Facial Emotion Recognition: State of the Art Performance on FER2013

- Aim:

Perform experiments with **optimizers** and **learning rate schedulers** for FER with a single network on the FER2013 dataset introduced at the International Conference on Machine Learning (ICML) in 2013.

- Methodology:

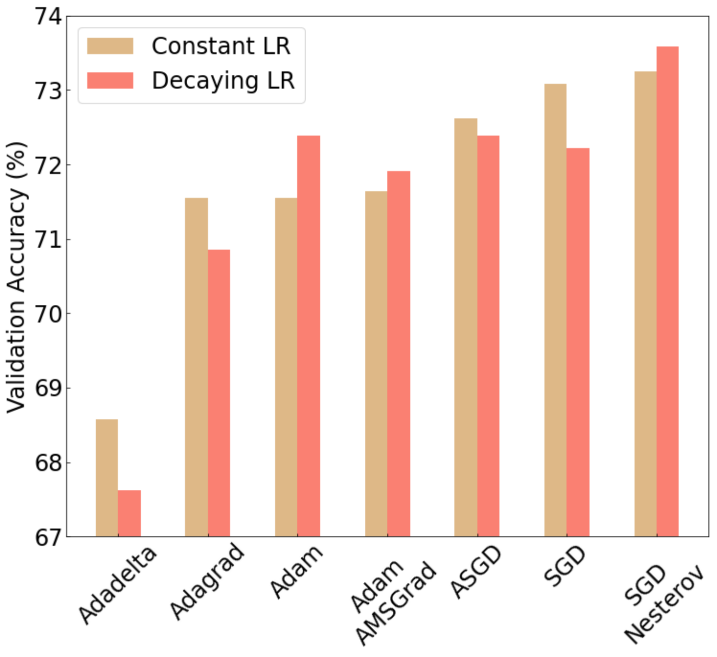
- Images rescaled to ± 20%, shifted horizontally and vertically by ± 20% of original size and rotated up to ± 10° with a 50% probability. Each image is then ten-cropped to 40x40 size and random portions of each crop are erased with 50% probability. Each crop is normalized by dividing by 255.
- The **VGGNet** architecture with Convolution, ReLU, Batchnorm, Max pooling and FC layers is trained for 300 epochs for each experiment optimizing the cross-entropy loss using a fixed momentum (0.9) and weight decay (10^{-4}).
- Grid search is used to determine the optimal batch size and drop-out rate.
- 6 diff. optimizers are experimented with: SGD, SGD with Nestrov, Avg. SGD, Adam, Adam w/ AMSGrad, Adadelta and Adagrad.

- Methodology (Contd.):

- 2 diff. variations of the experiment are run. The 1st one with constant learning rate of 0.001 for all the optimizers. The 2nd with a learning rate scheduler with initial learning rate of 0.01, reduced by 0.75 if val. acc. plateaus for 5 epochs.
- The next experiment is to find the optimal learning rate scheduler out of RLRP, Cosine Annealing, Cosine WR, OneCycleLR, StepLR. All schedulers have an initial learning rate of 0.01.
- The model wts. are finally hyper-tuned for a final 50 epochs with a small learning rate of 0.0001. This is run using Cosine and Cosine WR. Another variation of this experiment is run after combining the val. and train datasets to allow for the model to train on a larger amt. of data.

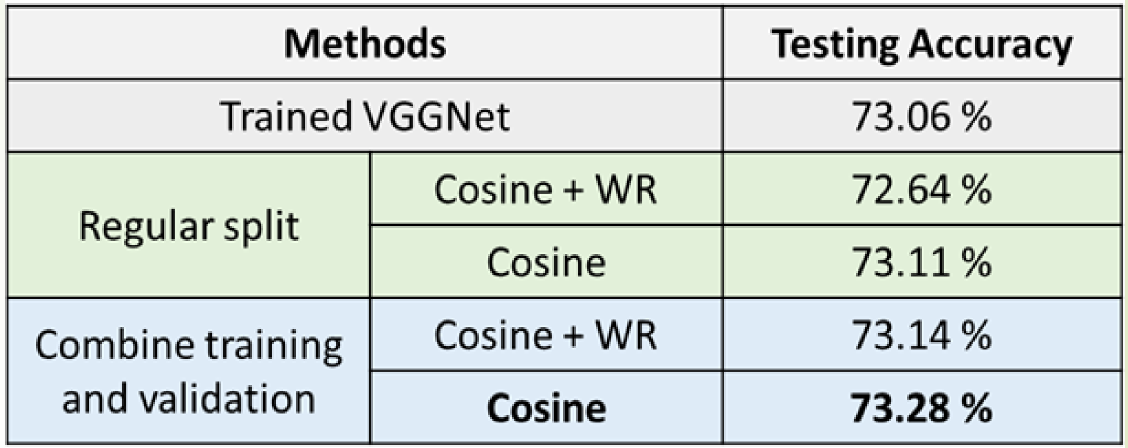
- Results:

- As far as the optimizers are concerned, the model with SGD + Nestrov performs the best.



Results (Contd.):

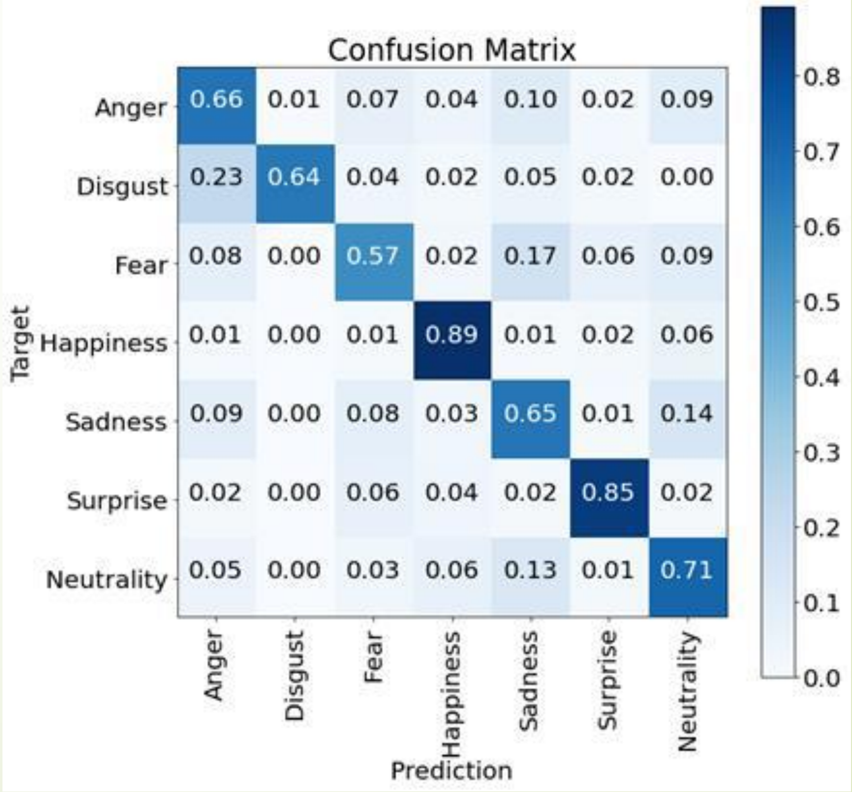
- The RLRP learning rate scheduler performs best with SGD + Nestrov. The test acc. at this point is 73.06%.
- Shown here are the results through fine-tuning.



Methods		Testing Accuracy
Trained VGGNet		73.06 %
Regular split	Cosine + WR	72.64 %
	Cosine	73.11 %
Combine training and validation	Cosine + WR	73.14 %
	Cosine	**73.28 %**

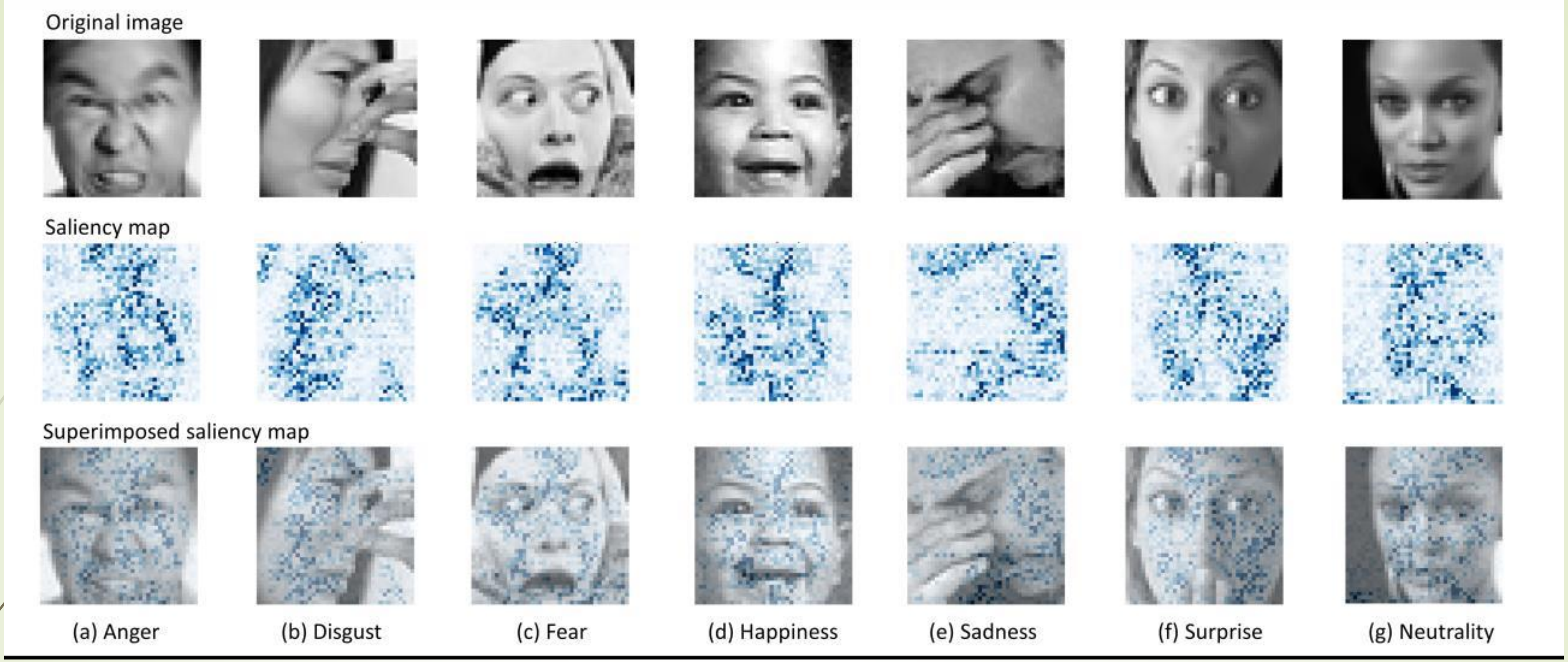
Conclusion:

- Through the confusion matrix, it is seen that the model performs best on ‘happiness’ and ‘surprise’ and worst on ‘disgust’ and ‘anger’. The misclassification b/w ‘fear’ and ‘sadness’ can be attributed to inter-class similarities.
- The model places a large importance on almost all facial features, seen in (f) where the saliency map (next slide) almost perfectly maps the entire face. The model also effectively drops non-informative regions like the background in (a), (d), (g), hair in (a), (c), (g), and the hand in (e). Some mistakes in the saliency maps are seen in (b), (e), (g) where the model highlights some of the background pixels. A model that can more effectively identify facial features and drop all useless information will perform better.



Confusion Matrix

Target \ Prediction	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutrality
Anger	0.66	0.01	0.07	0.04	0.10	0.02	0.09
Disgust	0.23	0.64	0.04	0.02	0.05	0.02	0.00
Fear	0.08	0.00	0.57	0.02	0.17	0.06	0.09
Happiness	0.01	0.00	0.01	0.89	0.01	0.02	0.06
Sadness	0.09	0.00	0.08	0.03	0.65	0.01	0.14
Surprise	0.02	0.00	0.06	0.04	0.02	0.85	0.02
Neutrality	0.05	0.00	0.03	0.06	0.13	0.01	0.71



- This model gives the highest single network acc. ever recorded at 73.28% (state of the art) using transfer learning with the VGGNet architecture, marginally better than estimated human performance (~65.5%) on the same dataset.