BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Kids Emotion Recognition

– ITSG report –

Team members
Almasan-Tigau Alexandra
Gorgos Andreea
Corman Robert-Marian

1

2019

Contents

Chapter 1

Neural Networks

1.1 Introduction of neural networks

Neural Networks are a set of algorithms, modeled mainly after the human brain, that are designed to identify and recognize different types of patterns. The patterns that the algorithms recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated. Neural networks can be seen as a clustering and classification layer on top of the data that needs to be stored and managed. They help to group unlabeled data according to similarities between the example inputs. All classification tasks depend upon labeled datasets. In other words, humans must transfer their existing knowledge to the dataset so a neural network to learn the correlation between labels and data. This is known as supervised learning and brings value for:

- Detect faces, identify people in images, **recognize facial expressions**
- Recognize gestures in video
- Identify objects in images

and many others.[5]

1.2 Neural Networks Elements

Neural networks are the networks compose of several **layers**. The layers are made of **nodes**. A node is a place where computation happens, meaning that it combines input from the data with a set of coefficients, or weights, that either amplify of dampen that input. These input-weight products are summed and then the sum is passed to an **activation function** (of a node) to determine if the signal should progress further through the network in order to affect the final outcome. If this happens and the signals passes, then the neuron's status is active.[5, 2]

Chapter 2

Algorithm: Convolutional Neural Networks

2.1 What?

A **Convolutional Neural Network (CNN)** is a type of artificial neural network used in image recognition and processing that is specifically designed to process data based on its pixels. A CNN is a Deep Learning algorithm which can take in an input image, assign importance (weights) to various aspects in the image and be able to differentiate one from the other. Traditional neural networks are not ideal for image processing and must be fed images in reduced-resolution pieces. CNN have their ”neurons” (nodes) arranged more like those of the frontal lobe, the area responsible for processing visual stimuli in humans.[3]

2.2 Architecture

A Convolutional Neural Network consists of a number of convolutional layers optionally followed by fully connected layers.The input of a layer is a $\boldsymbol{m \times m \times r}$ image, where m is the height and width of the image and r is the number of channels (for example, in case of an RGB image, $\boldsymbol{r = 3)}$. The convolutional layer will have $\boldsymbol{k}$ filter/kernels of size $\boldsymbol{n \times n \times q}$ where n is smaller than the dimension of the image and q can either be the same as the number of channels r or smaller. Out of these data there can be produces **k** feature maps of size **m-n+1**.[1].

2.3 How?

In the context of a convolutional neural network, a convolution is a linear operation that involves multiplication of a set of inputs with a set of weights, just as described in the previous chapter of Neural Networks. Given that the technique was designed for two-dimensional input, the multiplication is performed between an array of input data and a two-dimensional array if weights called filter. The results of previous multiplications are then summed and in the end a single value is returned, that is why the operation is often called **scalar product** (between inputs and weights). Usually, the filter is smaller than the input and this detail allows the same set of weights to be multiplied by the input array multiple times at different points of the input. As the filter is applied multiple times to the input array, the result is a two-dimensional array of output values that represents a filtering of the input. As such, the two-dimensional output array from this operation is called a **feature-map**. So convolutional networks perform a sort of search among the provided image and try to find the previously generated feature-map among the initial image. Each time a match is found, it is mapped onto a feature space particular to that visual element. In that space, the location of each match is recorded. [2]

Our aim was to find an algorithm that has already attained good results in detecting adult facial emotions and train it to recognize children facial emotions while maintaining the highest level of results' accuracy.

Adults' emotion recognition resultse

The ratio used was: 80-10-10 for training-validation-test sets.

The best result obtained: 71.161% accuracy [**?**]

Children emotion recognition results

The ratio used was: 80-10-10 for training-validation-test sets.

The best result obtained: 65.5335% accuracy

2.4 Methodology

- What are criteria you are using to evaluate your method?

  Accuracy. The accuracy of a model is usually determined after the model parameters are learned and fixed and no learning is taking place.

- What specific hypotheses does your experiment test? Describe the experimental methodology that you used.

  It's testing how well algorithm that does classification for emotions on people faces.

- What are the dependent and independent variables?

  - Learning Rate: 0.001
  - Adam optimizer
  - categorical_crossentropy loss

- What is the training/test data that was used, and why is it realistic or interesting? Exactly what performance data did you collect and how are you presenting and analyzing it? Comparisons to competing methods that address the same problem are particularly useful.

  FER 2013 Dataset

  - Does not contain kids images
  - Traning acc (90% from dataset): 98.64% - Test acc (10% from dataset): 67.24% (a bit of overfitting)

  CAFFE Dataset

  - Data are unbalanced
  - Labeling seems to be wrong (sad seems neutral or others)
  - Has open mouth faces (e.g. for sad, neutral)
  - Traning acc (90% from dataset): 89.94% - Test acc (10% from dataset): 67.59% (a bit of overfitting) - also removing the open mouth images
  - Contains only kids images

2.5 Data

FER The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression in to one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

train.csv contains two columns, ”emotion” and ”pixels”. The ”emotion” column contains a numeric code ranging from 0 to 6, inclusive, for the emotion that is present in the image. The ”pixels” column contains a string surrounded in quotes for each image. The contents of this string a space-separated pixel values in row major order. test.csv contains only the ”pixels” column and your task is to predict the emotion column.

CAFFE To train the algorithm for facial recognition we have used the data set provided by the Child Study Center at Rutgers University, New Jersey. The Child Affective Facial Expressions Set (CAFE)[**?**] represents the first large representative database of children posing using many facial expressions. This data set consists of approximately 1200 photographs of over 100 children racially and ethnically diverse (90 female models and 64 male models: 27 African American, 16 Asian, 77 Caucasian/European American, 23 Latino, and 11 South Asian)with ages between 2 and 8 years old, that are posing 7 different facial expressions: happy, angry, sad, fearful, surprise, neutral and disgust.

The institute focuses their research on the cognitive, emotional and perceptual development of infants and children. To gather all these visual records, the institute is inviting all families having a child with age between 3 and 8 years, to take part in the study by bringing their child to a session of 30-45 minutes of computer interactive games or of a stimulating reading session. The children are verbally invited to pose each of the emotion with their mouth open and with their mouth closed.

2.6 Results

- FER: Traning acc (90% from dataset): 98.64% - Test acc (10% from dataset): 67.24% (a bit of overfitting)
- CAFFE: Traning acc (90% from dataset): 89.94% - Test acc (10% from dataset): 67.59% (a bit of overfitting) - also removing the open mouth images

2.7 Discussion

- Is your hypothesis supported?

  Yes

- What conclusions do the results support about the strengths and weaknesses of your method compared to other methods?

  A more larger balanced dataset with kids faces is needed to obtain better results.

- How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

  Overfitting could happen sometimes (depends on the dataset), but the algorithm is pretty accurate as a human can be.

Bibliography

[1] http://deeplearning.stanford.edu/tutorial/supervised/ConvolutionalNeuralNetwork/

[2] https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/

[3] https://searchenterpriseai.techtarget.com/definition/convolutional-neural-network

[4] https://skymind.ai/wiki/convolutional-network

[5] https://skymind.ai/wiki/neural-network

[6] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53