Regular paper

Journal of the KIECS. pp. 75-80, vol. 16, no. 1, Feb. 28. 2021, t. 105, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2021.16.1.75

저해상도 영상 자료를 사용하는 얼굴 표정 인식을 위한 소규모 심층 합성곱 신경망 모델 설계

살리모프 시로지딘* · 류재홍**

A Design of Small Scale Deep CNN Model for Facial Expression Recognition using the Low Resolution Image Datasets

Sirojiddin Salimov* · Jae Hung Yoo**

요 약

인공 지능은 놀라운 혜택을 제공하는 우리 삶의 중요한 부분이 되고 있다. 이와 관련하여 얼굴 표정 인식은 최근 수십 년 동안 컴퓨터 비전 연구자들 사이에서 뜨거운 주제 중 하나였다. 저해상도 이미지의 작은 데이터 세트를 분류하려면 새로운 소규모 심층 합성곱 신경망 모델을 개발해야 한다. 이를 위해 소규모 데이터 세트에 적합한 방법을 제안한다. 이 모델은 기존 심층 합성곱 신경망 모델에 비해 총 학습 가능 가중치 측면에서 메모리의 일부만 사용하지만 FER2013 및 FERPlus 데이터 세트에서 매우 유사한 결과를 보여준다.

ABSTRACT

Artificial intelligence is becoming an important part of our lives providing incredible benefits. In this respect, facial expression recognition has been one of the hot topics among computer vision researchers in recent decades. Classifying small dataset of low resolution images requires the development of a new small scale deep CNN model. To do this, we propose a method suitable for small datasets. Compared to the traditional deep CNN models, this model uses only a fraction of the memory in terms of total learnable weights, but it shows very similar results for the FER2013 and FERPlus datasets.

키워드

Convolutional Neural Networks, Facial Expression Recognition, Design of CNN architecture, Low Resolution Image
합성곱 신경망, 얼굴 표정 인식, 합성곱 신경망 구조 설계, 저해상도 영상

Ⅰ. Introduction

Artificial Intelligence has important roles in our lives providing incredible benefits to humanity. This requires researchers to develop new and better models. In this perspective, the demand for facial expression recognition systems has been increasing, since many companies require automated systems with high performance. Many researchers have been working on image classification systems on

* 전남대학교 컴퓨터공학과 석사과정 재학
(sirojiddin1224@gmail.com)
** 교신저자: 전남대학교 컴퓨터공학과
· 접 수 일 : 2020. 12. 02
· 수정완료일 : 2021. 01. 09
· 게재확정일 : 2021. 02. 17
· Received : Dec. 02. 2020, Revised : Jan. 09, 2021, Accepted : Feb. 17, 2021
· Corresponding Author : Jae-Hung Yoo
Dept. of Computer Engineering, Chonnam Nat. Univ.
Email : jhy@jnu.ac.kr

big dataset of high resolution images. Classifying small dataset of low resolution images would need small scale model development. To this point, we believe that a design of small scale model is very crucial domain of research giving us motivation to tackle the current issue.

Pictures of customers are taken and our system tries to classify customers' mood. This will evaluate customers' satisfaction and provide the reference for the customer relationship management (CRM). Installed camera might be a bit far from customer, our task would be classifying face expressions using low resolution images.

In this paper, a small scale CNN networks are investigated for the image classification problems. In section II, typical CNN are introduced. In section III, image datasets are reviewed. In section IV, proposed architecture is explained. In section V, experimental results show the effectiveness of the proposed method followed by the conclusion and reference sections[1-12].

II. Related works

Proposing CNN models was the huge breakthrough in Computer Vision field enabling and encouraging researchers to dig into this area deeper and explore more possibilities. In this point of view, we assume that mentioning the first CNN model would be appropriate. The first CNN model is Alex Net [1] demonstrated how effective are CNNs. The Alex Net used Imagenet dataset. Followed by Alex Net many other CNN models are proposed later. Later in 2018, SE-ResNet model [2] is considered as the state of the model so far to tackle the VGG-face dataset. However, these models that depicted promising results on Imagenet and VGG-face (224x224x3 input size) datasets with the cardinality over one million are not the right ones on FER2013 and FERPlus datasets, both contains 48x48x1 images with the cardinality of 35887.

State of the art CNN-based work on FER2013 dataset is fine-tuned VGG-face showing accuracy of 72.11%[3]. Latest works with very promising results on FERPlus dataset have been reported[4-6].

III. Datasets

3.1 FER2013

Originally this dataset is announced in Kaggle challenge, in 2013. The highest score was 71.16% in the competition. The dataset contains small face images. The dataset contains faces with different poses. This makes dataset more challenging.

Table 1 shows the distribution of both FER2013 and FERPlus datasets. The 3rd row in the Table 1 represents the majority training mode[6]. They filtered images with certain rules and follow their rules to sort out the reduced testing images

Table 1. Cardinality of FER2013 and FERPlus datasets

	Training	Validation	Testing
FER2013	28709	3589	3589
FER+	28561	3579	3574
majority mode[6]	25060	3199	3153

One might wonder why the highest accuracy was only 71.16%. The reason is that the dataset contains low resolution images (48x48), invalid images (not face images), wrongly labelled images (given label to a particular image is wrong), partly correct images (it is hard to decide label for an image). Thus, the highest performance on this dataset was not higher than 71.16% and even works proposed later (not in the challenge) are not high. Here are some examples for invalid images.

Fig. 1 Invalid example images from FER2013

3.2 FERPlus

This dataset contains almost same images that FER2013 dataset has. The difference is that invalid images are deleted and relabeled and FERPlus dataset is created. Here are some example cases from [6] of how authors of FERPlus relabeled the images from FER2013. Top labels are from FER2013 and bottom labels are from FERPlus.

Fig. 2 Every single face image has two labels below it; top ones are showing FER2013 labels, while below ones are showing labels for FERPlus

Ⅳ. Proposed Architecture

Normally, CNN architectures start model layers with 5x5 or 3x3 to extract more general features first. However, unlike other models, we utilize very small window sizes from the initial convolutional layers. Our assumption is that since given images have a very small window size – 48x48, we might lose some important data by performing 5x5 or 3x3 kernel size convolutions. Therefore, we used 2x2 size windows to perform convolution operation for all layers.

The model has 6 convolutional layers, 3 fully connected layers and the output layer. Output layer represents number of classes, which is 7 in case of FER2013 dataset and 8 in case of FERPlus dataset. Each layer has convolutional layer followed by max pooling layer, except the first layer. The pooling decreases number of features for computing faster, however we wanted to keep initial features and thus didn't perform pooling operation after first layer.

After convolutional layers, we utilized 3 fully connected layers. All fully connected layers have 512 nodes. The last layer is output layer, which has 7 nodes for FER2013 dataset and has 8 nodes for FERPlus dataset.

Number of filters in the first layer is 32, second layer has 64 filters, while later layers have 128, 256, 512 and 1024 filters respectively. Number of filters are doubled compared to previous layer. The model uses only a fraction of the memory in terms of total learnable weights as in the next table and is given in the next figure.

Table 2. Model complexity by the number of layers and parameters

Model	No. of Layers	No. of Parameters
AlexNet[1]	8	61M
SENet-50[2,5]	50	27.5M
VGG16[3,4]	17	138M
ResNet18[4]	18	11M
VGG16[4]	17	138M
VGG13[6]	14	133M
Our model	10	**3.8M**

Ⅴ. Experimental Results

To find the optimal model on FER2013 dataset we made experiments on different kernel sizes and layers with augmented data. Next table shows the results.

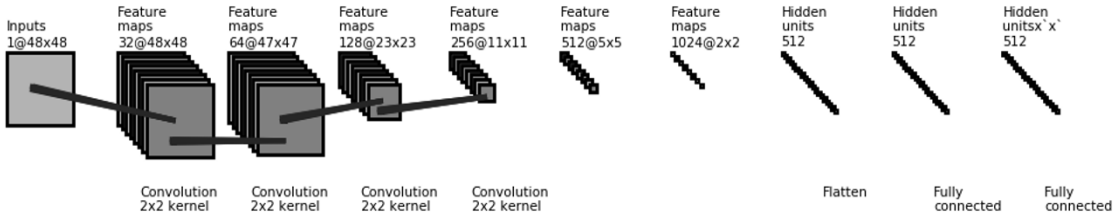

Fig. 3 Our model architecture

Table 3. Experiments on FER2013 to find the model using augmented images

Layers	Kernel size			
	2	3	4	5
2	57.64	59.48	62.05	61.35
3	64.67	64.97	66.03	66.59
4	65.64	67.67	68.34	67.15
5	68.15	68.17	67.11	67.86
6	**69.32**	68.29	68.24	68.05
7	66.98	67.6	66.85	67.32

As it is depicted, utilizing 6 convolutional layers and using 2by2 windows in each layer shows the highest score.

Here are our model's learning curves for FER2013 dataset. We saved the weight set of model that achieved the best performance on validation data.

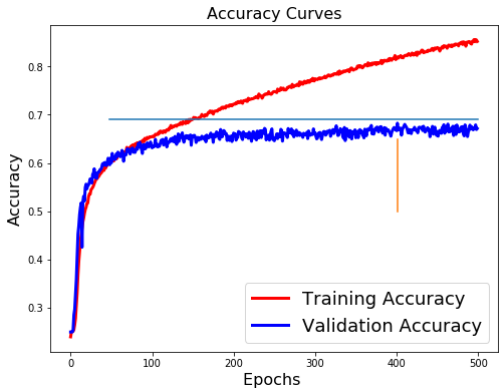

Fig. 4 Learning curves for FER2013 dataset

Without augmentation our model showed 60.82% testing data accuracy for FER2013, training data accuracy was over 98% on the epoch 50. We did data augmentation to deal with overfitting problem. With the data augmentation our model depicted 69.32% accuracy. Training data accuracy was 83.95% on the epoch 500.

We next show the comparison of our model and previous best models on FER2013 and FERPlus datasets. Among CNN models, the best model so far is fine-tuned VGG-face with accuracy of 72.11%[3].

Table 4. Model accuracy on FER2013 dataset

Model	Accuracy (%)
Fine-tuned VGG-face (aug.) [3]	72.11
Our model	69.32

The models with the highest performances on FERPlus are Kai Wang et al. [4] Comparison results are given in the following table.

Table 5. Model accuracy on FERPlus dataset

Model	Accuracy (%)
RAN-VGG16[4]	89.16
SENet-50[5]	88.8
RAN-ResNet18[4]	88.55
Our model	85.35
VGG13[6]	85.1
Fine-tuned VGG-face (aug.) [3]	84.79

Confusion matrices of our model on FER2013 and FER2013 datasets are provided to demonstrate how well our model is.

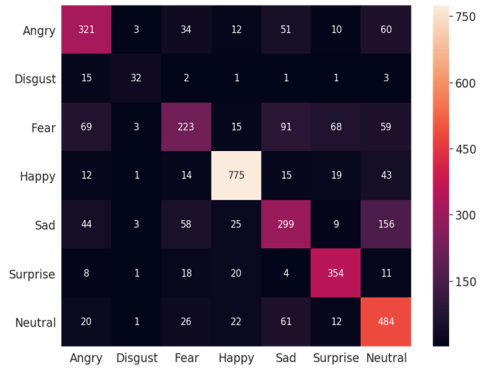

Fig. 5 Confusion matrix for FER2013

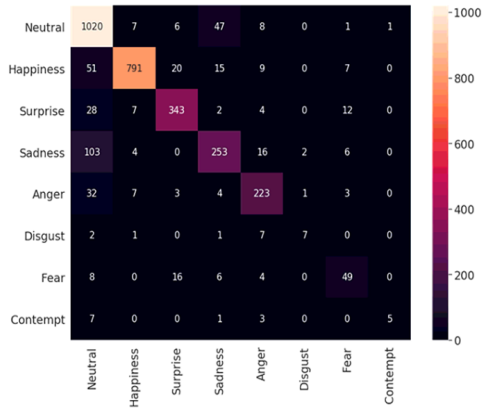

Fig. 6 Confusion matrix for FERPlus

As it's shown in the tables above, our model is highly competitive with other models on both datasets. In the figure 5 and 6 confusion matrices are demonstrating which classes are learned well by our model. It's noticeable that model is being confused while predicting "sad" class samples as "neutral" and "neutral" classes as "sad". Indeed, many of the images that are labelled as "neutral" and "sad" are very similar and hard to classify even by human.

Ⅵ. Conclusions

The proposed model uses only a fraction of the memory in terms of total learnable weights compared to the conventional CNN models. The model is highly competitive showing good performance for FER2013 and FerPlus datasets. We plan to explore the possibility of transfer learning for these datasets.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", In *Advances in Neural Information Processing Systems 25: 26th Annual Conf. on Neural Information Processing Systems,* New York, USA, Dec. 2012, pp. 1097-1105.

[2] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," In *IEEE Conf. on Computer Vision and Pattern Recognition,* Salt Lake City, USA, 2018, pp. 7132-7141.

[3] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local Learning with Deep and Handcrafted Features for Facial Expression Recognition," *IEEE Access,* vol. 7, 2019, pp. 64827-64836.

[4] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE Trans. on Image Processing,* vol. 29, 2020, pp. 4057-4069.

[5] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion Recognition in Speech using Cross-Modal Transfer in the Wild," In *Proc. of the 26th ACM Int. conf. on Multimedia,* Seoul, Korea, 2018, pp. 292-301.

[6] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," In *Proc. of the 18th ACM Inte. Conf. on Multimodal Interaction,* Tokyo, Japan, 2016, pp. 279-283.

[7] J. Yoo, "An Extension of Unified Bayesian

Tikhonov Regularization Method and Application to Image Restoration," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 01, 2020, pp. 161-166.
[8] J. Yoo, "A Unified Bayesian Tikhonov Regularization Method for Image Restoration," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 11, 2016, pp. 1129-1134.
[9] J. Yoo, "Self-Regularization Method for Image Restoration," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 1, 2016, pp. 45-52.
[10] S. Kim and D. Cho, "Design and Implementation of Hashtag Recommendation System Based on Image Label Extraction using Deep Learning," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 4, 2020, pp. 709-716.
[11] Y. Kim, D. Kim, and S. Park, "Research on Robust Face Recognition against Lighting Variation using CNN," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 12, no. 2, 2017, pp. 325-330.
[12] M. Ying and K. Kim, "CNN Based 2D and 2.5D Face Recognition For Home Security System," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 14, no. 6, 2019, pp. 1207-1214.

저자 소개

살리모프 시로지딘(Sirojiddin Salimov)

2016년 Tashkent University of Information, Software Engineering공학과 졸업(공학사)
2020년 현재 전남대학교 대학원 컴퓨터공학과 재학
※ 관심분야 : 심층 신경망, 영상처리, 컴퓨터 비전

류재흥(Jae-Hung Yoo)

1981년 한양대학교 기계공학과 졸업(공학사) (BE in Mechanical Engineering from Hanyang Univ. in 1981)
1986년 디트로이트 대학교 대학원 전산학과 졸업 (MA in Computer Science from Univ. of Detroit in 1986)
1993년 웨인주립 대학교 대학원 전산학과 졸업 (PhD in Computer Science from Wayne State Univ. in 1993)
1994년 여수대학교 컴퓨터공학과 교수 (Joined as a faculty member in the Dept. of Computer Engineering, Yosu Nat. Univ. in 1994)
2006년～현재 : 전남대학교 컴퓨터공학과 교수 (Became a faculty member in the Dept. of Computer Engineering, Chonnam Nat. Univ. in 2006)
※ 관심분야 : 인공신경망, 패턴인식, 기계학습, 영상처리 및 컴퓨터비전 (Main research areas: Artificial Neural Networks, Pattern Recognition, Machine Learning, Image Processing and Computer Vision)