석 사 학 위 논 문
Master's Thesis

캡슐 네트워크를 이용한 얼굴 표정 인식

Facial Expression Recognition Using Capsule Networks

2019

막슴축 바슬 (Maksymchuk, Vasyl)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

캡슐 네트워크를 이용한 얼굴 표정 인식

2019

막슴축 바슬

한 국 과 학 기 술 원

전기및전자공학부

캡슐 네트워크를 이용한 얼굴 표정 인식

막슴축 바슬

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2018년 12월 19일

심사위원장 김 대 식 (인)

심 사 위 원 정 송 (인)

심 사 위 원 이 현 주 (인)

Facial Expression Recognition Using Capsule Networks

Vasyl Maksymchuk

Advisor: Dae-Shik Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Electrical Engineering

Daejeon, Korea
December 19, 2018

Approved by

Dae-Shik Kim
Professor of Electrical Engineering

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MEE
20174099

막슴축바슬. 캡슐 네트워크를 이용한 얼굴 표정 인식. 전기및전자공학부 . 2019년. 25+iv 쪽. 지도교수: 김대식. (영문 논문)
Vasyl Maksymchuk. Facial Expression Recognition Using Capsule Networks. School of Electrical Engineering . 2019. 25+iv pages. Advisor: Dae-Shik Kim. (Text in English)

초 록

얼굴 표정을 이해하는 것은 사람들간의 비언어적 의사소통의 가장 기본 보편적인 구조중의 하나이다. 이러한 감정 표현들을 분류하는 능력은 더 나은 기계-사람 상호작용에 있어서 중대하다. 본 논문에서는 다양한 데이터 집합의 학습된 특성을 일반화하는 능력으로 알려진 캡슐 네트워크 구조를 사용하여 감정 분류 문제에 대해 연구한다. 알고 있는 한, 이것은 깊은 신경 네트워크를 사용하여 인간 얼굴의 정서적인 변화 인코딩을 배우는 첫 번째 접근법이다. 제안된 모델은 감정 분류기가 중요한 얼굴 특징을 학습하도록 유도하는 얼굴 핵심 검출 단위를 갖는다. 제안된 방법을 사용하여 우리는 보편적인 인간의 감정표현을 구분지을 수 있었고, 신경 네트워크가 어떠한 통제 없이 몇 가지 감정표현 활성화 단위를 학습할 수 있다는 것을 보여주었다.

핵 심 낱 말 얼굴 표정, 분류, 캡슐 네트워크

Abstract

Facial expression understanding is one of the basic universal constructions of nonverbal inter-human communication. The ability to classify facial expressions is crucial for better machine-human interaction. In this thesis, we study emotion classification problem using Capsule Network architecture, which is known for ability to generalize learned characteristics of various datasets. To the best of our knowledge, this is a first approach to learn emotional variance encoding of human face using deep neural networks. The proposed model has facial keypoint detection unit, which encourages emotion classifier to learn critical facial attributes. Using the proposed method, we were able to disentangle universal human expressions and we showed that the neural network could learn several expression action units without any supervision.

Keywords Facial expression, Classification, Capsule network

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Facial expression understanding is one of the basic universal constructions of nonverbal inter-human communication. Recently, with the advance of machine learning field, the problem of facial expression recognition became a hot topic in computer vision area. Solving this problem is crucial for a better human-machine communication systems. As Mehrabian's communication research showed, 55% of the meaning of the sentence is pertain to the facial expressions while only 7% and 38% are relevant to the words and para-linguistics (the quality of voice and the way of vocalization), respectively [1]. Apart from the significance in a design of communication systems, a better understanding of human emotions could be beneficial in a large number of other applications. For instance, detecting mental disorders, remote detection of people in trouble, detection of malicious behavior, obtaining satisfaction feedback in customer services and others.

Some researchers [2] argued that emotions are universal and genetically inscribed in human body showing the cross-cultural similarity between different subjects. However, the other studies [?] claimed that emotions are not only biologically determined but also are culturally affected. A person tend to show higher or lower emotional arousal level based on one's cultural background. Nevertheless, all people are lean to react in a similar manner in a positive antecedent situations (for instance accomplishing one's goals) and negative antecedent circumstances (for instance one's failure) without regard to individual culture. Despite these similarities, the task of emotion classification could vary within three different aspects: classification, recognition method, and classifier construction.

The classification task could be divided into two theories. The first theory considers all emotions to be discrete and fundamentally differ from each other. In 1978, Ekman and Friesen[3], proposed a Facial Action Coding System (FACS), a system that is used to taxonomize human facial movements. This system defines a set of Action Units (AU) which are the basic building blocks for emotion formation. Using this system, all emotions could be categorized into six groups, called the six universal emotions: joy, surprise, sadness, anger, disgust, and fear. The second theory tries to classify emotions on a dimensional basis. To support this theory, Schlosberg[4] proposed a method to classify all emotions within three dimensions: "pleasantness-unpleasantness", "attention-rejection", and "level of activation".

Based on a recognition method the task of emotion classification could be split into three categories[5]. The first method is to use geometrical features of the face, such as facial keypoints (for instance, position of the nose, mouth and eyes), geometrical shapes, etc. for classification. The second method uses an entire image and recognizes emotions based on the appearance of the picture. The third method combines the first two methods. It uses a low-level geometrical features and simultaneously benefits from the global features of the image.

Finally, the classifier construction is separated into two categories: static and dynamic[6]. The static methods deals with a single image, typically used for an instant emotion recognition based on a single image. The dynamic method requires a sequential input, usually a short video clips featuring facial expression progression.

In this work, we try to classify emotions within the discrete space of six universal facial expressions. We propose a geometry and appearance-based method for a single-shot static image classification. Our method is based on the Capsule networks[7], which are know for their ability to generalize various representation features of many different datasets. Using facial keypoints, we were able to disentangle

universal human expressions and our experiment showed that the capsules could encode numerous action units (inner and outer brow movements, brow lowering, lip corner pulling and depressing, and mouth stretching) within themselves without any supervision. Unlike other approaches, out method not only learns the classification of facial expressions, but also learns the variation of facial emotions. To the best of our knowledge, this is the first approach to recognize facial expression using Capsule networks.

Chapter 2. Theoretical Background

This section describes important key architectures used in the experiment. It also provides theoretical details about training of the described networks and compares them showing their advantages and drawbacks.

2.1 Convolutional Neural Networks

Convolutional neural networks became a very popular and widely-used algorithm in computer vision area after LeCun et al.[8] introduced the first successful architecture of LeNet5, which was composed out of three convolutional layers combined with max-pooling layers followed by three sequential fully-connected layers. This network could easily classify hand-written digits unlike previously designed hand-crafted object classification algorithms. The advantage of the proposed architecture was the existence of the shared sliding window of parameters which could see the entire image and learn the needed features for object classification. Unlike multilayer perceptron neural networks, convolutional networks do not require the storage of a vast number of parameters. Instead, the sliding windows play a role of important feature detectors, providing a summary of the detected features as an output. As mentioned in [9], these features could be a presence of the curves, edges, color gradient or others. The other advantage of this architecture is that the higher level feature detection is based on the previously detected features, which in the end allows the network to work with entire image.

However, there is a drawback in this model. The proposed max-pooling layer, which provides the summary of the presence of the features, does not take into account the translational relationship between the features. Therefore, the new approach, called Capsule networks, which considers the loss of important information, was proposed.

2.2 Capsule Networks

In 2017, Geoffrey Hinton et al.[7] demonstrated a novel method of viewing convolutional neural networks. The original idea of capsules[10] was published long time before however only with advance of machine learning the algorithm, called dynamic routing between capsules, was proposed allowing to perform training of capsule networks.

The Capsule networks are based on the convolutional neural networks. The proposed original architecture of Capsule Network is shown on the Figure 2.1. The input of the network is a single 28x28 gray scale image which is also an input to a convolutional layer with 256 filters. The output is put into another convolutional layer, however this time with the stride of 2 in order to reduce the size of the maps, and the output of this layer is transformed into 32 features maps 6x6x8. These features maps are called Primary Capsules. After, these maps are converted into 1,152 vectors of length 8 which will be the input to the next DigitCaps layer. Unlike the traditional convolutional networks, the input to the last layer is not scalar but vector. The communication between Primary Capsules and Digit Capsules is performed with the dynamic routing mechanism. The description of this procedure is shown on the Figure 2.2.

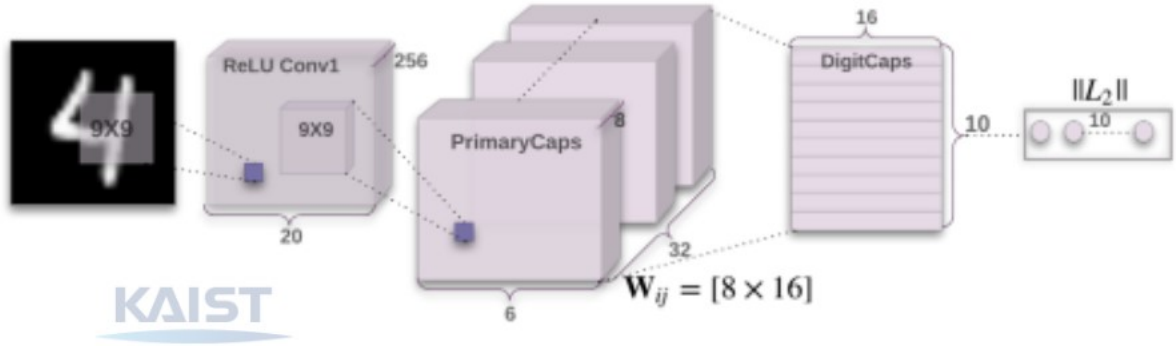

Figure 2.1: Capsule Network architecture (from [7]).

Procedure 1 Routing algorithm.

1: **procedure** ROUTING($\hat{\boldsymbol{u}}_{j|i}, r, l$)
2: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.
3: **for** r iterations **do**
4: for all capsule i in layer l: $\mathbf{c}_i \leftarrow \texttt{softmax}(\mathbf{b}_i)$
5: for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$
6: for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$
7: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$
return $\mathbf{v}_j$

Figure 2.2: Routing algorithm (from [7]).

The transformation process starts with taking the Primary Capsule input vector u_i and multiplying it by a translation matrix W_{ij}. Then the routing process starts between $\hat{u}_{j|i}$ and vector s_j. First, all coefficients b_{ij} are assigned with zero values. The coupling coefficients c_i are calculated as a softmax output of b_i. Once the coupling coefficients are known, the previously obtained vectors $\hat{u}_{j|i}$ are multiplied by these coupling coefficients c_i. These multiplied vectors are then summed up and serve as the inputs to the next capsule layer. Before proceeding forward these new input vectors are normalized using squash function.

$$\hat{u}_{j|i} = W_{ij} u_i$$

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})}$$

$$s_j = \sum_i c_{ij} \hat{u}_{j|i}$$

$$v_j = \text{squash}(s_j) = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||}$$

The coupling coefficients are updated with each iteration in the routing algorithm and are determined by this process. To obtain a better performance of the network, it requires non-linear activation function. In the multidimensional case, the commonly used functions like hyperbolic tangent, ReLU, softmax or

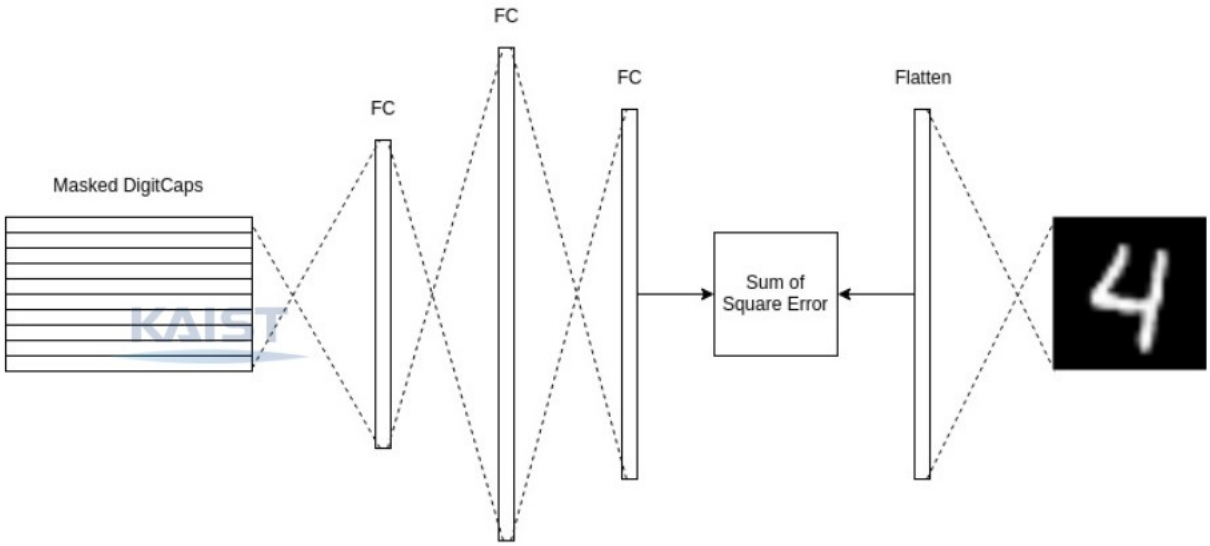

Figure 2.3: Reconstruction error.

others cannot be used. Therefore, the authors proposed a new non-linear function for a vector space, called squash, that does not inference on the direction of the vectors. Additionally, the squash function is used to guarantee that each input vector to the next capsule layer is strictly bounded between 0 and 1. Unlike in traditional convolutional neural network algorithm, the dynamic routing process determines how the capsules communicate between each other. The iterative update of the coupling coefficients determines in which way the higher level capsules will output a better prediction based on the lower level capsule inputs.

Finally, the length of the output capsules, Digit Capsules, will determine the probability of the presence of particular class.

The loss function for this network is divided into two pars: the marginal and reconstruction losses. The marginal loss increases if the predicted class does not correspond to the input image. It can be calculated as

$$L_M = T_k \max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k)\max(0, ||v_k|| - m^-)^2$$

where T_k is 1 if an input image belongs to class k and T_k is 0 otherwise. The upper and lower boundaries of vector v_k, m^+ and m^-, are set to be 0.9 and 0.1, respectively.

To calculate the reconstruction error, additional decoder network is trained. This network consists out of three consecutive fully connected layer. The architecture of the decoder network is shown on the Figure 2.3. The input to this network is a masked DigitCaps output. Only a true labeled capsule is used and the rest of the capsules are masked with zero values. The output of the decoder network is 784-dimensional vector which is compared to the flattened original image vector. The reconstruction error, L_R, is equal to the sum of the square error between reconstructed and input images. The reconstruction loss is used to force capsules to learn representational features like angle, stroke width, narrowness, etc.[7] of the input images. This loss is also used with the purpose of regularization.

The total loss could be calculated as

$$loss = L_M + \lambda L_R$$

where λ controls the image reconstruction rate. Higher values of λ will put more importance to the image reconstruction. It is a hyper-parameter of the network and should be chosen so that it does not

dominate the marginal loss during training. In the original work, this hyper-parameter was set to be equal to 0.0005.

This method achieves state-of-the-art classification results on MNIST[11] dataset. The classification error is 0.25%.

2.3 Comparison between CNN and CapsNet

Even though the two architectures look similar, there are a few significant differences between them. The key differences are listed as follows.

- Convolutional neural networks are translation invariant. For example, the Capsule networks will not classify a randomly assembled facial parts as a face, while convolutional networks will identify different facial features and classify it as a face.
- Convolutional neural networks require a lot of training samples to generalize[12].
- Capsule networks mimic human vision system. A human could easily identify the same object if it is rotated while convolutional neural networks will output totally different result for spatial translation of an object.

Chapter 3. Related Work

3.1 Geometry-based methods

Many different works use facial landmarks to classify facial expressions. For example, [13] uses a feed forward convolutional network to detect eyes, nose, eyebrows and lips and based on these defections they classify an input image.

Another common approach is to use action units to classify emotions. [14] uses joint-patch and multi-label learning for classification purposes. [15, 16] introduced an attention module for action units. The training of these architectures is performed in semi-supervised manner. However all these algorithms are strictly dependant on the action units, and it has been observed[17] that certain emotions could be expressed independently to action units.

3.2 Appearance-based methods

[18] uses a Siamese neural network to classify histogram oriented gradients of the image. Unlike geometry-based approaches, this methods takes a whole image as an input and calculates the appearance features for classification. [19] used an ensemble of 8 networks proposed in [18]. This network contains more than 170 million parameters and achieves state-of-the-art of 75.2% classification accuracy performance on the fer2013 dataset classification.

Another common algorithms of expression classification are based on the fine-tuning of the pre-trained object classifiers[20, 21, 22]. However, these methods do not learn how to generalize the variance of the facial expression. Moreover, these models often lack ability to generalize in-the-wild conditions (bad lightning, atypical angles, poor image quality).

Chapter 4. Proposed Method

This section explains the proposed method to solve the initial problem of facial expression classification. Section 4.1 describes the preliminary steps for data management. Section 4.2 shows the overall architecture of the proposed method.

4.1 Preparation Steps

4.1.1 Feature extraction

The task of image recognition and object classification has an important role in machine learning area. Since the beginning of the ImageNet competition[23] in 2012, AlexNet architecture[24] showed the power of the machine learning algorithms applied to object classification task. After that, various models like GoogLeNet[25], VGG16[20], InceptionV3[22], ResNet50[21] and other were proposed, and some of these models achieved state-of-the-art classification results on ImageNet dataset. The weights of these pre-trained on ImageNet models are available online and could be downloaded for usage.

The VGG16 model's architecture is shown on the Figure 4.1. This model was selected for basic feature extraction mechanism due to its simple architecture and impressive performance (it achieves 90.1% top-5 accuracy). It is constructed out of five convolutional blocks (13 convolutional and 4 pooling layers in total) followed by three fully connected layers.

In this experiment, the weights of the first three blocks of VGG16 network were left unchanged and the output of this cut-network is used as an input of the next block. This process, called transfer learning, is used when a learned knowledge needs to be transferred to the other network. It is also know as fine-tuning of the network. It is a common technique used in computer vision field. The weights of the lower layers are usually left unchanged since they learned how to capture universal features like curves and edges which could be used to learn a relevant task.

4.1.2 Keypoints extraction

Convolutional neural networks are one of most popular architectures in computer vision field for image analyzing. These networks could easily learn many different features of the images. Many successful architectures[20, 21, 22, 26] for face detection, object tracking or object detection are based on these convolutional networks. Therefore, the proposed facial keypoint extracting unit is also based on CNNs. The keypoint extracting module architecture can be seen on the Figure 4.2. It is constructed out of three sequential convolutional layers with max-pooling layers in between. The number of convolutional filters doubles with each layer. The output of the third convolutional block is flattened and propagated to the three fully-connected layers. Dropout regularization was used between the last two fully-connected layers. The output of this network is 30-dimensional vector representing 15 (x,y) pairs of facial keypoints. This network was pre-trained on Kaggle Facial Keypoint Detection dataset and achieved mean squared detection error on validation set of 0.0016. The labels were normalized before the beginning of training.

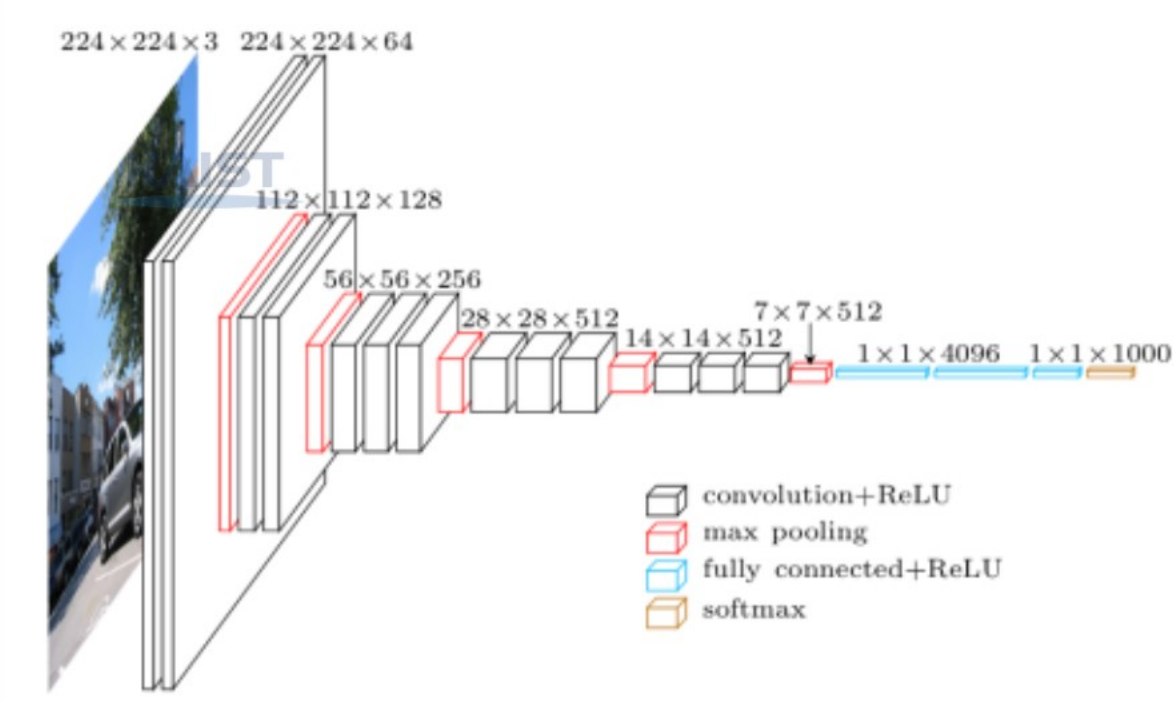

Figure 4.1: VGG16 model architecture (from [27]).

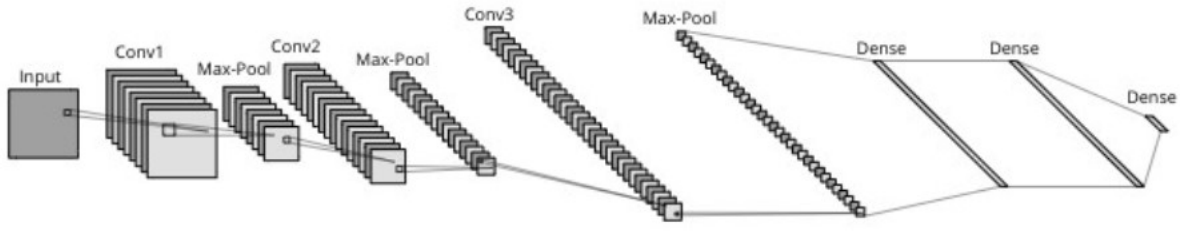

Figure 4.2: Facial keypoint extracting module architecture.

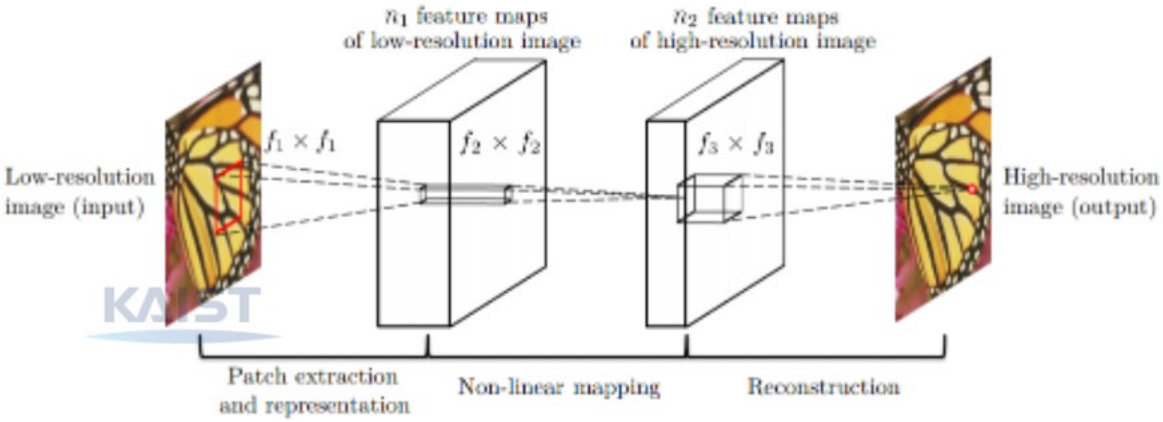

Figure 4.3: SRCNN architecture (from [28]).

4.1.3 Super Resolution

Super resolution is a technique of increasing a resolution of a given image. The easiest way to do it is to apply bicubic interpolation process to the image, however this process is known to output not a high quality high-resolution images due to presence of the noise resulting in a blur effect on an image. One of the ways to overcome the problem mentioned before is to apply machine learning algorithms. Particularly, Dong et al.[28] proposed super-resolution convolutional neural network (SRCNN) architecture to generate upscaled images. The architecture of this network is shown on the Figure 4.3.

First the image is upscaled to the desired size (upscaling coefficient is 4 to obtain the desired resolution of input image of 192x192) using bicubic interpolation. This upscaled image is then put to the convolution layer in order to learn its representation. Then it proceeds to the mapping layer which is followed by reconstruction layer.

In this experiment, the super-resolution block was pre-trained on RaFD[29] dataset. Unlike other datasets, this dataset consists only out of high-resolution face images representing different facial expressions from various viewing points. This is the ideal dataset to use in order to learn needed features for facial super resolution upscaling. All images in this dataset are rectangular. However the faces are aligned which allows to perform square cropping to get 681x681 square images containing faces. These images are then downscaled to 192x192 and 48x48 images representing high-resolution (HR) and low-resolution (LP) training datasets, respectively.

4.2 Method

4.2.1 Overview

The overall architecture is shown on the Figure 4.4. First, the input image is put into the super resolution unit. The output of this unit is then proceeds to the Feature and Keypoints Extraction units. The extracted features from the pre-trained lower layers of the VGG16 object classifier are then passed into the chain of convolutional layers for fine-tuning the extracted features. Then the final extracted features are formed into a Primary capsules which finally advance into the Emotion capsules. The Emotion capsules are used for the classification purpose. The capsule with the greatest length wins, and therefore labels an image with the corresponding emotion. The evaluation network is used in order

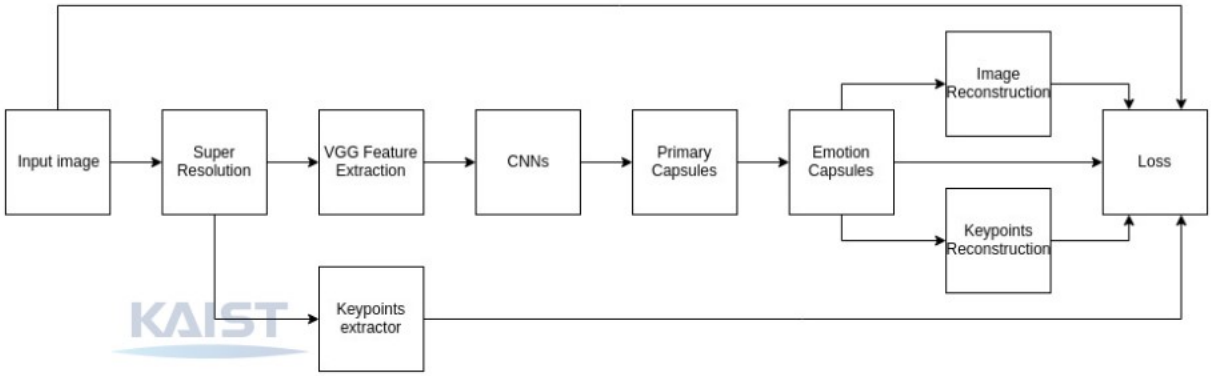

Figure 4.4: Proposed overall architecture.

to force capsules to learn critical information about facial expression. The output of the classification network, the Emotion Capsule, is used as an input to Image and Keypoint Reconstruction units. The closer to the ground truth the reconstructed images and keypoints get, the better. These reconstruction losses and the classification loss are combined and used as a total loss for training purpose.

4.2.2 Details

The super resolution unit is the three layer convolutional network with 128, 64 and 1 filters in each layer, respectively. The network takes already pre-scaled (using bicubic interpolation) image (in the case of the experiment, scale factor is set to be 4) and outputs the same size output high-resolution 192x192 image. This super-resolution unit was pre-trained on RaFD dataset and the weights of it are fixed while the overall emotion classifier is trained. In order to accelerate training process, the pre-processed high-resolution fer2013 images were stored beforehand the experiment.

The next step of the experiment is to obtain 15 (x,y) pairs of crucial keypoints of the face. The keypoint extraction unit contains three convolutional layers with 32, 64, and 128 filters, respectively. Each convolution operation is followed by max-pooling operation with the size of the pool being set to 2. The output of the final max-pooling layer is flattened and proceeded to the three sequential fully-connected layers with 500, 500 and 30 activation units, respectively. The last 30 units represent 15 points on the image plane. The input to this network should have dimensions of 96x96, therefore, the output of the super-resolution unit is downscaled by the factor of 2. This unit is trained on the Kaggle Facial Keypoints Detection dataset. During the overall training of emotion classifier, the weights remains unchanged. Additionally, to accelerate the process, the pre-detected facial keypoints of each fer2013 image ware stored separately in a text file.

The feature extraction unit is composed out of two parts. First, the pre-obtained high-resolution image is inputted into the VGG16 classifier. The initial high-resolution image has 192x192x1 size and VGG16 classifier requires three channels of the image (RGB channels) for its classification. Therefore, the only channel of the gray scale image is tripled to represent gray scale image in RGB color space. The image is processed through first three convolutional blocks of VGG16 classifier and the output is 256 48x48 feature maps. During the training process, the weights of the VGG16 classifier remain unchanged. These 256 feature maps become an input to the next sequence of four convolutional layers. Each layer has 256 filters for feature map extraction. The weights of the convolutional filters are not fixed and are learned during the process of overall training.

The obtained feature maps are transformed into 32 12x12x8-dimensional primary capsules which are then viewed as 4608 8-dimensional input vectors to the emotion capsules. There are 7 emotion capsules (each representing one of the six universal facial expressions plus neutral face) and each of them has 16 axis of freedom. The communication between capsules is performed using dynamic routing algorithm (the routing number is set to be 3).

Finally, the length of each capsule is computed using L2-norm. The capsule with the greatest length wins and it corresponds to the presence of the certain trait on the input image. In the case of the experiment, the traits are six universal facial expressions and a neutral face.

The additional two networks are image and keypoints reconstruction units. Both of them have a role of regularizers. The first network (image reconstruction) takes as an input the targeted capsule (the capsule corresponding to the true label) and transforms it into 48x48 input using four fully-connected layers with 512, 512, 1024, and 2304 units, respectively. The output of this network is reshaped into 48x48 input and is compared to the initial input image. In this step the reconstruction error is computed as a sum of square errors (the distance between reconstructed and original pixels).

The second network encourages capsules to learn emotional variance of the image. It takes the targeted capsule, which can be viewed as encoded image, and transforms it into 15 facial keypoints. The architecture of this unit is composed out of three fully-connected layers with 512, 512 and 30 units, respectively. The keypoint reconstruction error is computed by comparing pre-detected and reconstructed points using mean square error.

The total loss for overall training is computed as a sum of marginal, reconstruction, and keypoint reconstruction losses scaled by pre-determined factors.

$$L_{total} = L_{marginal} + \alpha L_{reconstruction} + \beta L_{keypoint}$$

where α was set to be 0.0005 and β to be 0.001. The value of α controls the spatial features learning (rotation, scale, etc.), while β regulates emotion action units (inner brow riser, outer brow riser, brow lower, lip corner ruler, lip corner depressor, mouth stretcher, etc.).

Chapter 5. Data

Throughout this experiment, multiple datasets were used. All of these datasets are related to face emotion recognition and facial keypoints detection tasks. This section describes the datasets used for the experiment and the purpose of their application. In addition, the second part of this section will demonstrate the reasons why Kaggle's open source fer2013 dataset is challenging.

5.1 Dataset

5.1.1 fer2013 Dataset

In 2013, Goodfellow et al.[30] introduced a new dataset for the Kaggle facial expression recognition challenge. All images were obtained from the internet. The dataset consists of 35,887 images. Every image is constructed out of 48x48 gray scale pixels representing face and is labeled with one on the seven category labels: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral. The sample images can be seen on the Figure 5.1.

This dataset was used for the main purpose of this research, emotion classification. The other two datasets are auxiliary.

5.1.2 Kaggle Facial Keypoints Detection Dataset

In 2016, Bengio et al.[31] provided a new dataset for the Kaggle facial keypoints detection challenge. The dataset consist out of 7,049 and 1783 gray scale 96x96 pixel training and test images, respectively. Each image has 15 (x,y) labels, representing real-valued pairs in the space of pixels. These representations are the following keypoints of the face: left and right eye centers (2), left and right inner and outer eye corners (4), left and right inner and outer eyebrow ends (4), nose tip (1), left and right mouth corners (2), and center of the top and bottom lips (2). The sample images of the dataset can be seen on the Figure 5.2.

This dataset was used to train a facial keypoint detection network.

5.1.3 Radboud Faces Database

The Radboud Faces Database (RaFD)[29] is provided by the Radboud University Nijmegen. This dataset consist out of images of 67 models displaying 8 facial expressions (anger, disgust, fear, happiness,

Figure 5.1: Sample of images from fer2013 dataset.

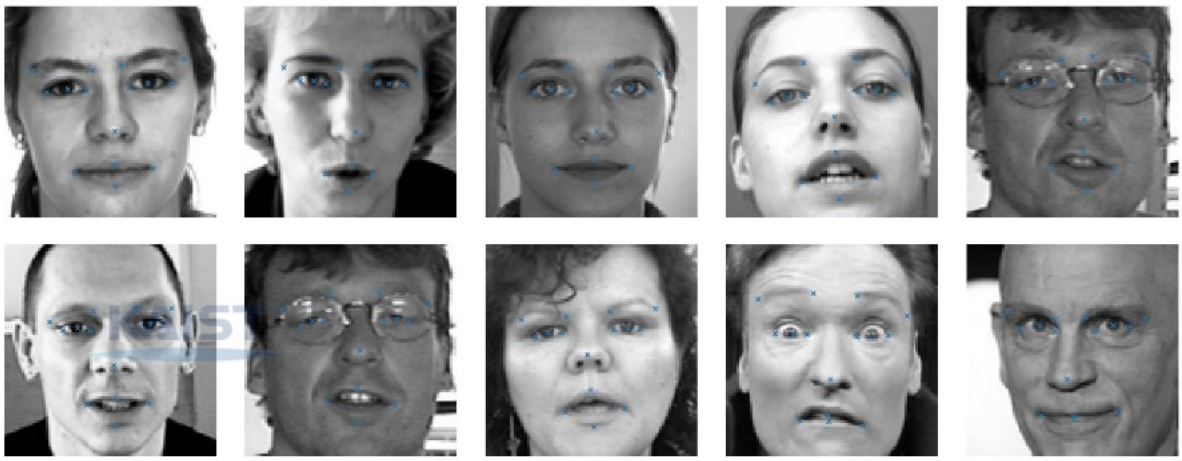

Figure 5.2: Sample of images from Kaggle Facial Keypoints Detection dataset.

Figure 5.3: Sample of images from RaFD dataset.

sadness, surprise, contempt, and neutral). Each emotion is presented with three different gaze directions and it image was taken from five different angles (full face, left and right 45 degrees rotation, and left and right profiles). The dataset has 8,040 high-resolution (681x1024) color images. The sample images of the dataset can be seen on the Figure 5.3.

This dataset was used to train a super resolution network.

5.1.4 Other Existing Datasets

There are many other existing datasets for facial expression recognition task. Some of the popular datasets are Extended Cohn-Kanade Dataset (CK+)[32] which consists out of 593 high-resolution gray scale image sequences representing eight universal emotions, Japanese Female Facial Expression (JAFFE)[33] which consist out of 213 static gray scale 256x256 resolution images classified between seven different classes of universal facial expression (including neutral facial expression). Facial Expression Research Group Dataset (FERG)[34] contains 55,767 high resolution color images of generated stylized characters representing six universal emotions. However the last dataset only contains the images of six characters and the images are generated from the short clips representing emotions. Finally, the AffectNet[35] is a recent dataset which contains more than a million different various size color images labeled with the eight universal emotions. This dataset was created using search engines with 1250 emotion related keywords.

Unlike other datasets, fer2013 has a compact size while containing a large number of fixed-size elements resembling wild settings. Therefore, due to its advantages over other datasets, it was selected for the experiment.

Figure 5.4: Sample of unacceptable images from fer2013 dataset.

Figure 5.5: Sample of mislabeled images from fer2013 dataset (from [36]).

5.2 Difficulty of fer2013 Dataset

The datasets was constructed with the images obtained from the internet. However, some of the images were mistakenly chosen to represent face. Among those pictures, the images of text, numbers, geometric figures, group of people, plain background, cartoon characters, faces covered with other objects, faces covered with hands and others could be found. Figure 5.4 displays some of the unacceptable images.

Apart from the presence of wrongly selected images, some of the dataset components are also mislabeled. As shown on the Figure 5.5, some of the images which a person can easily classify, are labeled with obviously incorrect label [36]. The upper label is taken from the original fer2013 label list, while the lower one is obtained from the public survey.

In addition, all images in the fer2013 dataset are not aligned, and taking into account the fact of a gray scale and low resolution of the images, it makes the task of representation learning more difficult for this dataset.

Finally, the dataset is not equally distributed. Some of the classes have a vast number of images while the other classes barely contain any images. The number of elements of each class are as follows: angry (4953), disgust (547), fear (5121), happy (8989), sad (6077), surprise (4002), and neutral (6198). Only 1.5% of the entire dataset images belong to 'disgust' class, while 25% of it are are contained within 'happy' class.

Chapter 6. Results

This section provides numerical results of fer2013 dataset classification. The second section provides examples of reconstructed images, while the third section shows examples of learned variance of facial expressions.

6.1 Emotion Classification

For this experiment we compared performance of the proposed method to the several existing classifiers and to the baseline. We chose a baseline to be a convolutional neural network with a VGG-like architecture. It is constructed out of four blocks, where each block contains convolutional layer, batch normalization layer, another convolutional layer, follower by batch normalization and max-pooling layers. The output of the fourth block is followed by two fully-connected layers. The results of image classification are shown in the Table 6.1.

The accuracy of the baseline model is higher than originally proposed Capsule network (with the same architecture), however our proposed method outperforms the baseline and some of the state-of-the-art object classification models.

6.2 Face Reconstruction

As a part of the experiment, the image reconstruction unit was also trained. Figure 6.1 shows the examples of reconstructed images. The input images (shown above) are classified with the emotion classification network, and the output classification capsule is used as an input to reconstruct image (shown below).

Due to complexity of a human face, the face reconstruction unit could not learn the fine characteristics of the face (wrinkles, position of the teeth, freckles, etc.), however, all reconstructed faces belonging to one class have a similar traits, which could be recognized by a human.

It is visible that face reconstruction unit learned spatial variance of the images (face angle, face size, etc.), however, as shown in the Section 6.3, the decoding unit also learned emotional variance.

6.3 Emotion Disentanglement

The final experiment was conducted to determine whether the Capsule network learned the facial expression variance. Once the classification capsule is obtained, we check the appearance of the reconstructed images when the dimensions of the capsules are perturbed. A random noise, varying from -0.5 to 0.5, was added to one of the dimensions of the capsule and then the capsule was used as an input to reconstruction unit. The results are shown on the Figure 6.2.

The Capsule network learned several different features. For instance, on Figure 6.2(a), it is visible that one dimension of the capsule corresponds to FACS-4[3] (brow lowerer), which is the key action unit for 'anger' emotion. 6.2(b) shows that capsule learned FACS-1 (inner brow raiser), FACS-2 (outer brow raiser) and FACS-27 (mouth stretcher), which are the key action units of the 'fear' emotion. 6.2(c) displays a correlation between one of the dimensions and FACS-12 (lip corner puller), activation of which

(a) disgust (b) surprise

(c) fear (d) happy

(e) angry (f) sad

(g) neutral

Figure 6.1: Real and reconstructed images for different emotions.

Table 6.1: Classification accuracy (%) on fer2013.

Method	Accuracy, %
Human accuracy	65±5
Baseline	66.20
VGG16 [20]	68.44
InceptionV3 [22]	68.12
CapsNet [7]	64.71
Modified CapsNet + Keypoints	69.23
Proposed method	70.68

could be seen in 'happiness' emotion. 6.2(d) shows the correlation between perturbation of one dimension and FACS-15 (lip corner depressor), which is the main action unit of the 'sad' emotion. Finally, 6.2(e) shows that the capsule could learn 'surprise' emotion. This example shows that one capsule could learn the activation of FACS-1, FACS-2, FACS-27.

Despite the fact that the universal facial expressions share many common traits (for instance, 'fear' and 'surprise' share four action units), it is possible for a human to see that the reconstructed images differ from each other depending on their class. This proves that the Capsule network could learn the emotional variance of the face.

(a) anger

(b) fear

(c) happiness

(d) sadness

(e) surprise

Figure 6.2: Emotion capsule dimension perturbation.

Chapter 7. Conclusion

In this thesis, we proposed a model to classify universal facial expressions which could learn emotional variance. The proposed model consists out of three pre-processing units (super resolution, feature extraction, and keypoint detection) and two layers of capsules. The experiment showed that the proposed model outperform the baseline and some of existing state-of-the-art object classifiers. One of the experiments also proved the hypothesis that Capsule networks could learn the emotion action units despite the fact that no examples were provided. To the best of our knowledge, this is the first approach to learn the origin of the emotion formation using deep neural networks. Unlike other approaches of emotion classification, our method showed that capsule network could generalize the variance of each class. For the future work, we have a plan to investigate the way of enforcing capsules to separately learn different emotion action units.

Bibliography

[1] A. Mehrabian, *Communication without words, Communication theory*, 2008.

[2] Ekman P. Universals and cultural differences in facial expressions of emotion. University of Nebraska Press; Lincoln, NE: 1972.

[3] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, 1978.

[4] Schlosberg, H. (1954). "Three dimensions of emotion". Psychological Review. 61: 81–8.

[5] Wenyun Sun, Haitao Zhao, Zhong Jin. An Efficient Unconstrained Facial Expression Recognition Algorithm based on Stack Binarized Auto-encoders and Binarized Neural Networks. Neurocomputing, 2017.

[6] CA Corneanu, MO Simón, JF Cohn, et al. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. IEEE Transactions on Pattern Analysis Machine Intelligence, 2016.

[7] S. Sabour, N. Frosst, G. Hinton. "Dynamic Routing Between Capsules". In: Advances in Neural Information Processing Systems (NIPS), pp 3859-3869, 2017.

[8] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. "Gradient-Based Learning Applied to Document Recognition". Proceeding of the IEEE, November 1998.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus. "Intriguing properties of neural networks", arXiv:1312.6199, 2013

[10] G. Hinton, A. Krizhevsky, S. Wang. "Transforming Auto-Encoders". In: International Conference on Artificial Neural Networks, pages 44-51. Springer, 2011.

[11] Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits, 1998.

[12] P. Hensman, D. Masko. "The Impact of Imbalanced Training Data for Convolutional Neural Networks", 2015

[13] F. Khan. "Facial Expression Recognition using Facial Landmark Detection and Feature Extraction via Neural Networks", 2018

[14] K. Zhao, W.-S. Chu, F. Torre, J. F. Cohn, and Z. H, "Joint patch and multi-label learning for facial action unit detection," In CVPR, 2015.

[15] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," In CoRR, 2018.

[16] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," In ECCV, 2018.

[17] I. Tautkute, T. Trzcinski. "Classifying and Visualizing Emotions with Emotional DAN", 2018

[18] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images," in Proc. IEEE Int. Conference on Computer Vision (ICCV), 2015, pp. 3631–3639.

[19] C. Pramerdorfer, M. Kampel. "Facial Expression Recognition using Convolutional Neural Networks: State of the Art", 2016

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", arXiv:1512.03385, 2015

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567

[23] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015. paper — bibtex — paper content on arxiv — attribute annotations

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". arXiv:1506.02640, 2015

[27] https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/

[28] Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang. Image Super-Resolution Using Deep Convolutional Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Preprint, 2015

[29] 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

[30] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. 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, "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

[31] https://www.kaggle.com/c/facial-keypoints-detection

[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression," in 3rd IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010.

[33] M. J. Lyons, M. Kamachi and J. Gyoba, "Japanese Female Facial Expressions (JAFFE)," Database of digital images, 1997.

[34] "Facial Expression Research Group Database (FERG-DB)". grail.cs.washington.edu. Retrieved 2016-12-06.

[35] Mollahosseini, A.; Hasani, B.; Mahoor, M. H. (2017). "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". IEEE Transactions on Affective Computing. PP (99): 1–1. doi:10.1109/TAFFC.2017.2740923. ISSN 1949-3045.

[36] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer and Zhengyou Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution", arXiv:1608.01041v2 [cs.CV], Sep. 2016.