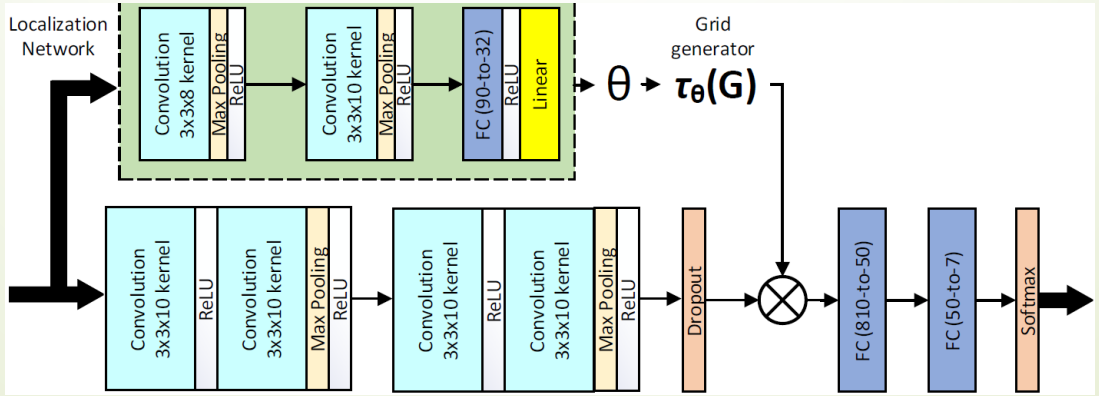
Deep-Emotion: FER with Attentional CNN

- **Aim:**

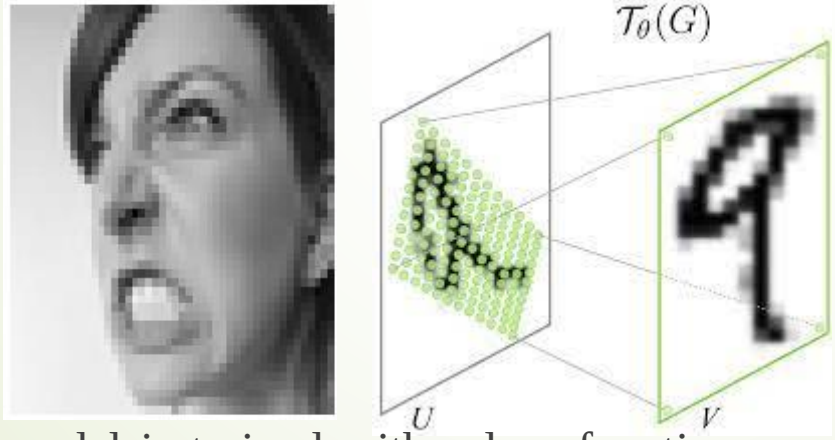
Perform FER using the proposed Attentional Convolutional Network.

- **Methodology:**

  - Facial Expressions are determined by only certain features of a face e.g. mouth and eyebrows and not so much by others e.g. hair and ears. Taking this observation into consideration, an end-to-end deep learning framework is proposed employing a Spatial Transformer Network (STN) Module which learns to focus on only the relevant parts of the face for emotion detection.
  - The model architecture comprises of a feature extraction part consisting of four convolutional layers, each two followed by a max-pooling and ReLU layer. These are then followed by a dropout and two FC layers. The spatial transformer (localization network) consists of two convolutional layers, each followed by a max-pooling and ReLU layers and finally two FC layers.



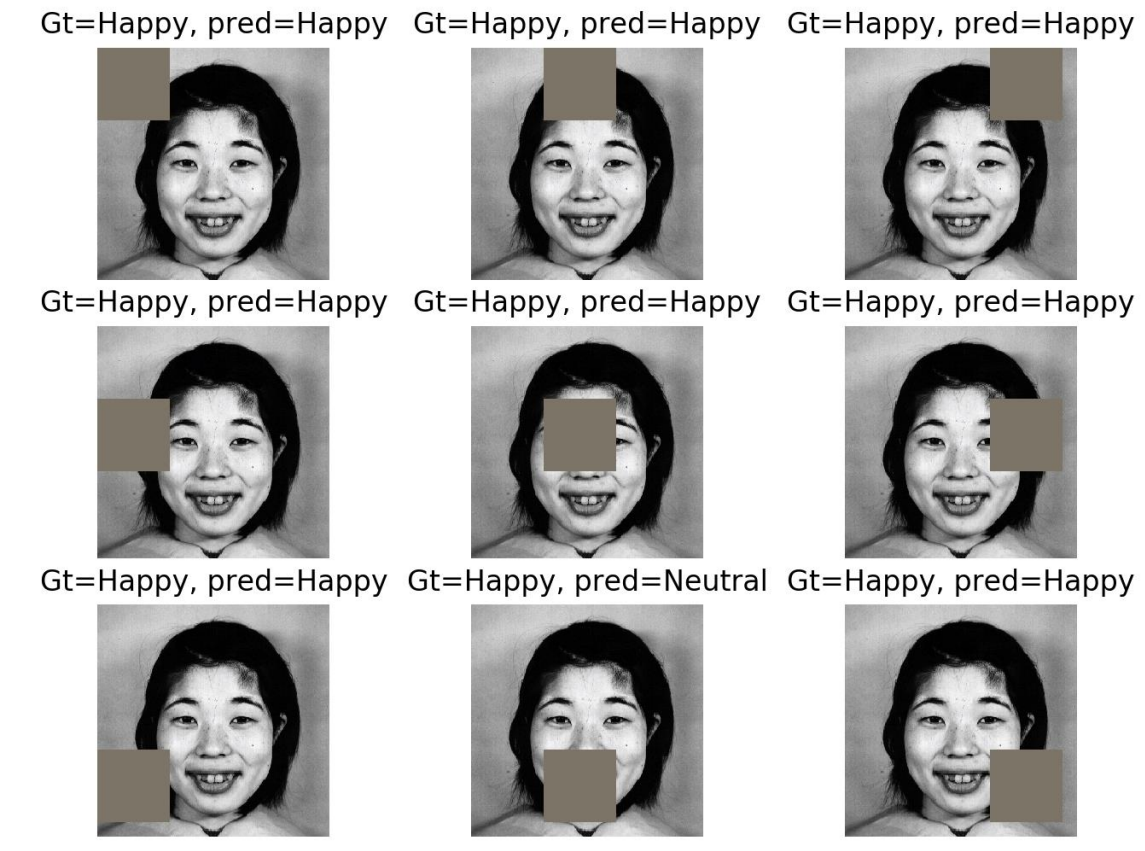
- The Spatial Transformer learns to produce (by backpropagating the localization net) a transformation matrix Θ corresponding to each new input image (U). A mesh grid comprising of a set of (x, y) coordinates is generated by the Grid Generator of the STN and the transformation matrix Θ is applied to this mesh grid (applied here means matrix multiplication is carried out) to produce a new grid $T_\Theta(G)$ comprising of a new set of sampling points. Finally, using the initial input image, transformed mesh grid and a differentiable interpolation function, the sampled output image is produced.
- The sampled output (V) produced by the STN module is essentially translated/rotated/warped in such a way as to make the entire model more robust to input images taken in uncontrolled environments wherein the subjects face is, say, posed at an angle or occluded. Therefore, the STN module tries to remove spatial invariance from images, focusing only on the most important features of the input image.



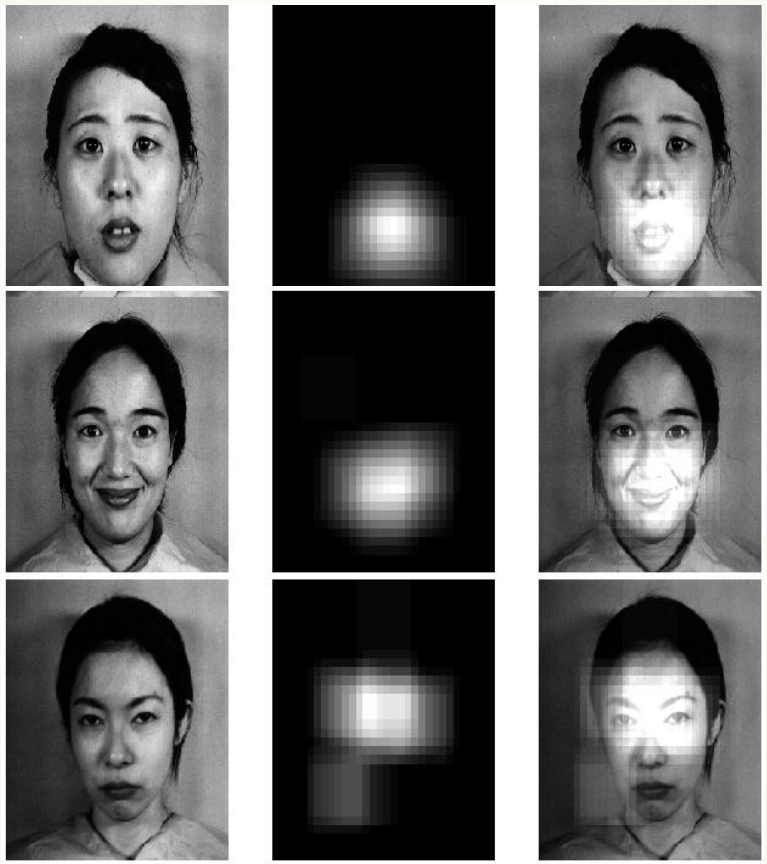
- Finally, the model is trained with a loss function, comprising of a classification and a regularizing term.

$$\mathcal{L}_{overall} = \mathcal{L}_{classifier} + \lambda \|w_{(fc)}\|_2^2$$

- In order to visualize the important regions for different emotions, square regions of size N x N inside the image are zeroed out starting from the top left corner of the image. If occluding a certain region causes the model to produce the wrong label, that region is considered to be important in determining that emotion, else not. As is observed, zeroing out regions around the mouth for the ‘Happy’ emotion, causes the model to make incorrect predictions.



- The detected important regions by the proposed model for different emotions shown through saliency maps:



- **Results:**
  - 4 datasets: FER13, JAFFE, FERG and CK+ were used. One model was trained per dataset and the architecture and hyper-parameters were kept constant across. State of the art results were obtained in 3 out of the 4 chosen datasets.
  - Performances obtained are:

BLE I: Classification Accuracies on FER 2013 da

Method	Accuracy Rate
Bag of Words [52]	67.4%
VGG+SVM [53]	66.31%
GoogleNet [54]	65.2%
Mollahosseini et al [19]	66.4%
The proposed algorithm	70.02%

TABLE IV: Classification Accuracy on CK+

Method	Accuracy Rate
MSR [39]	91.4%
3DCNN-DAP [40]	92.4%
Inception [19]	93.2%
IB-CNN [41]	95.1%
IACNN [42]	95.37%
DTAGN [43]	97.2%
ST-RNN [44]	97.2%
PPDN [45]	97.3%
The proposed algorithm	98.0%

  - As can be seen, the model improves performance on in-the-wild datasets such as FER13 and CK+ owing to the STN module which makes the model more robust to spatial variances.