Dana Goghberg, Sapphire Elimelech, Moran Hait

Facial Expression Recognition in Children

Spring 2019

GOALS

The facial expression recognition project is a joint project between the Educational Neuroimaging Center and the Computer Science Department at the Technion.

The goal of the project is to create a facial emotion recognition (FER) system, which will be used by the educational neuroimaging center lab for analyzing children's responses to various tasks, while being monitored by EEG/fMRI.

In addition we wanted to extract and visualize the facial features responsible for every facial expression, and compare them to adults' features.

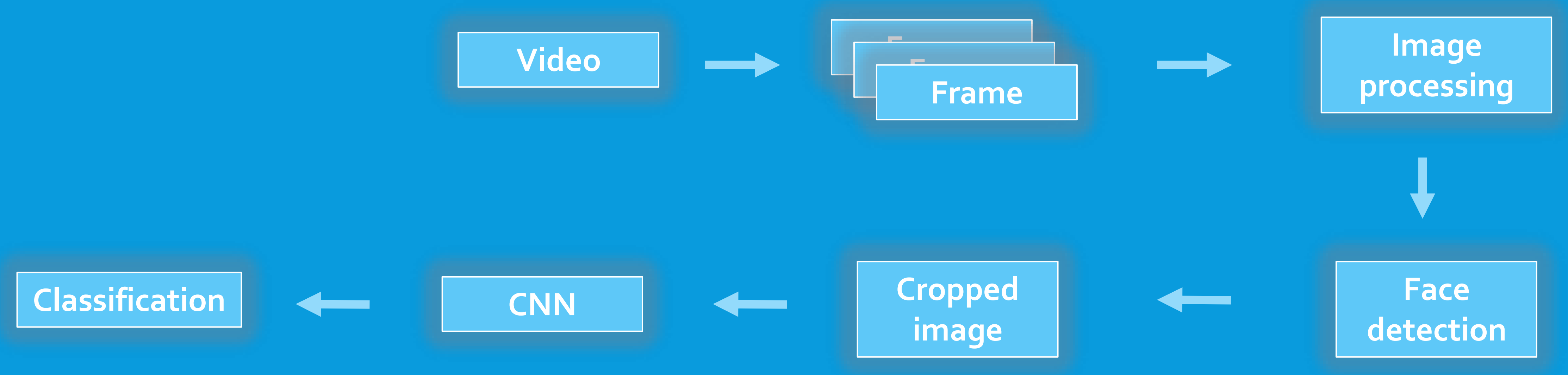
Our model will classify 5 emotions: angry, fearful, happy, sad, neutral.

CHALLENGES

Although well studied on adults, there is only a few FER studies involving children, and therefore only a few small datasets to be found. Another challenge we had to overcome is that different children's emotions representation can be similar, and it can be difficult to differentiate them. We will discuss the methods used to overcome these obstacles in depth later.

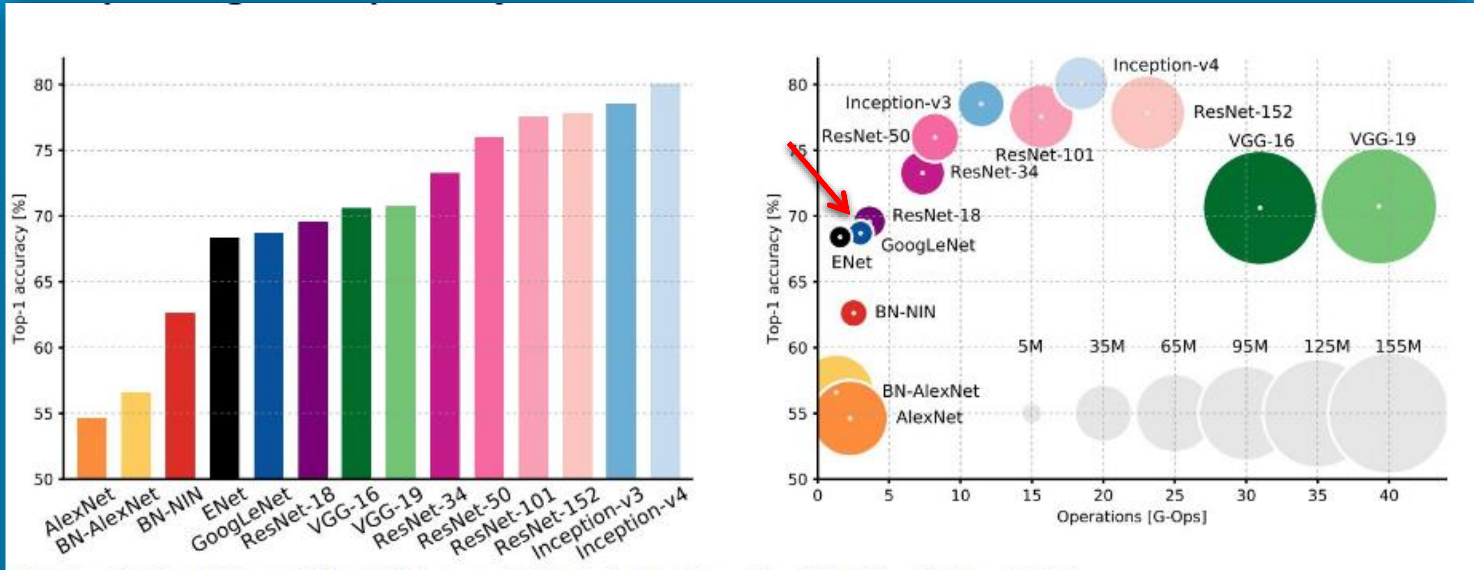
METHOD

Several recent papers were utilizing convolutional neural networks (CNNs) for feature extraction and inference. In our project we use convolutional neural network to implement a new facial emotion recognition system (FER) which is specifically designed for children.



CHOOSING THE MODEL

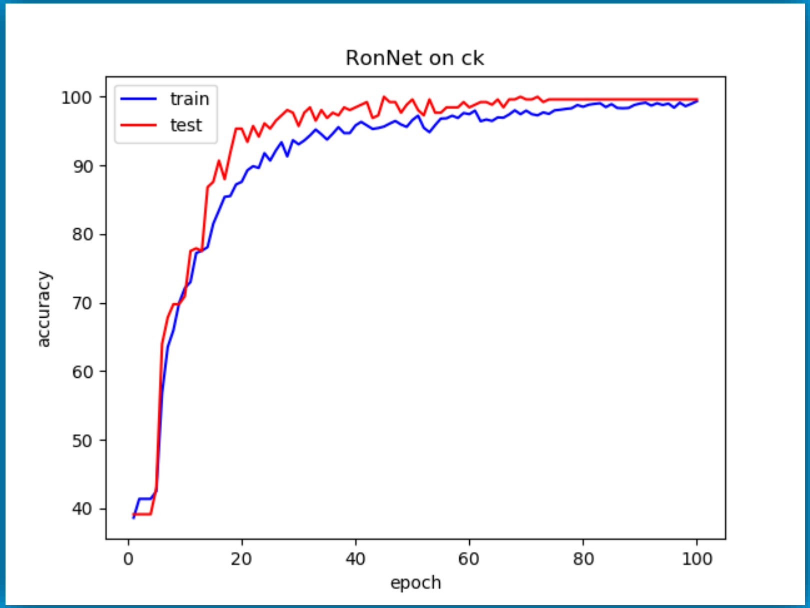
In our project, we compared Ran Breuer's convolutional neural network (**RonNet**) which he used for facial expression recognition in adults as part of his master thesis, "A Deep Learning Perspective on the Origin of Facial Expressions ", to one of the state-of-the-art CNNs used in the ImageNet project. We chose to use **GoogLeNet** due to its good performances and relatively small number of operations.

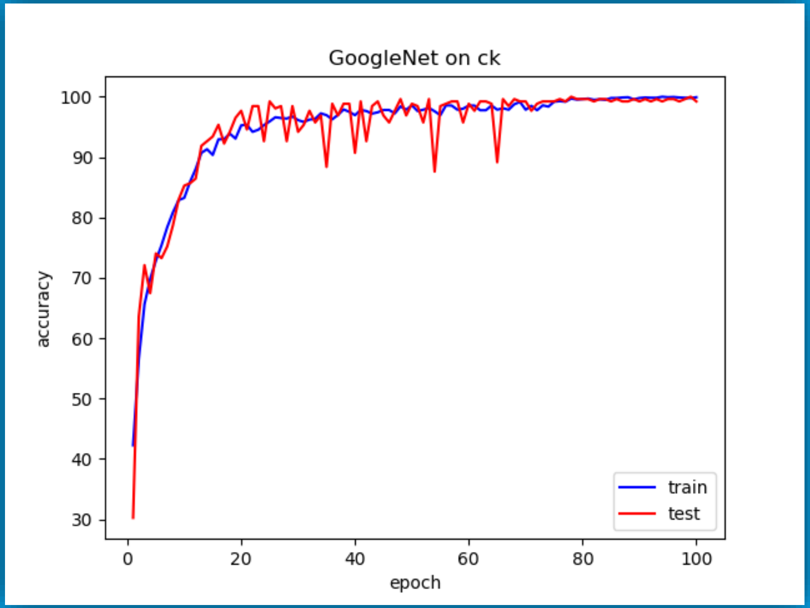


An Analysis of Deep Neural Network Models for Practical Applications, 2017.

CHOOSING THE MODEL – RESULTS

After comparing the accuracy of both **RonNet** and **GooLeNet** on three different datasets (CK+, FER2013, NovaEmotions), neither model proved to be better than the other. We chose to work with **RonNet**, which composed of three convolution layers and two fully-connected layers, due to its simplicity and small number of operations.





CHILDREN'S DATASETS

Educational Neuroimaging Center (ENIC) dataset:

787 Israeli children facial expressions tagged by the ENIC lab staff during EEG scan.

The Child Affective Facial Expression (CAFE) dataset - Databrary:

1180 total photographs taken of 2 to 8 year-old American children.

Dartmouth Database of Children's Faces (DDCF):

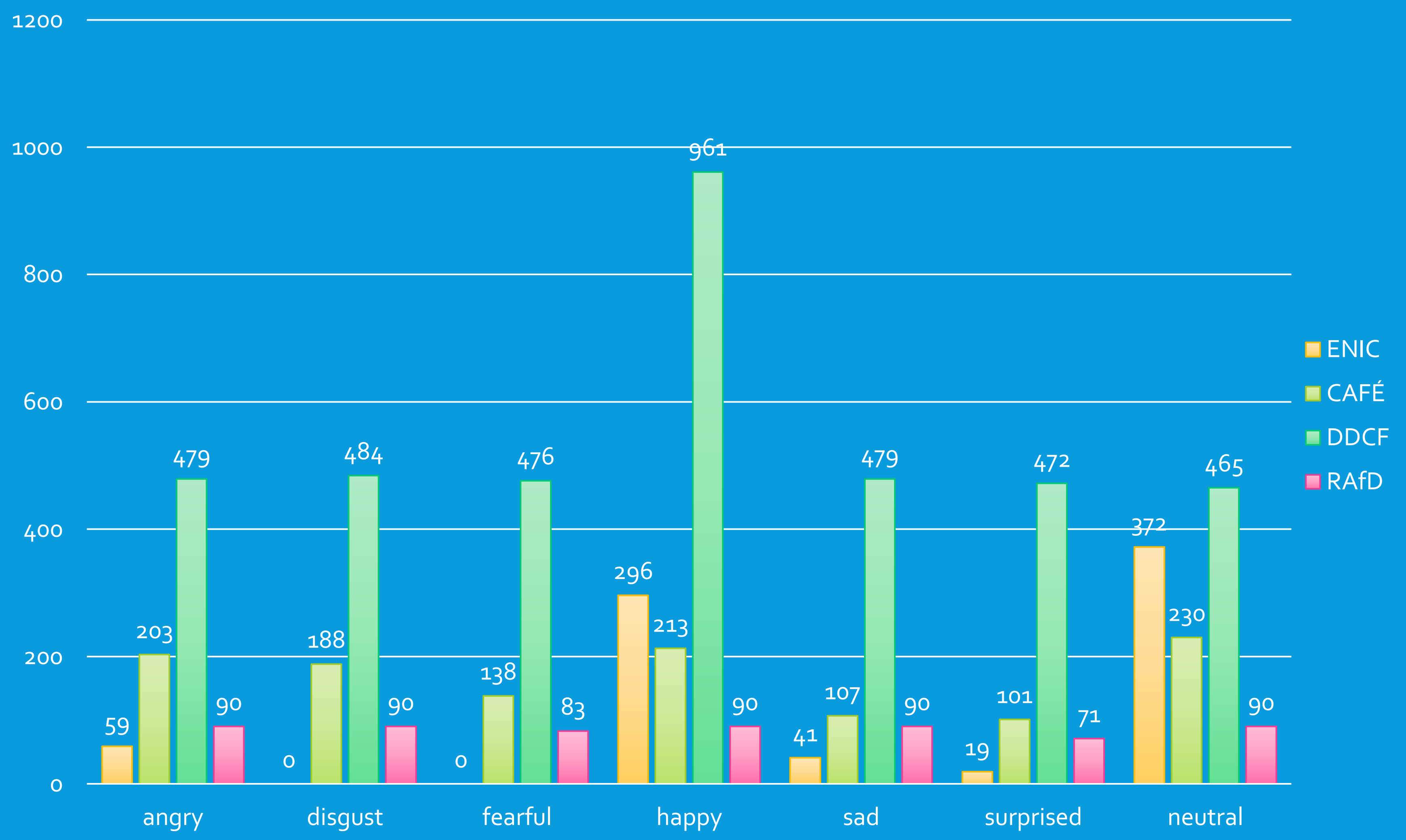
3816 pictures of 6-16 year old children.

Radboud Faces Database (RAfD) dataset:

604 pictures of 10 children (4 boys and 6 girls).

Examples from CAFÉ dataset

EMOTIONS DISTRIBUTION



PRE-PROCESSING

We used several pre-processing methods on the data:

- Face crop: we used a face detection model in order to extract only the face from every image. This method helped filter background noises that might have influenced the model, and helped us gather a relatively uniform dataset even though it is composed of four different sub-datasets.
- Data augmentation: In order to maximize generalization of the model, we used data augmentation methods to generate synthetic data and enlarge the training set: a combinations of **random flips** and **affine transforms**, e.g. **rotation**, **translation**, **scaling** and **shear.**
- RGB vs greyscale: we hypothesized that converting every picture to greyscale will result in higher accuracy since emotions do not correlate with the colors in the image, but we received similar results on RGB and greyscale images.

TRAINING THE MODEL

Our dataset is composed of a relatively small number of tagged children's emotions images (6395 images). We tried three approaches and compared their accuracy on the test set:

- Train the model on adults' datasets and test the accuracy on children's images.
- Train the model on a large adults' emotions dataset (CK+), and then use transfer learning on the children dataset by freezing the convolution layers.
- Train the model solely on the children dataset.

TRAINING ON ADULTS AND TESTING ON CHILDREN

In this approach, we used both models: **RonNet** and **GoogLeNet**, which were pretrained on three different datasets separately: CK+, NovaEmotions and FER2013

Figure 3.2: Images from CK+ (top), NovaEmotions (middle) and FER2013 (bottom) datasets.

RESULTS

Although both models gave decent results on adults images, the accuracy rate for children's emotions were low.

RonNet

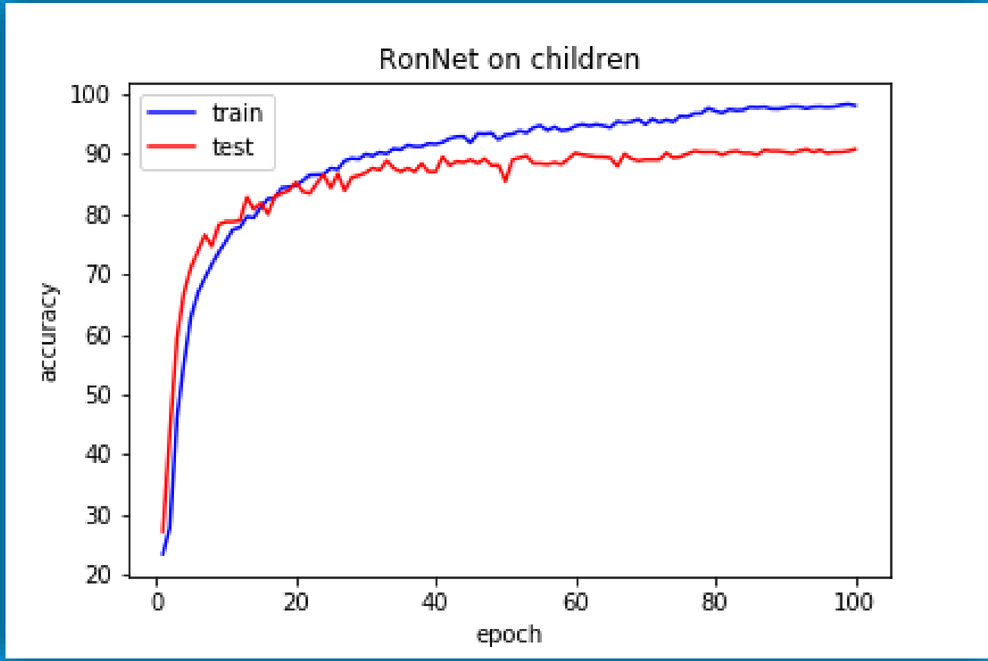
	CK+	FER	Nova	Children
CK+	99.612%	-	-	**25.415%**
FER	-	61.884%	-	**23.754%**
Nova	-	-	83.534%	**19.934%**

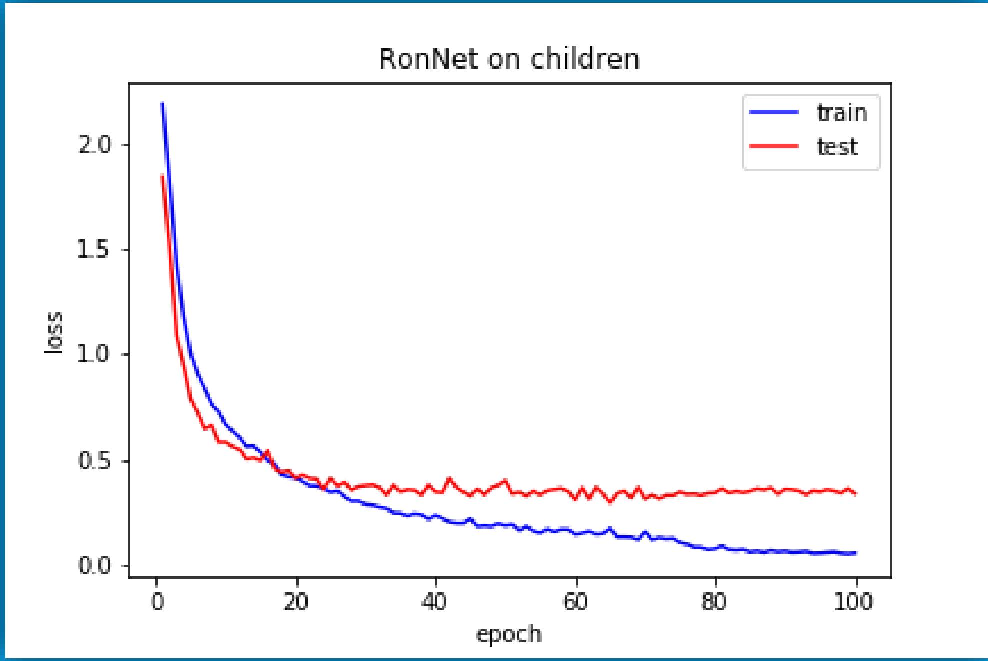
GoogLeNet

	CK+	FER	Nova	Children
CK+	99.225%	-	-	**13.123%**
FER	-	67.93%	-	**25.415%**
Nova	-	-	84.342%	**9.136%**

TRAINING DIRECTLY ON CHILDREN DATASET

Although we used a relatively small dataset of tagged children’s emotions, we received the highest test accuracy score when we trained the model directly on the children dataset.





USING TRANSFER LEARNING

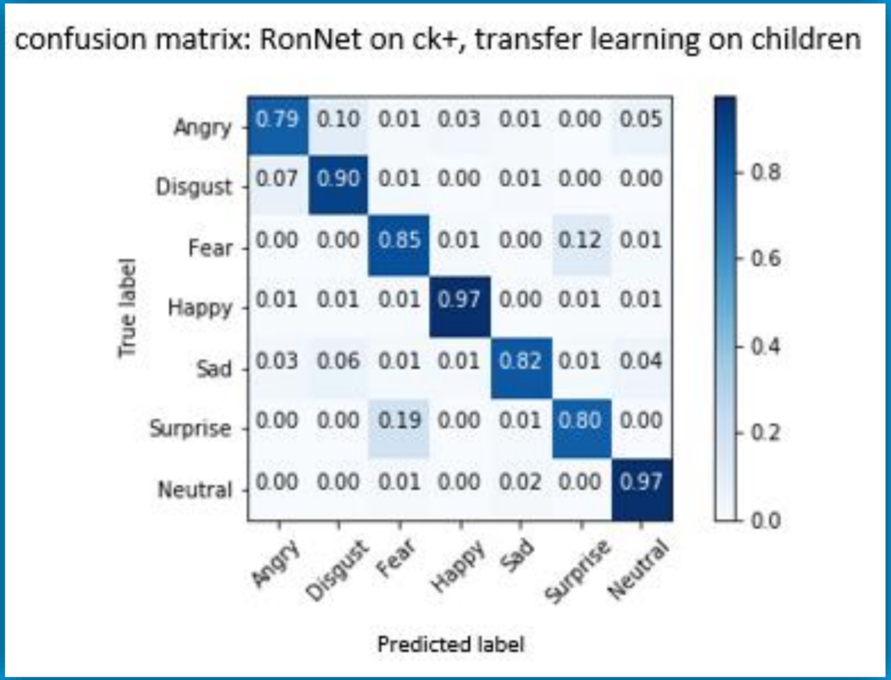
Transfer learning, or knowledge transfer, aims to use models that were pre-trained on different data for new tasks.

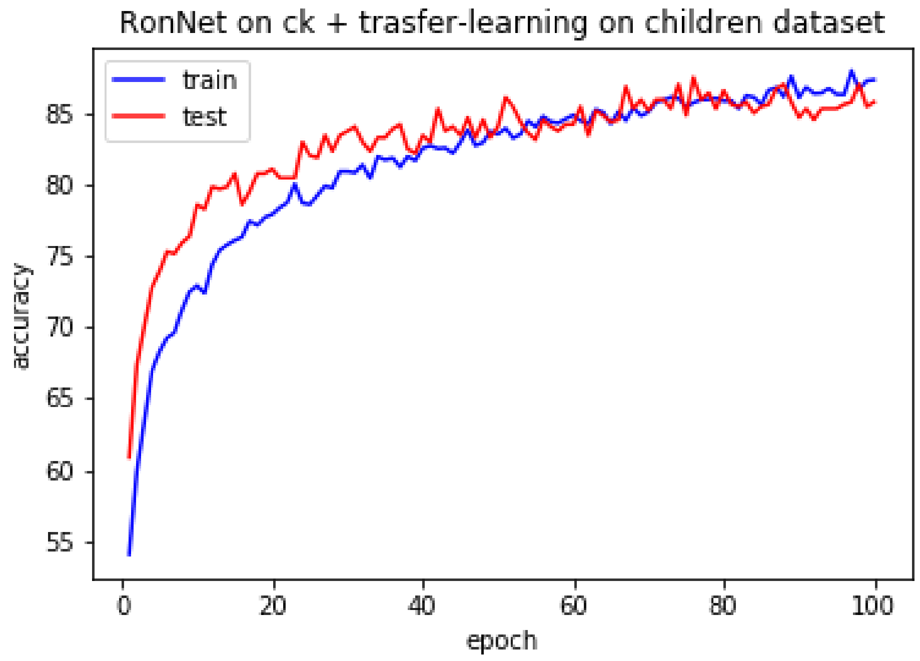
Neural network models often require large training sets. However, in some scenarios the size of the training set is insufficient for proper training.

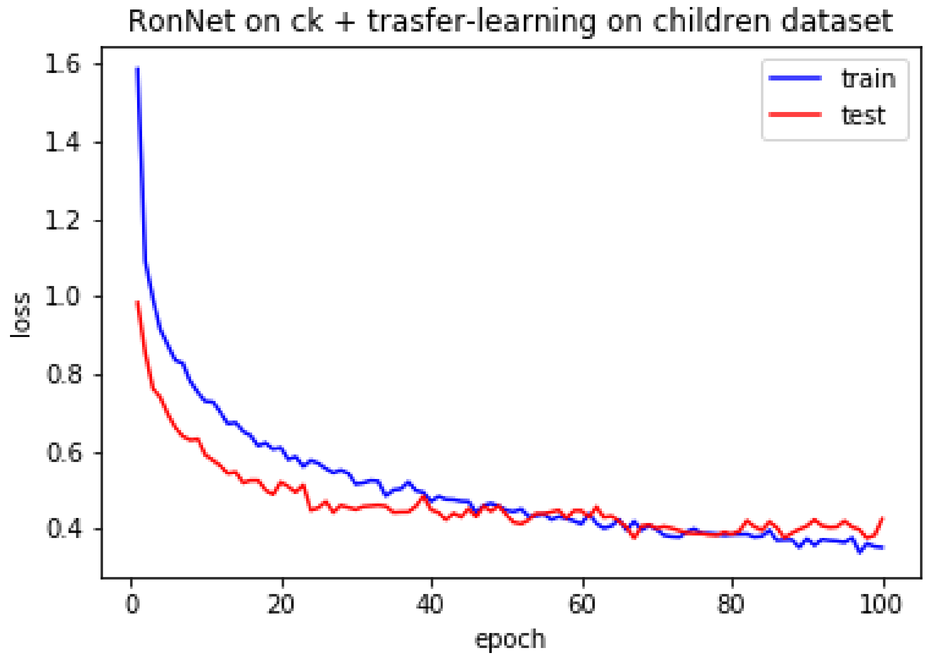
Transfer learning allows using the convolutional layers as pre-trained feature extractors, with only the output layers being replaced or modified according to the task at hand. That is, the first layers are treated as pre-defined features, while the last layers, that define the task at hand, are adapted by learning based on the available training set.

RESULTS

Using transfer learning should have seemingly fit our problem, but the accuracy achieved using transfer learning on a model which was trained on CK+ dataset was 85.759% on children compared to 99.612% on adults.







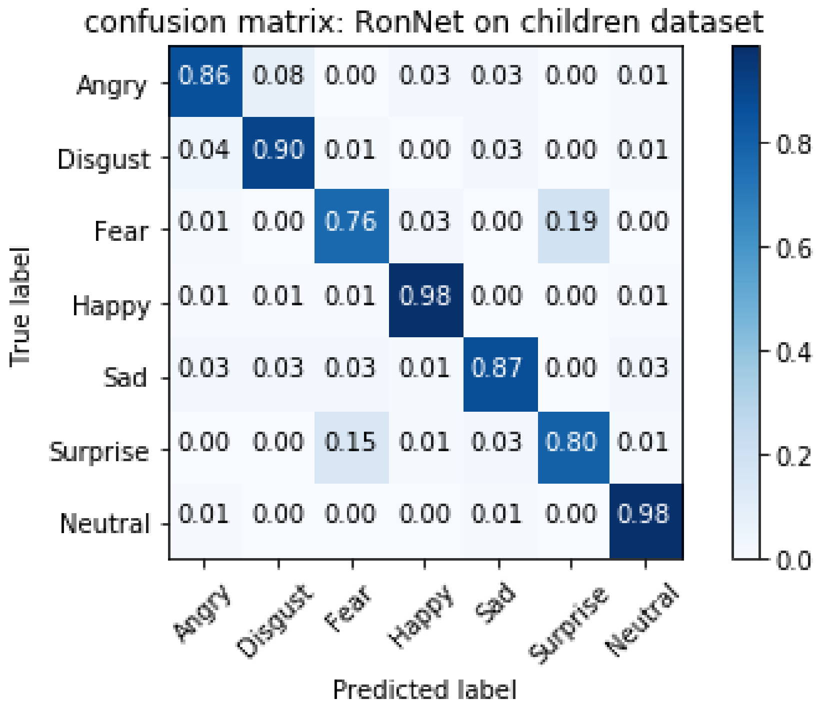
ANALYZING THE RESULTS

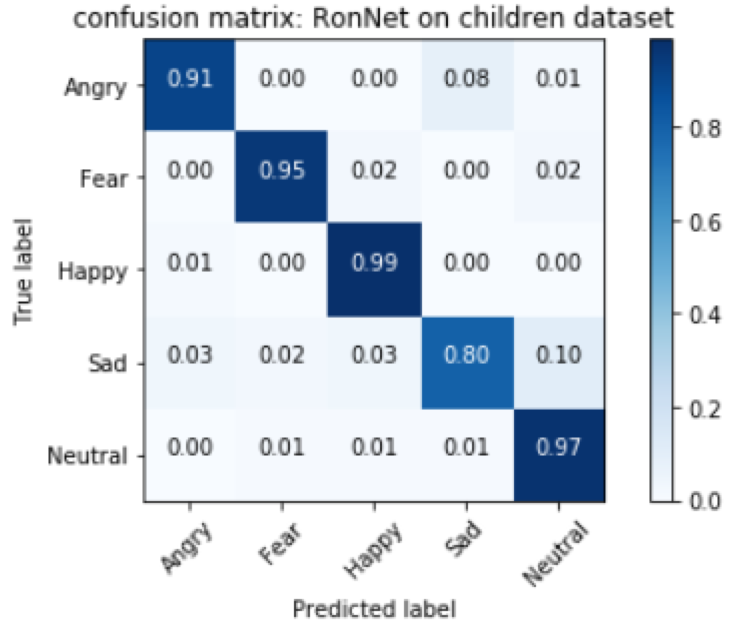
In order to analyze the results, we extracted a confusion matrix that provided us information regarding misclassifications.

As can be seen from the confusion matrix attached, the model tends to misclassify **fear** with **surprise** and **disgust** with **anger**.

We decided to omit "surprise" and "disgust" from our dataset and train the model again in order to receive higher accuracy.

One can notice that when using less emotions, the average accuracy does not increase dramatically.





VISUALIZATION

After choosing the method which gave us better results, we would like to visualize the feature map and examine which parts of the image helped classify every emotion.

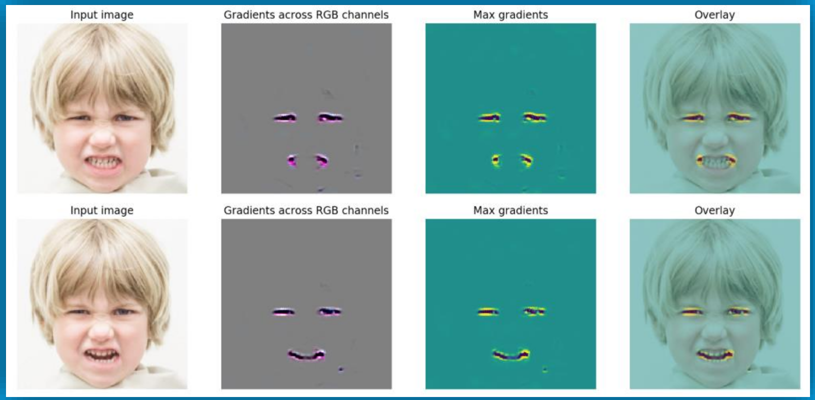
We used a **saliency map** approach which calculates the gradient of the score of interest (typically the highest score, corresponding to the classified class) with respect to the pixels of the input image.

Since the saliency map has the same shape as the input image, the saliency map can be used as a mask to highlight the important part of the input image, with regards to the predicted label. As per the definition of gradients, this saliency map tells us how much the prediction score for class would change if the RGB pixel intensities in the highlighted area of the image are slightly increased.

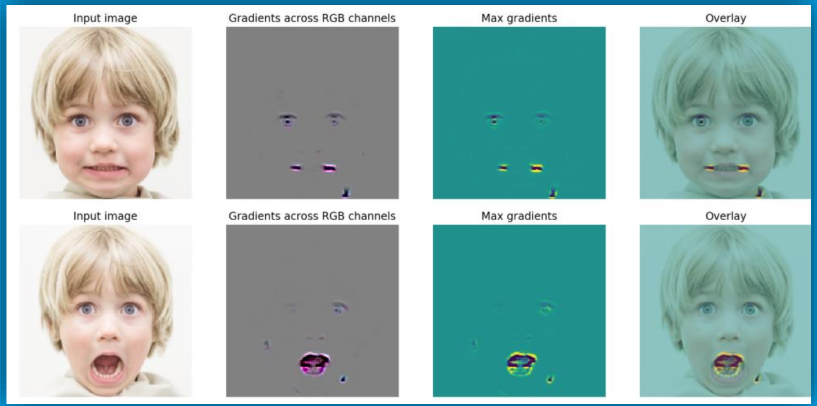
This approach helped us analyze the result and ascertain that our model has learned the expected feature maps.

VISUALIZATION

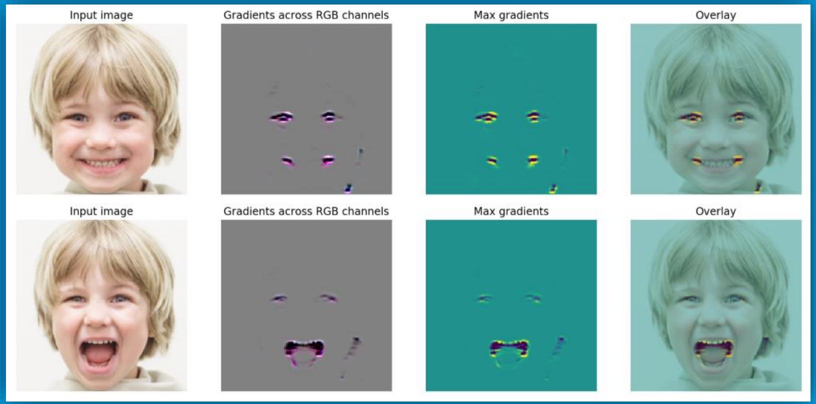
As can be seen from the following saliency maps, our model was more influenced by the eyes and mouth area as expected and has been demonstrated in Ran Breuer's research on adults.



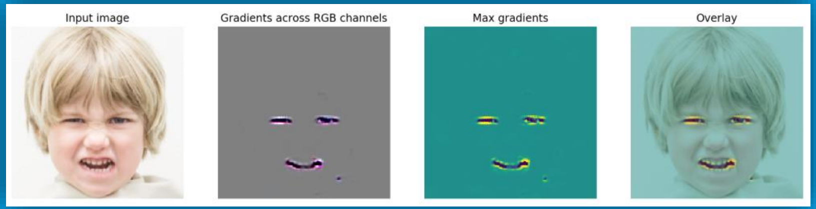
Angry



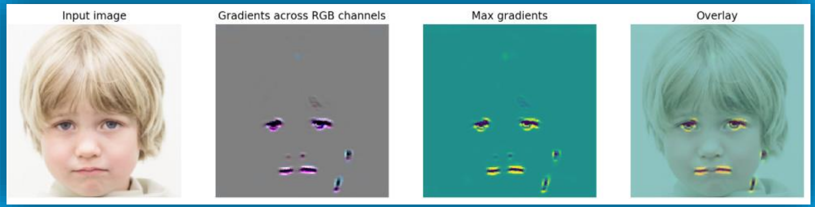
Fearful



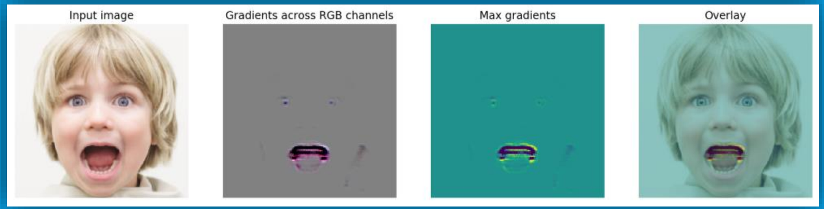
Happy



Disgusted



Sad



Surprised

VISUALIZATION