these weights to our facial expression dataset(the paper uses CK+, we planned to use **FER 2013**). In the second stage of the framework, there are two branches; as shown, the paper uses **LFE**(Local Feature Extraction) and **GFE**(Global Feature Extraction) but doesn't specify the implementation details of LFE and GFE; we used SIFT for LFE and HOG for GFE. The outputs of the first stage and LFE Feature maps are element-wise multiplied and connected to a fully connected layer, and the loss function is calculated based on whether the two images from the two different branches are of the same expression or not.

Similarly, the outputs of the GFE are connected to a fully connected layer, and a global sensitive contrastive loss is calculated. The element-wise multiplied feature maps and attention maps of

eyes and mouth and the GFE feature map are then inputted into the RRU(Relational Reasoning Unit) , which calculates the weights for classification and calculates the total loss of the framework. The output of the RRU is then fed to the classifier, which gives us the expression of the image.

Contributions:
Our research paper did not have an existing public code repository. We tried to implement the whole paper by ourselves. We implemented the first stage of the paper, which is AMG(Attention Map Generation) exactly as the paper described it. We did not get the Attention map that the paper claims to get. The paper also uses LFE(Local Feature Extraction) and GFE (Global Feature Extraction) but doesn't specify how it is done in this particular case. There are various methods to do LFE and GFE like LFNet,SIFT, SURF,etc and GFNet,HOG(Histogram Oriented Gradients),etc. We used SIFT for LFE and HOG for GFE, but could not use it after that due to lack of good Attention Maps.

Results:

- We got around 53% for eyes and 58% for mouth training accuracy of the AMG, and the attention maps that we got were pretty random.
- We highlighted the Local Features using SIFT and Global Features using HOG(Histogram Oriented Gradients) in our Colab Notebook
- The paper claims to use a Fully Connected(FC) layer and various loss functions after LFE and GFE, the paper neither specifies whether a neural net is used nor the architecture used for LFE and GFE.

Salient Features: None.

These are the training results we got:
EYE:

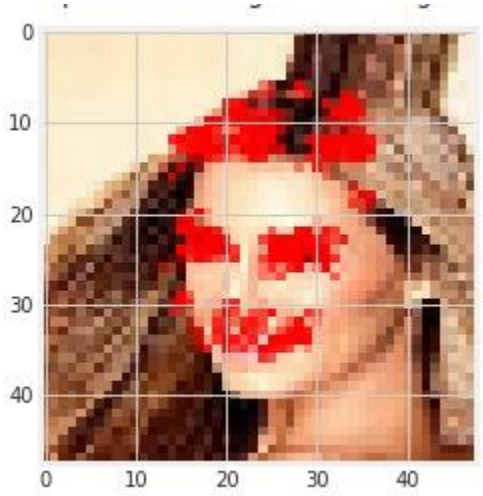
epoch	LR	tr-BCE	tr-Acc	tr-AUC	val-BCE	val-Acc	val-AUC
0	0.1000	0.6937	50.62	0.5231	0.6925	52.20	0.5452
1	0.1000	0.6926	51.46	0.5322	0.6887	54.60	0.5751
2	0.1000	0.6920	51.95	0.5361	0.6889	53.53	0.5755
3	0.1000	0.6919	51.83	0.5367	0.7038	49.67	0.5901
4	0.1000	0.6920	51.82	0.5354	0.6941	49.77	0.5828
5	0.1000	0.6927	51.42	0.5288	0.6928	50.03	0.5406
6	0.1000	0.6926	51.47	0.5307	0.6902	52.23	0.5687
7	0.1000	0.6927	51.49	0.5310	0.7060	50.10	0.5683
8	0.1000	0.6933	50.68	0.5207	0.6915	52.63	0.5392
9	0.1000	0.6928	51.34	0.5261	0.6930	50.73	0.5423

execution-time of function "train_model": 3h 3m 17s

Mouth:

epoch	LR	tr-BCE	tr-Acc	tr-AUC	val-BCE	val-Acc	val-AUC
0	0.1000	0.6921	52.15	0.5355	0.7095	35.57	0.5502

This is what we got from our LFE: (same as LFE*AGM)



This is what we got from our GFE: