Facial Expression Recognition in Children

GIP Project, Spring 2019

Dana Goghberg, Sapphire Elimelech, Moran Hait

Supervisors: Ron Kimmel, Tzipi Horowitz Kraus, Michal Zivan, Gary Mataev

Computer Science Department, Technion - Israel Institute of Technology, Haifa

Educational Neuroimaging Center, Technion - Israel Institute of Technology, Haifa

Table of Contents

Abstract

Our project is the product of a collaboration between the Geometric Image Processing Lab (GIP) at the Computer Science department and the Educational Neuroimaging Center (ENIC) at the Technion. The goal of the project is to create an emotion recognition system based on facial expression recognition (FER) algorithm, which will be used by ENIC lab for analyzing children's responses to various tasks, while being monitored by EEG/fMRI. The facial expression of a child can provide information about their emotional state during an activity. We track the children along a video using face recognition and image processing tools, and output their emotion using a classifier.

We are deploying a state-of-the-art model which was developed by the GIP lab for facial expression recognition on adults. Several recent papers were utilizing convolutional neural networks (CNNs) for feature extraction and inference. In this project, we used convolutional neural network to implement a new facial expression recognition system which is specifically designed for children. Although well studied on adults, only a few facial expression recognition studies have been done involving children, and consequently only a handful of small relevant datasets exist. This difficulty was added to the fact that some children's emotions have similar representation and are therefore difficult to differentiate.

To overcome the lack of data, we initially planned using transfer learning so that models that were already trained on adults' expressions will be re-trained on children's expressions. This way we hoped to utilize the existing networks' results for the benefit of our new study. Unfortunately, this led to less-than-desirable results since there is little similarity between adults' and children's facial expressions, as we later learned.

Still, these results were compared to other attempts: training on adults and testing on children, and training solely on children, while comparing the accuracy of each model using different networks.

We tested our chosen model using cross-validation and a real-time emotion detection from a video recorder. The results exceeded our expected goals.

Datasets

We used the following datasets:

1. **Children Dataset:**

   Assembled from the following Datasets:

   Educational Neuroimaging Center (ENIC) Dataset: 409 Israeli children facial expressions which were recorded during EEG scans and tagged by the ENIC lab staff. The facial expressions were classified into seven emotions categories - angry, disgusted, fearful, happy, sad, surprised and a neutral expression. The children were recorded while performing a variety of interactive tasks, similar to the tasks they will perform when using our trained model to automatically detect their emotions. To make the images more specific (without irrelevant background and other faces), we used a face detection algorithm to bound the faces in uniform boxes.

   The Child Affective Facial Expression (CAFE) Dataset - Databrary:
   The Child Affective Facial Expression Set [2] is a collection of 1192 photographs of 2 to 8 year-old children (median = 5.3 years; range = 2.7-8.7 years) posing six emotional facial expressions - sadness, happiness, surprise, anger, disgust, fear and a neutral expression. The full set features 90 female models and 64 male models (27 African American, 16 Asian, 77 Caucasian/European American, 23 Latino, and 11 South Asian).

   With the exception of surprise, children were asked to pose for each expression with their mouths open and with their mouths closed. Surprised faces were only posed with their mouths open. Open mouth disgusted faces generally included a tongue protrusion.

Figure 1: Images from CAFE dataset

DDCF Dataset: Dartmouth Database of Children's Faces[12], a set of photographs of 40 male and 40 female Caucasian children between 6 and 16 years-of-age. Models posed eight facial expressions and were photographed from five camera angles under two lighting conditions. Models wore black hats and black gowns to minimize extra-facial variables. In our project, only the seven relevant emotions were used.

Radboud Faces Database (RAfD) Dataset: The Radboud Faces Database (RaFD) [1] is a set of pictures of 67 models (including Caucasian males and females, Caucasian children, both boys and girls, and Moroccan Dutch males) displaying eight emotional expressions. The RaFD is an initiative of the Behavioral Science Institute of the Radboud University Nijmegen (the Netherlands).

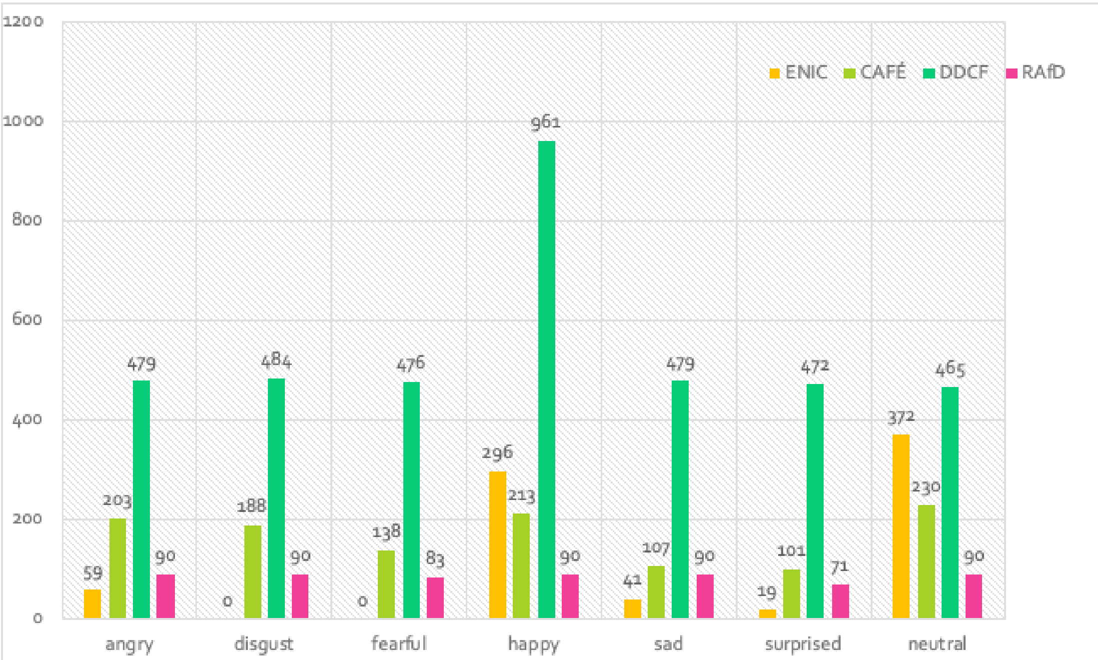


Figure 2: Children datasets emotion distribution

2. FER 2013 Dataset

The *FER 2013* challenge [13] was created using Google image search API containing 35,685 grayscale photographs of classified facial expressions, 48x48 pixels each. Expression are divided into seven categories - angry, disgusted, fearful, happy, sad, surprised and a neutral expression(8 classes in total).

Figure 3: Images from FER2013 dataset

3. CK+ dataset

The Extended Cohn-Kanade Dataset [14] is comprised of video sequences of 210 adults, labeled with seven possible emotional classes. Includes both posed and non-posed (spontaneous) expressions and additional types of metadata. Participant ages range from 18 to 50. 69% are female, 81% Euro-American, 13% Afro-American, and 6% belong to other groups. The dataset is composed of 593 sequences from 123 subjects containing posed facial expressions. Another 107 sequences were added after the initial dataset was released.

Figure 4: Images from CK+ dataset

4. NOVA dataset

This dataset contains the facial expression images captured using the Novaemotions game [15]. It contains over 42,000 images taken from 40 different people. The majority of the participants were college students with ages ranging from 18 to 25, labeled with the challenged expression and the expression recognized by the game algorithm, augmented with labels obtained through crowdsourcing. Here is a small sample with the images and crowdsourced labels:

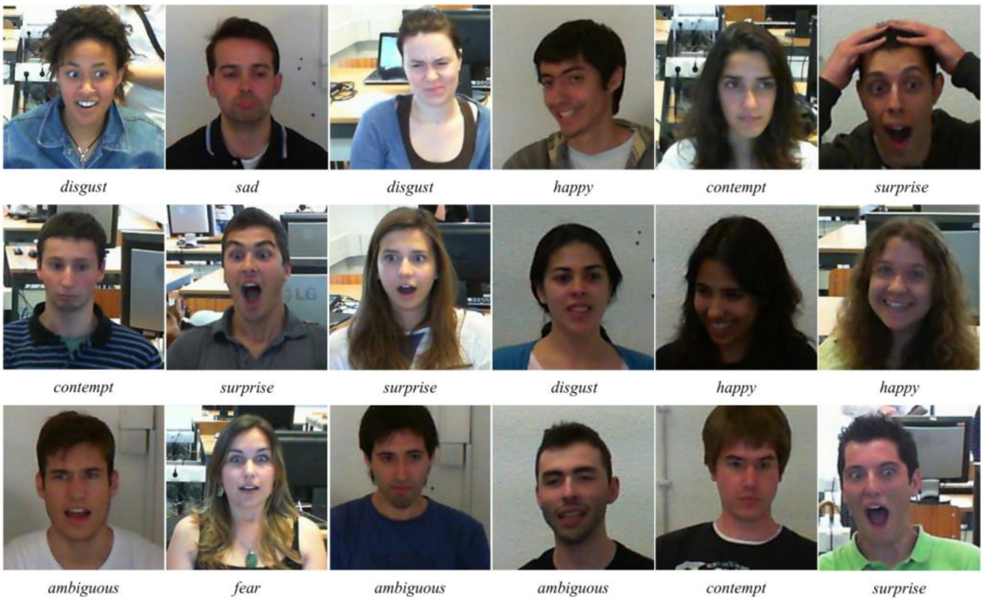


Figure 5: Images from NOVA dataset

Workflow

Theoretical Knowledge Acquirement

Starting this project, we had no knowledge in image processing or machine learning. The Stanford 231n course (Convolutional Neural Networks for Visual Recognition) gave us some understanding of deep learning with an image processing approach. We choose to use Pycharm IDE and worked with Anaconda3 as virtual environment, using Pytorch and OpenCV as our main tools.

This project was also based on Ran Breuer's paper, "A Deep Learning Perspective on the Origin of Facial Expressions" and his neural network RonNet [11], which gave excellent results for emotion classification in adults. Therefore, we needed to have a rather thorough understanding of both.

Figure 6: System workflow

CNN Research

To find our best option, we compared Ran Breuer's convolutional neural network RonNet with one of the state-of-the-art CNNs used in the ImageNet project: GoogLeNet.

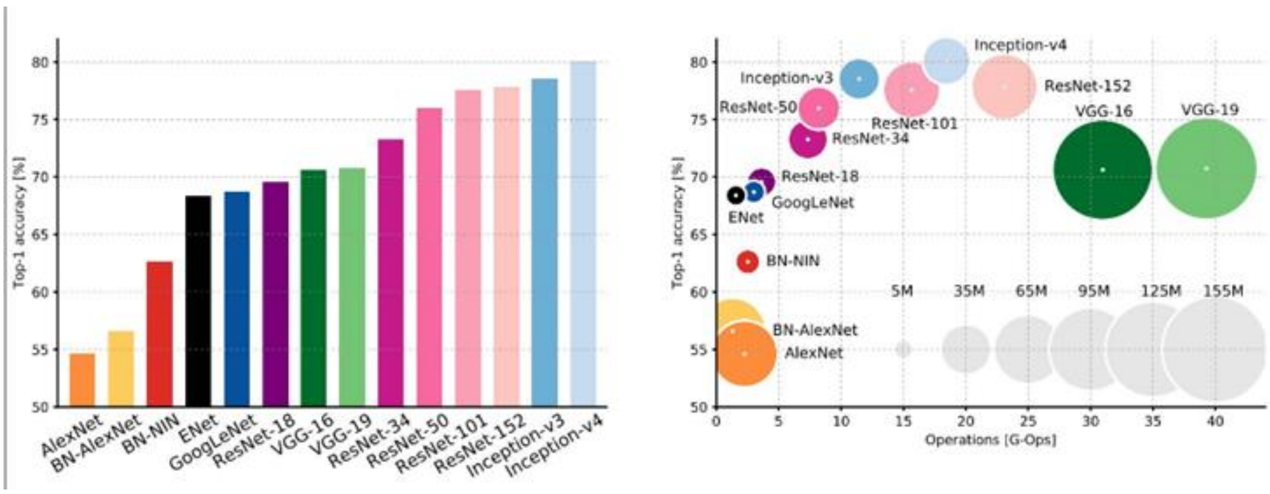


Figure 7: An analysis of Deep Neural Networks Models for Practical Applications 2017

Comparing the CNN Models

GoogLeNet Model

GoogLeNet [16] was based on a deep convolutional neural network architecture codenamed "Inception", which was responsible for setting the new state-of-the-art model for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC 2014). The single crop error rates on the ImageNet dataset with a pretrained model are listed below.

Model structure	Top-1 error	Top-5 error
googlenet	30.22	10.47

All pre-trained models expect input images normalized in the same way, i.e. mini-batches of 3-channel RGB images of shape (3 x H x W), where H and W are expected to be at least 224 pixels. The images must be loaded into a range of [0, 1] and then normalized using mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225].

We chose to use GoogLeNet due to its good performances and relatively small number of operations.

(number of parameters is 4 million).

Here we can see a diagram of the GoogLenet structure (enlarged diagram here):

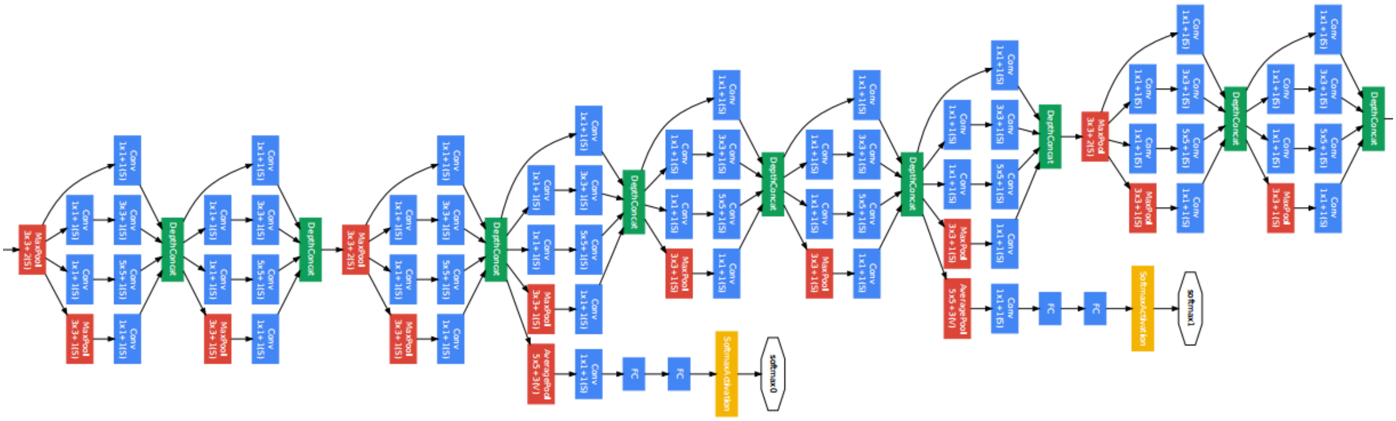
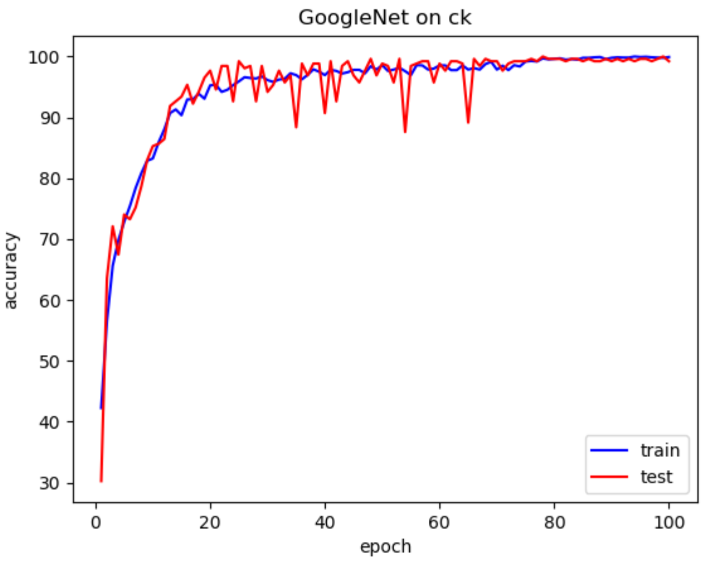
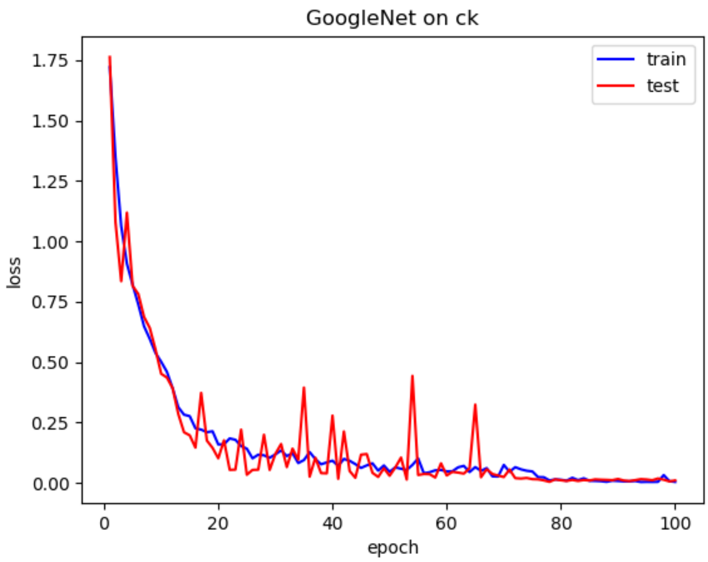


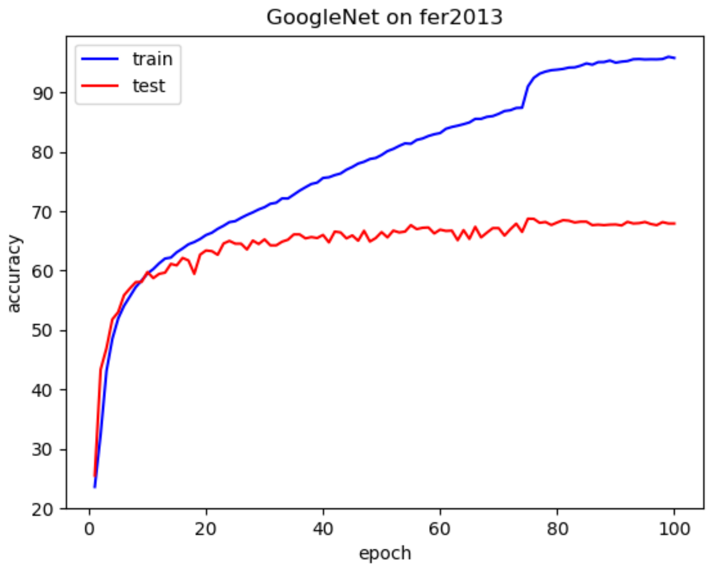
Figure 8: GoogLeNet Model, 2014

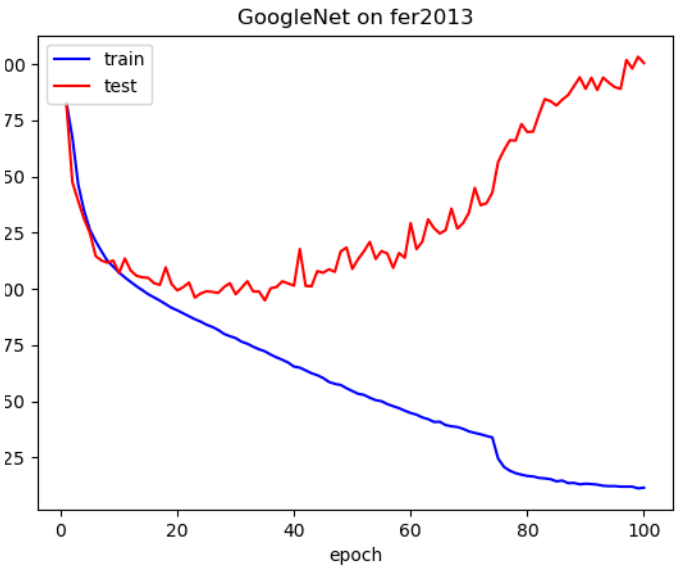
GoogleNet on ck+ dataset gained 99.914% accuracy on the train set and 99.225% accuracy on the test set:





GoogleNet on FER2013 dataset gained 95.782% accuracy on the train set and 67.93% accuracy on the test:





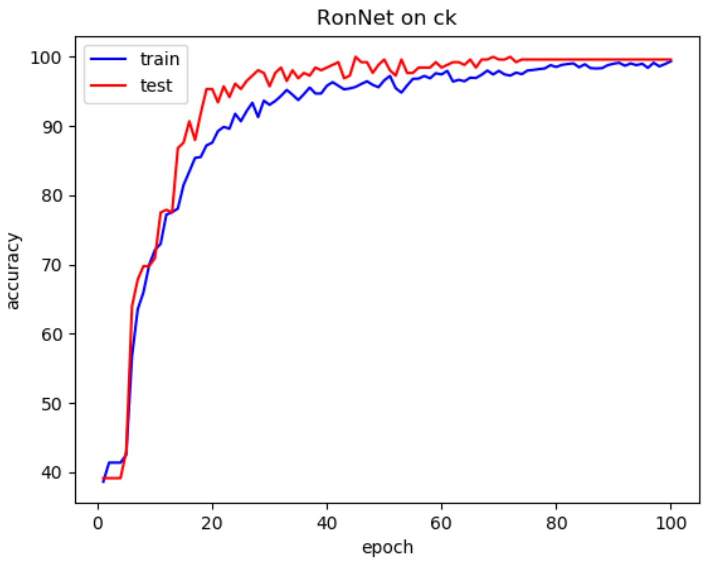
RonNet Model

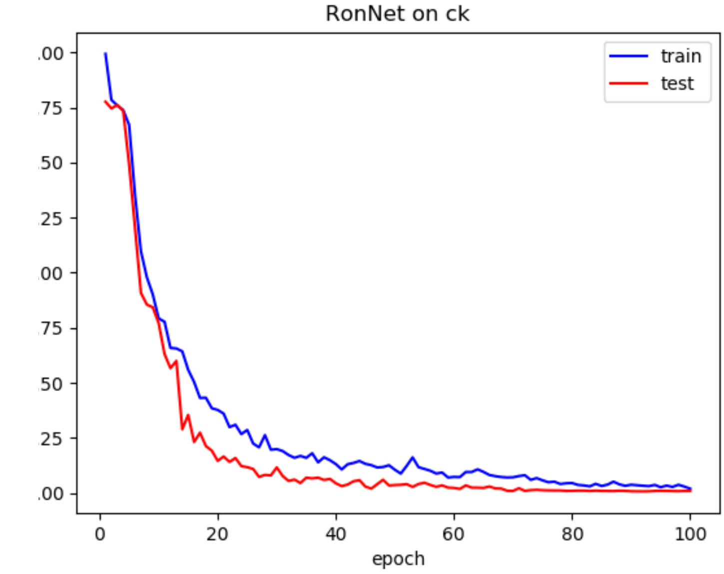
This classifier was developed by Mr. Ran Breuer during his master's work for prop Ron Kimmel [9]. This is a classic feedforward convolutional neural network. Ran's implementation consists of 3 convolutional blocks, each with a rectified linear unit (ReLU) activation and a pooling layer with 2×2 pool size. The convolutional layers have filter maps with increasing filter (neuron) count the deeper the layer is, resulting in a 64, 128 and 256 filter map sizes, respectively. Each filter supports 5×5 pixels.

The convolutional blocks are followed by a fully connected layer with 512 hidden neurons. The hidden layer's output is transferred to the output layer, whose size is affected by the task at hand - eight for emotion classification. Dropout was applied after the last convolutional layer and between the fully connected layers to reduce overfitting with probabilities of 0.25 and 0.5 respectively. A dropout probability *p* means that each neuron's output is set to 0 with probability *p*. For training ADAM optimizer was used with a learning rate of 1e−3 and a decay rate of 1e-5.

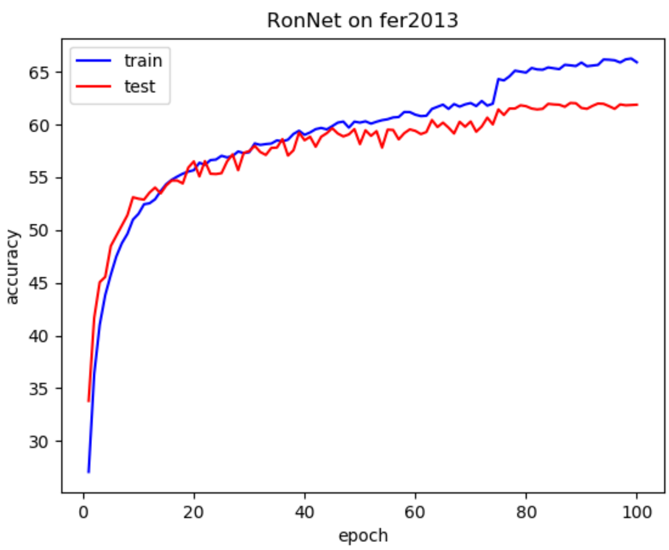
Figure 9: RonNet Model (Left to Right)

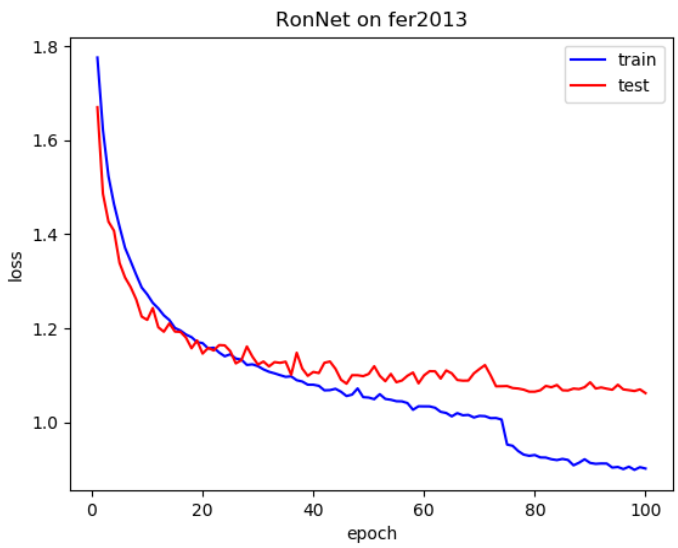
RonNet on FER2013 dataset gained 99.354% accuracy on the train set and 99.612% accuracy on the test set.





RonNet on ck+ dataset gained 65.913% accuracy on the train set and 61.884% accuracy on the test.



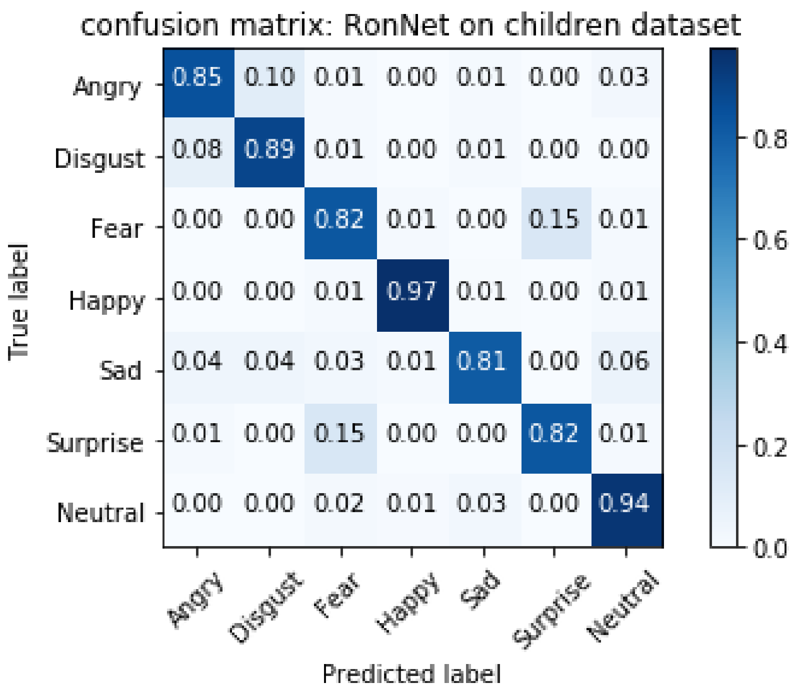


Data Preprocessing

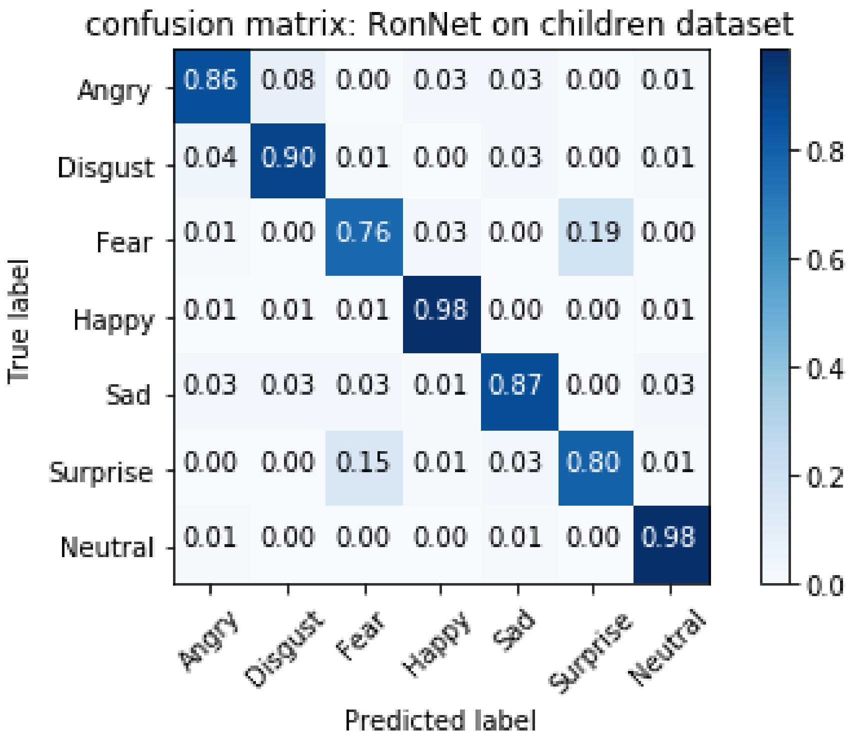
We used several preprocessing methods on the data:

- Face crop: using a face detection algorithm, we extracted only the face from every image. This 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 datasets. We used two different algorithms in order the maximize the number of detected faces: we first used *autocrop model* (see reference [10]) and then *poor_man_rekognition* (see reference [11]), for the images the first one could not recognize.
- Data augmentation: to maximize generalization of the model, we used data augmentation methods to generate synthetic data and enlarge the training set: a combination of random flips and affine transforms, e.g. rotation, translation, scaling and shear.
- RGB vs grayscale: we hypothesized that converting every picture to grayscale will result in higher accuracy since emotions do not correlate with the colors in the image, but we received similar results on RGB and grayscale images. See the result as presented below in a form of confusion matrix.

RonNet on children database when all pictures are converted to greyscale:



RonNet on children database in RGB (Original Images):



Therefore, and in order to speed up the learning process, we converted all the children's dataset to grayscale, thus “squeezing” our images’ 3 channels (RGB) into one channels.

Training the Model

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

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

Training the Model on Adults' Datasets 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.

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

	CK	**FER**	**NOVA**	**CHILDREN**
CK	99.225%			**13.123%**
FER		67.93%		**25.415%**
NOVA			84.342%	**9.136%**

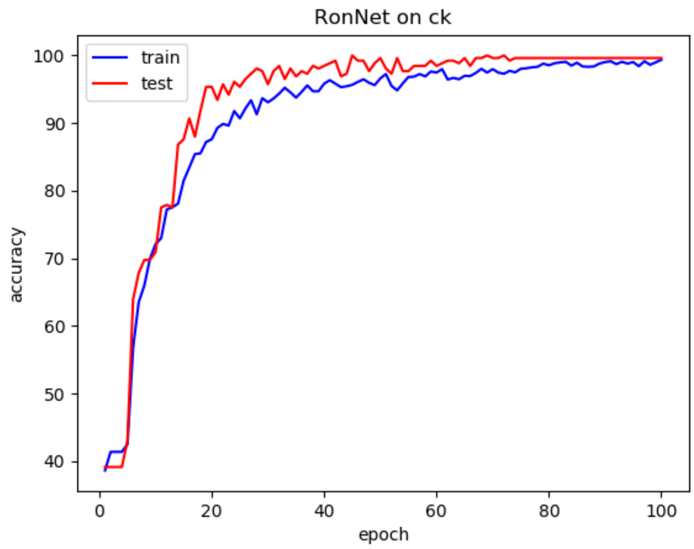
Figure 10: Test Accuracy running on GoogleNet (in %)

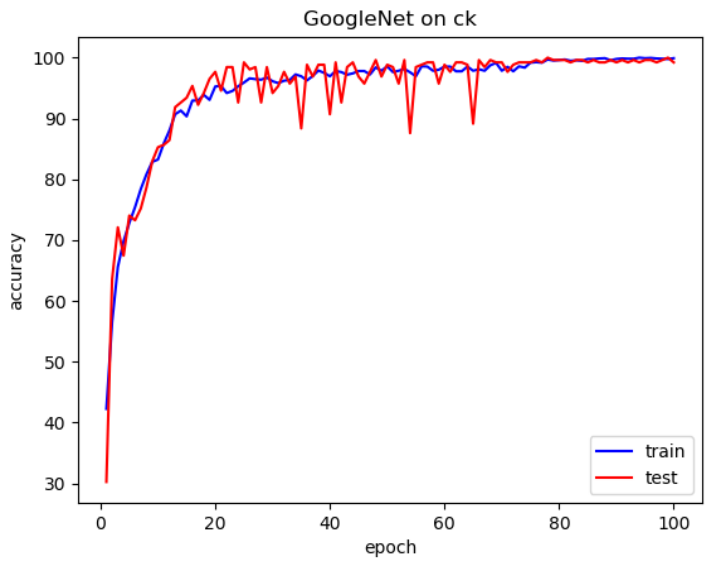
	CK	FER	NOVA	CHILDREN
CK	99.612%			**25.415%**
FER		61.884%		**23.754%**
NOVA			83.534%	**19.934%**

Figure 11: Test accuracy running on RonNet (in %)

Choosing the Model

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 (which contributes to faster training).





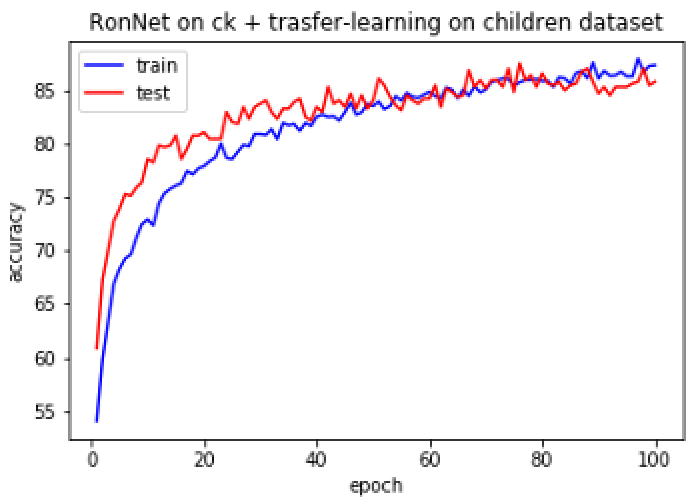
Train by Using Transfer Learning on Children Dataset and Freezing the Convolutional Layers

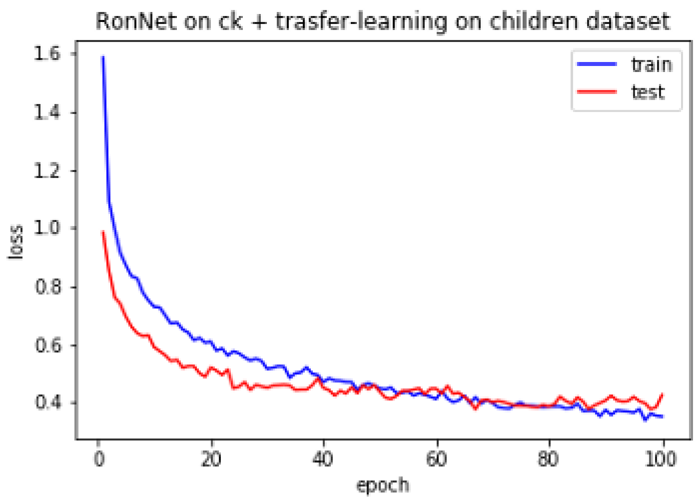
Transfer learning, or knowledge transfer, aims to use models that were pre-trained on different data for new tasks.

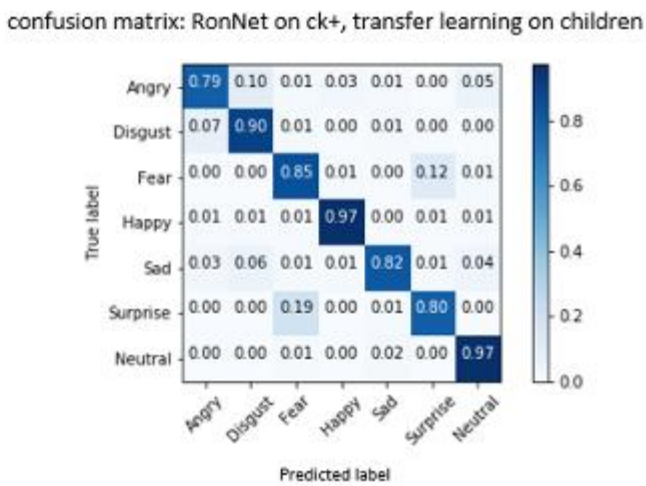
Neural network models often require large training sets. However, in some scenarios the size of the training set is insufficient for proper training.

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.

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.



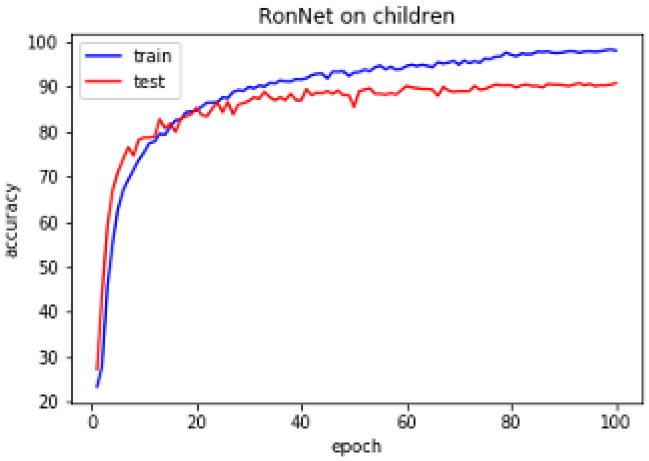


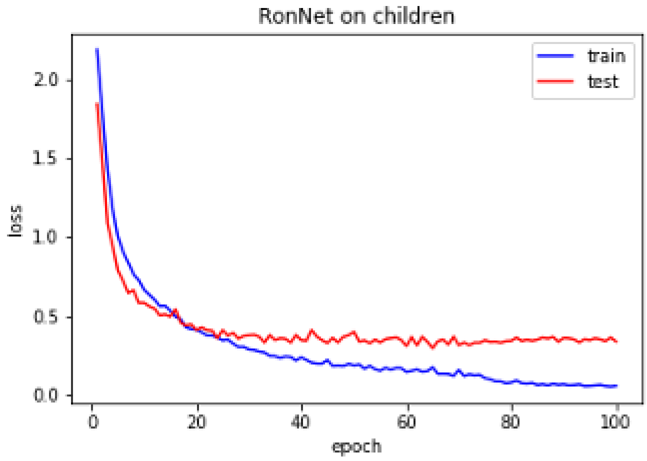


Training the Model Directly on Children

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.

Important note: the test set consisted of children who performed a wide range of emotions and were not included in the train set.





After comparing the results, we achieved using transfer learning to the results we achieved training solely on the children's data set, we chose the latter model - using RonNet classifier.

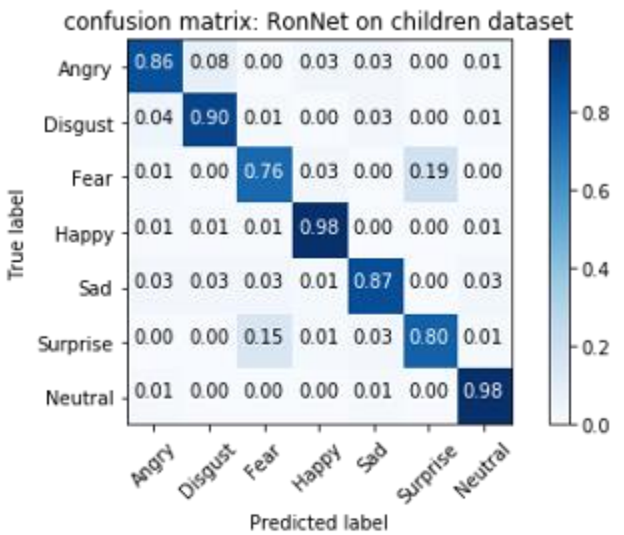
Analyzing the Results

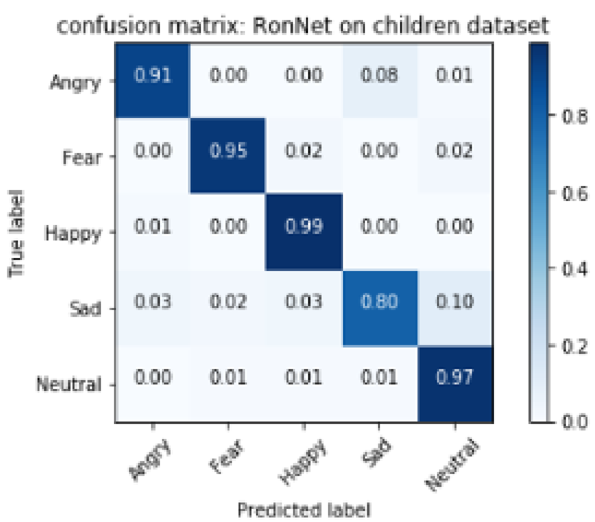
To analyze the results, we extracted a confusion matrix that provided us information regarding misclassifications.

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

After discussing these results with the ENIC lab, we decided to omit *surprise* and *disgust* from our dataset and train the model again to increase the accuracy.

One can notice that when using less emotions, the average accuracy does not increase dramatically (Acc: 91.274%).





Visualizations

Neural networks are often described as "black boxes". The lack of understanding on how neural networks make predictions enables unpredictable/biased models, causing real harm to society and a loss of trust in AI-assisted systems.

Feature visualization is an area of research, which aims to understand how neural networks perceive images.

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 *s*lightly increased.

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

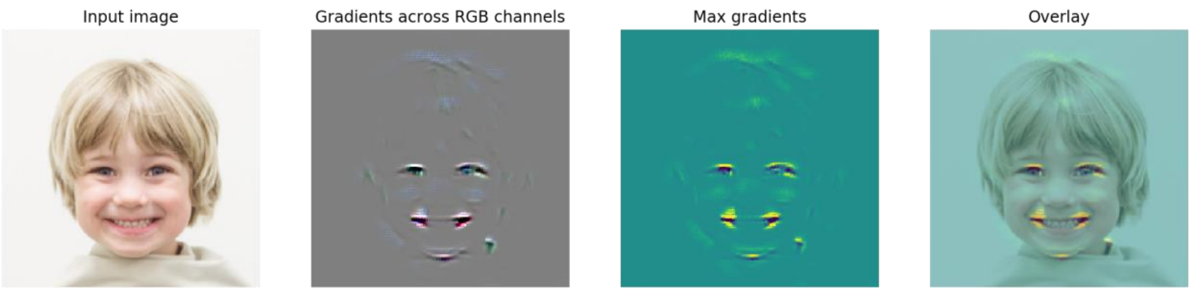


Figure 12: Happy feature map

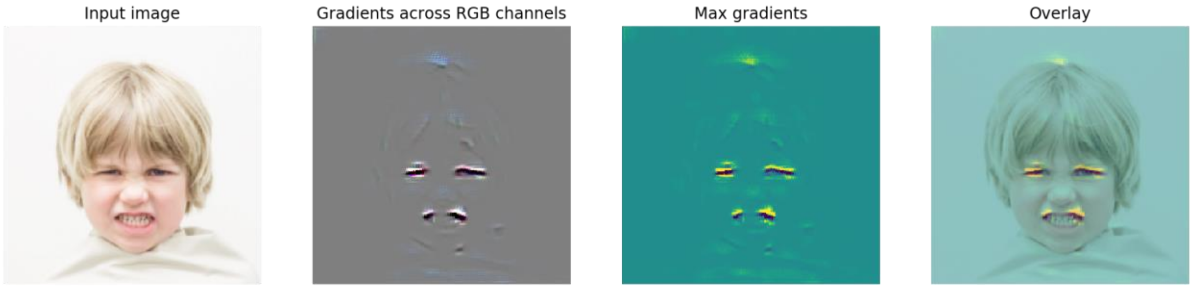


Figure 13: Angry feature map

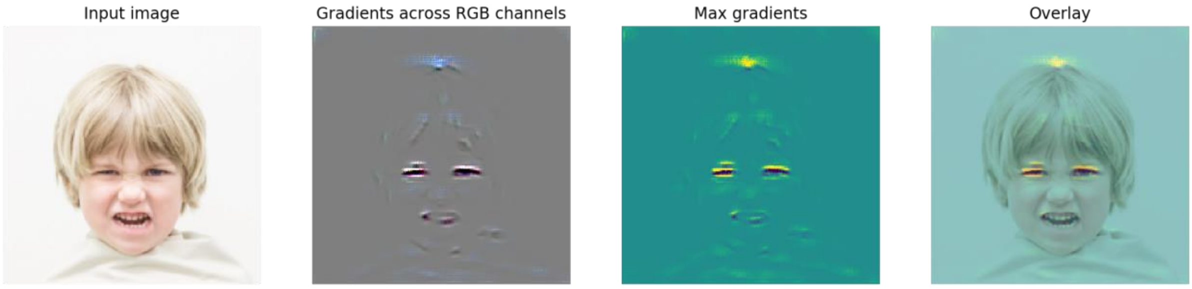


Figure 14: Open Angry feature map

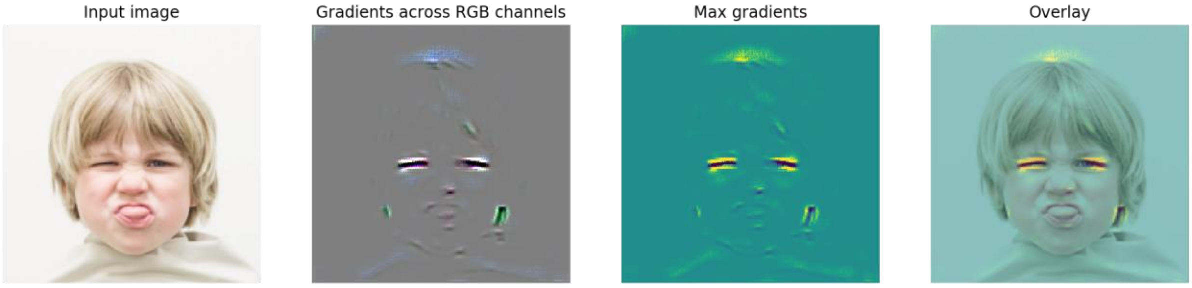


Figure 15: Disgust with tongue feature map

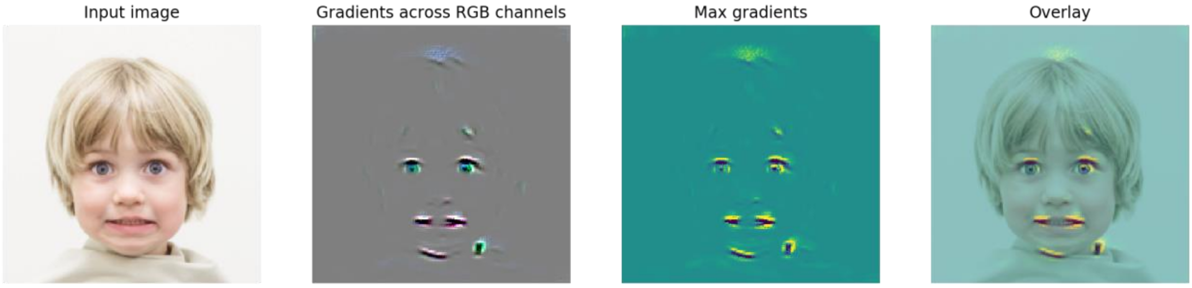


Figure 16: Fear feature map

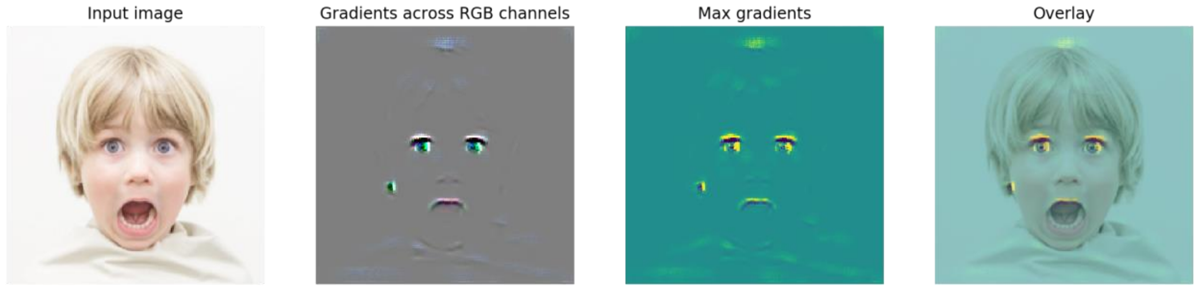


Figure 17: Fear open mouth feature map

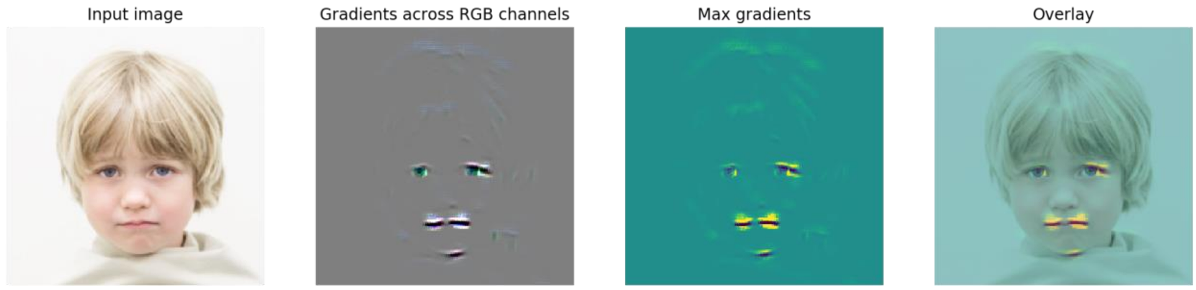


Figure 18: Sad feature map

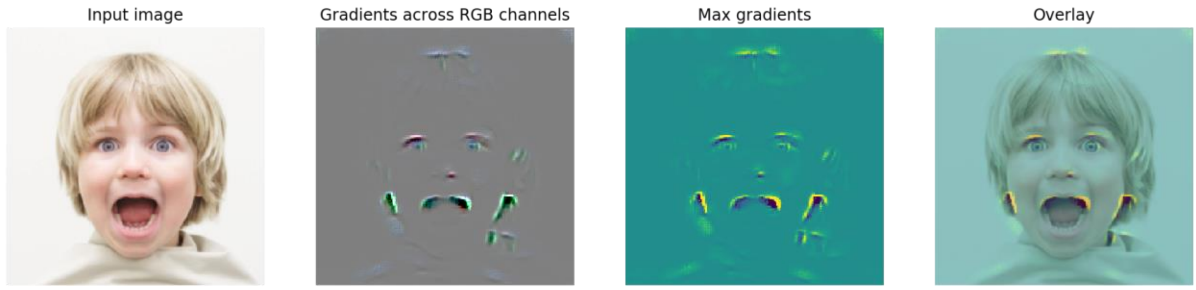


Figure 19: Surprise feature map

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

Conclusion and Next Steps

We created a FER system for children based on machine learning tools, to be used by the ENIC lab during the performance of EEG and fMRI scans. It can detect emotion from live video and display the results as a function of time for each child, to be synced with the EEG machine (with synchronized time stamps).

We introduce a novel machine-learning system that detects facial expressions in children with 91% accuracy. We are not aware of approaches yielding similar results in the literature.

As can be seen in the confusion matrix, some of the emotions are similar and sometimes hard to differentiate, such as *fear* and *surprise*, and removing one of the two leads to better results. We can assume that more training data can solve this confusion.

An interesting observation was that while testing an adult-trained model on children leads to poor results, testing a children-trained model on adults resulted in 63% accuracy. This implies that while this project proves there is a significant difference between the way children and adults express emotions, we can assume that children's facial expressions generalize better to adult emotions than the other way around.

This project can be extended in many ways to further serve the ENIC lab and a variety of other fields, such as child psychology, sociology, neuroscience and child development. Immediate candidates are:

1. Expanding the facial recognition algorithm to include a parent during an interaction and utilize an adult FER network to classify their emotional state alongside the child.
2. Enabling emotion recognition in a group of children during an interaction, in a way that continually monitors each child and their emotional state.
3. Researching the difference discussed earlier - how emotions in children allow for better recognition in adults that the opposite, and the social and cultural reasons that might cause this derivation.

Appendix - Facial Emotion Recognition System User Manual

The facial emotion recognition (FER) system will be used by the Educational Neuroimaging Center lab for analyzing children's responses to various tasks, while being monitored by EEG. The system is based on a convolutional neural network (CNN) which was trained on various kinds of datasets.

1. <u>Choosing the model</u>**:**

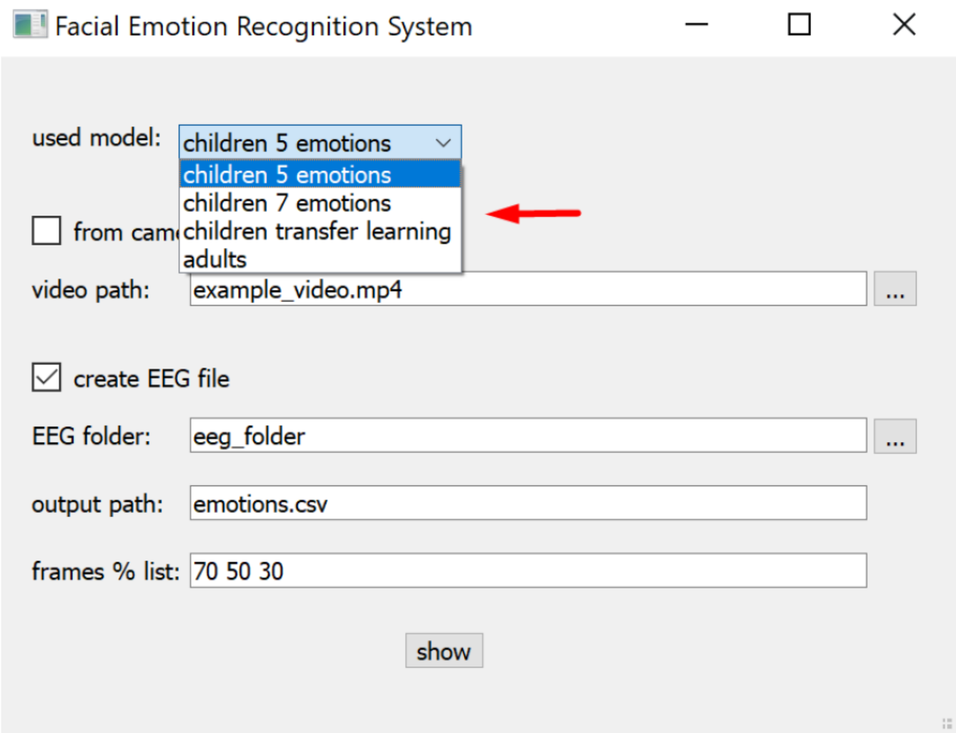
   The "used model" scroll bar enables the user to switch between four different models:

   children 5 emotions – classifies 5 emotions in children: Angry, Fearful, Happy, Sad, Neutral.

   children 7 emotions – classifies 7 emotions in children: Angry, Disgusted, Fearful, Happy, Sad, Surprised, Neutral.

   children transfer learning– classifies 7 emotions in children: Angry, Disgusted, Fearful, Happy, Sad, Surprised, Neutral. This model was trained using transfer learning and is different from "children 7 emotions" although it classified the same set of emotions.

   adults – classifies 8 emotions in adults: Neutral, Angry, Contempt, Disgusted, Fearful, Happy, Sad, Surprised.

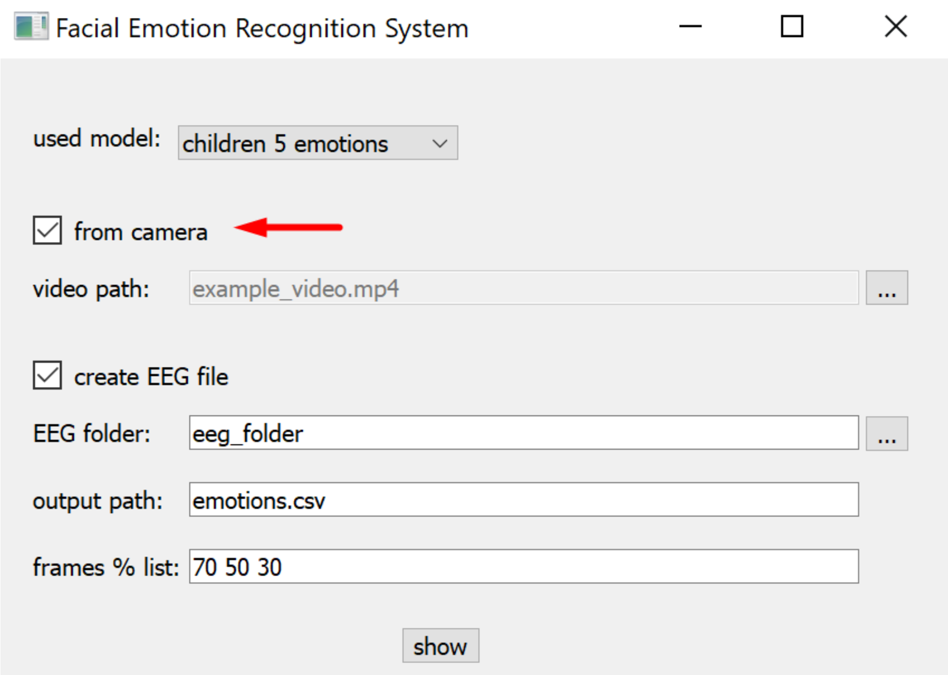


2. Choosing input:

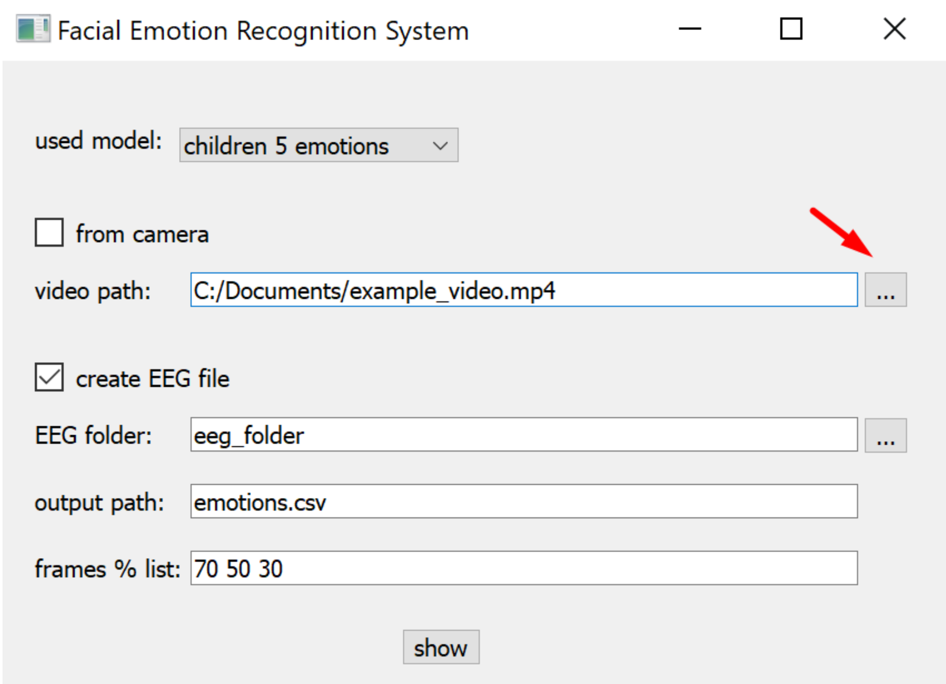
   Note: this application needs permission to use the computer webcam.

   The input to the system can be taken from the computer webcam or an existing video file.

   To choose to use the computer webcam, simply check “from camera” button.



To select an existing video from the computer, make sure the “from camera” button is unchecked and select the video path manually or select the file using the “…” button.



3. Creating EEG markers file:

   The system can be used to send EEG markers when a change in the subject's emotion is detected. In order to do so, provide the system with an existing EEG marker folder which stores the EEG marker files of the subject in the video. An example folder can be found in the system's code directory. Check the "create EEG file" button and specify the EEG folder path manually or choose it using the "…" button.

4. Creating output file:

   The system creates an output file with the following data in CSV format:

	A	B	C	D	E	F	G
1	Time	Angry	Fearful	Happy	Sad	Neutral	Classification
2	20191230203542000000	0.1	0.01	0	0	99.89	Neutral
3	20191230203542033333	0.01	0	0.03	0	99.96	Neutral
4	20191230203542066666	4.2	0.02	0.04	0.03	95.72	Neutral
5	20191230203542100000	1.21	0.01	0.02	0.01	98.76	Neutral
6							
7							
8	total	0	0	0	0	100	

   Column A (time): specifies the frame time beginning from the time the video was taken ("ctime").

   20191230203542033333 = 2019/12/30 20:35:42:033333 (microsecond)

   Columns B-F (emotion): specifies the percentage for every emotion as detected by the model.

   Column G (classification): specifies the classification which is the emotion with the highest percentage.

   Total: shows the percentages of every emotion throughout the video/webcam capture.

   Note that the output file name can be changed, but it must end with ".csv".

5. Frames % list:

   The frames % list is a list of numbers with backspace delimiter as shown in the example. Every frame sent to the model will output a percentages list which indicates the probability for every emotion. The classification is the emotion with the highest percentage.

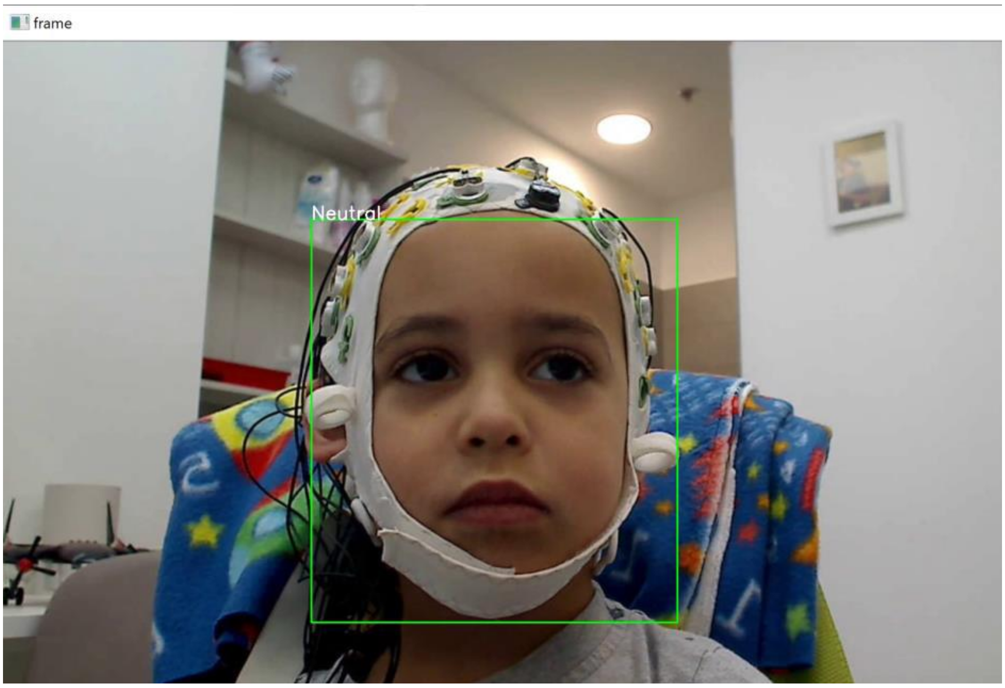
   The frames % list enables the user to take the percentage list of previous frames under consideration when classifying the current frame to overcome misclassifications. The % list specifies the percentage for every previous frame, i.e. how it will affect in the classification. For example, given the list "70 50 30", the system will take 70% of the previous frame's percentage list, 50% of the 2nd frames before the current frame and 30% of the 3rd frame before the current frame under consideration.

Formula:

100%$*current\ frame$+70%$*previous\ frame$+50%$*\ 2nd\ previous\ frame$+30%$*3rd\ previous\ frame$

Note that the list can be of any length, a longer list will use a higher number of frames before the current frame.

Once all the parameters are set, press “show” to evaluate the video/webcam capture. The system will show the video frame-by-frame, will display every frame’s classification and will create the files once the video is finished. If you are using a webcam camera or don’t want to evaluate the whole video, you can press ‘Q’ on the keyboard at any time to exit the video/webcam capture and create the files up to the time the video/webcam capture was stopped.



Acknowledgements

We want to thank Gary Mataev, Yaron Honen, Ron Kimmel, Tzipi Horowitz Kraus, Michal Zivan and Sapir Bar for their kind help and their contribution to the success of the project.

References

1. *Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. Cognition & Emotion, 24(8), 1377—1388. DOI: 10.1080/02699930903485076*
2. *LoBue, V., & Thrasher, C. (2015). The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults. Emotion Science, 5, 1532.*
3. *LoBue, V. & Thrasher, C. (2014). The Child Affective Facial Expression (CAFE) set. Databrary. Retrieved October 13, 2019 from http://doi.org/10.3389/fpsyg.2014.01532*
4. *Challenges in Representation Learning: A report on three machine learning contests." I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio. arXiv 2013.*
5. *R Vemulapalli, A Agarwala, "A Compact Embedding for Facial Expression Similarity", CoRR, abs/1811.11283, 2018.*
6. *Crowdsourcing facial expressions for affective-interaction, Gonçalo Tavares, André Mourão and João Magalhães, Computer Vision and Image Understanding, Volume 147, June 2016, 102-113*
7. *Competitive affective gaming: Winning with a smile, André Mourão and João Magalhães Proceedings of the 21st ACM international conference on Multimedia (ACM MM 2013), 83-92*
8. *Misa Ogura (2019, September 26). MisaOgura/flashtorch: 0.1.1 (Version v0.1.1). Zenodo. http://doi.org/10.5281/zenodo.3461737*
9. *A Deep Learning Perspective on the Origin of Facial Expression, Ran Breuer, Ron Kimmel - Department of Computer Science Technion - Israel Institute of Technology Technion City, Haifa, Israel*
10. *autocrop face detection and crop https://leblancfg.com/autocrop/*
11. *face_recognition tool https://github.com/backSpace001/poor_man_rekognition#poor_man_rekognition*
12. *Dartmouth Database of Children's Faces; Dalrymple et al., 2013)*
13. *Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing, pages 117–124. Springer, 2013.*

14. *Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, 13-18 June 2010, pages 94–101. IEEE Computer Society, 2010. doi: 10.1109/CVPRW.2010. 5543262. URL http://dx.doi.org/10.1109/CVPRW.2010.5543262.*
15. *André Mourão and João Magalhães. Competitive affective gaming: Winning with a smile. In Proceedings of the 21st ACM International Conference on Multimedia, MM '13, pages 83–92, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2404-5. doi: 10.1145/2502081.2502115. URL http://doi.acm.org/10.1145/2502081. 2502115.*
16. *Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich Going Deeper with Convolutions.*