CVSN Project Report

Arghadeep Ghosh, Ameya Anand Kamat, Shankar Ram Vasudevan

May 2022

Abstract

Facial expressions play an important role in daily human-human communication. Automatic facial expression analysis is an important area of artificial intelligence. Due to its potential applications in various fields, such as intelligent tutoring systems, service robots, driver fatigue monitoring, Facial Expression Recognition (FER) has attracted increasing attention in the computer vision community recently. Occlusion and pose variations, which can change facial appearance significantly, are two major obstacles for automatic Facial Expression Recognition (FER). This project involved building a Region Attention Network (RAN), to adaptively capture the importance of facial regions for occlusion and pose variant FER. The RAN aggregates and embeds varied number of region features produced by a backbone convolutional neural network into a compact fixed-length representation. Inspired by the fact that facial expressions are mainly defined by facial action units, we propose a region biased loss to encourage high attention weights for the most important regions.

1 Introduction

The Region Attention Network (RAN) is built to effectively capture the importance of facial regions for occlusion and pose robust FER. The RAN is comprised of a feature extraction module, a self-attention module, and a relation attention module. The later two modules aim to learn coarse attention weights and refine them with global context, respectively. Given a number of facial regions, our RAN learns attention weights for each regions in an end-to-end manner, and aggregates their CNN-based features into a compact fixed-length representation. Besides, the RAN model has two auxiliary effects on the face images. Cropping regions can enlarge the training data, which is important for those insufficient challenging samples and re-scaling the regions to the size of original images highlighting fine-grained facial features.

2 Preparing the Dataset

To build a dataset for training our RAN we use the popular and publicly available FERPlus dataset. The dataset contains 35710 48x48 size images which

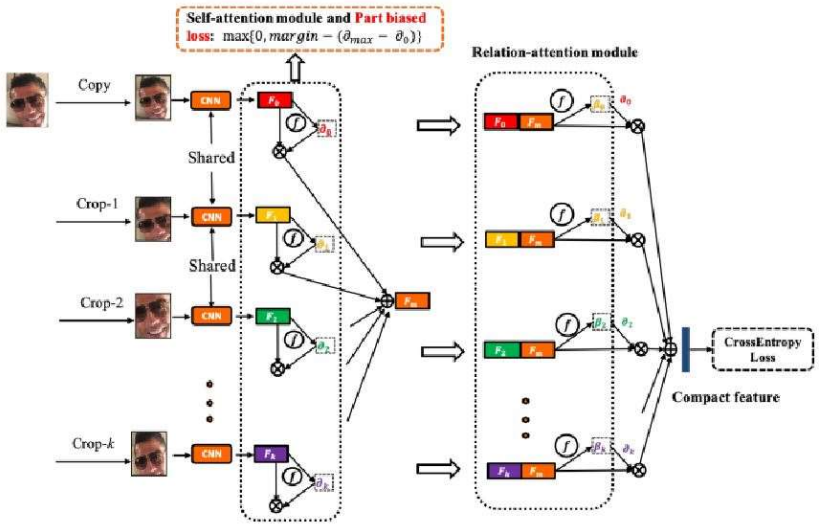

Figure 1: The RAN framework

are classified into nine categories: eight emotions and an unknown, catch-all exceptions category. To avoid overfitting the model to the FERplus dataset, the model is pre-trained on a facial recognition dataset and then fine-tuned to the emotion recognition dataset. We use the python library Dlib 68-point facial landmark detector to identify, extract and rescale faces in complex scenes.

3 The RAN framework

The first module, the Region Cropping module, crops the image into k sub-images to capture the importance of facial regions, which allows for a more robust classification in occluded and pose-variant environments. This module allows the model to adaptively capture the importance of each facial region in the image. In our implementation, we have chosen a methodology that selects and crops five fixed regions from the original image.

The cropped images are then passed through a base CNN which is used to generate feature vectors F_i from each of the cropped regions. The CNN architecture we have chosen is iResNet50. Residual networks are widely used for facial image classification problems since they have "skip-connections" which allows effective training of deep neural networks by tackling the vanishing gradient and accuracy saturation problems. iResNet50 is a residual network which consists of 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer.

The feature vectors extracted by the CNN are passed to the self-attention module. The self-attention module generates a Region Biased loss for each of the cropped regions in the image and adds a constraint that enforces one of the cropped regions has attention weights larger than the original image by a

small margin, which is chosen as a hyperparameter (which is set to 0.1 in our case). The module generates this loss by applying a fully connected layer and a sigmoid function to estimate the weightage of each of the k feature vectors as

$$\mu_i = f(F_I^T \mathbf{q}^0)$$

The features are then combined by taking a weighted average to produce a *summary feature vector* F_m

$$F_m = \frac{1}{\sum \mu_i} \sum \mu_i F_i$$

The second part of the network, the Relation Attention module, takes the individual and summary feature vectors to formulate new attention weights $\nu_i = f([F_i : F_m]^T \mathbf{q}^1)$ for the cropped regions which capture the relation between the individual vectors and the summary vector. This is then aggregated to produce a new compact feature P_{RAN} which is the final representation of the RAN.

$$P_{RAN} = \frac{1}{\sum \mu_i \nu_i} \sum \mu_i \nu_i [F_i : F_m]$$

4 Evaluating the model

The model is trained on the FERplus dataset for 70 epochs, with 32137 training images and 3173 test images.

5 Results and Observations

Upon running the trained network on the validation set, we get an accuracy score of 86%. To verify the robustness of the model in occluded and pose-variant environments, the training procedure was repeated on a subset of the training images which are selected as occluded or pose-variant with the same train-test split. Upon running the network on the new validation set, we get an accuracy score of 81%. The confusion matrices for both the experiments are given below.

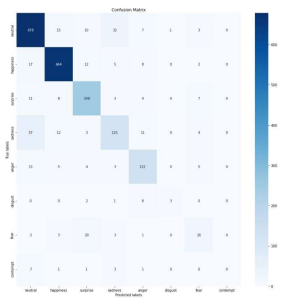

Figure 2: Confusion matrix for full dataset

We observe that the model performs very well on simple emotions such as happiness, surprise, anger and sadness while it struggles to accurately classify complex emotions such as disgust and fear.

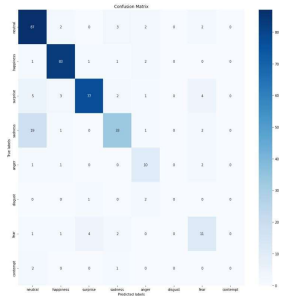

Figure 3: Confusion matrix for occluded or pose-variant dataset