합성곱 신경망과 장단기 기억 신경망을 이용한 이미지 내에서의 집단 수준의 행복한 감정의 강도 예측

박규태, 곽노준
융합과학기술대학원, 서울대학교
pgt4861@snu.ac.kr, nojunk@snu.ac.kr

Abstract

In this paper, we describe our method to predict group-level emotion intensity in the wild images. The group-level emotion recognition task is to infer a happiness mood intensity of the group image as a whole. This work was motivated by group-level emotion recognition sub-challenge of EmotiW2016 (Emotion recognition in the Wild). After detecting faces in an image, we use Convolutional Neural Network (CNN) to predict happiness intensities of faces. Then, we combined Long Short Term Memory (LSTM) to predict group-level happiness intensity of whole image from the outputs of CNN model. At the final stage, we ensemble several LSTM models, which resulted in improved Root Mean Square Error (RMSE) performance. The final achievement we made is 0.918 RMSE on the HAPPEI test dataset, which shows large improvement over the baseline 1.30 RMSE.

1. Introduction

Image analysis is to extract some useful and meaningful information from images. One of the image analysis which have attracted many researchers' attention for recent years is emotion recognition tasks. Recognizing user emotion from visual data has a great potential in a variety of fields, including marketing, entertainment and medical applications. Especially, this can be applied to social network service (SNS) which means that the platform to socially engage and interact with a large population. As the number of SNS users has been increasing dramatically, millions of images are being uploaded every day on the web. According to this trend, the demand for a system that can understand the emotional attributes of people from image data is likely to increase.

In past years, most of the works on emotion recognition was to predict the type of facial expressions (e.g. anger, disgust, happy, neutral etc.) from face images that were created from experimental environment. This can be seen as a classification problem for relatively simple facial expression image data like CK+ [11], FEEDTUM [12] and MultiPIE [13]. In recent years, the emotion recognition tasks were extended to predict facial expression not on conditioned images but on the wild image [14]. In the case of images in a wild environment, it is more difficult problem because it involves various pose, illumination, scale and occlusion. Furthermore, the EmotiW2016 sub-challenge [1] suggests new problem, which is defined that predicting group-level emotion intensity from whole image. These images came from Flickr with searching keyword related to 'happiness' and include more than two people in the wild images.

In this paper, we will describe our work for predicting group happiness intensity in the wild images on the HAPPEI dataset using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Details of the proposed method will be described in the following chapters.

2. Proposed Method

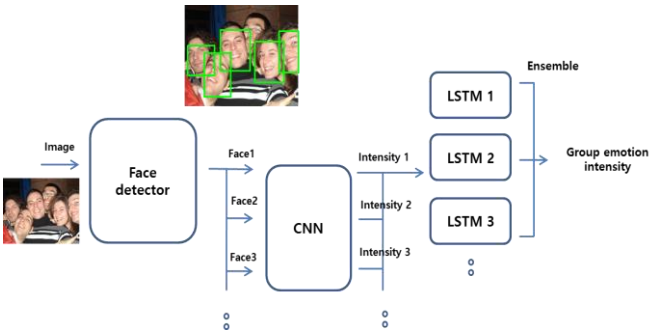

Figure 1. Our overall system for recognizing group-level happiness intensity in a wild image.

The overview of our proposed method is depicted in figure 1. We started from the assumption that each facial expression contained within the input image determines the entire mood. First, input image passes through the face detector and outputs regions of detected faces. To make a accurate face detector, we trained the WIDER Face [5] dataset with Faster-RCNN [4]. All detected faces on input image are cropped and resized in order to be input of the CNN model. The CNN model spits out the happiness intensities of all input faces. Then, these predicted intensities are enter the LSTM model and predict group-level happiness intensity. To decide the order of input

sequence, we calculated a factor for each face. Finally, we use several LSTM models to ensemble the result to increase the performance by performing majority vote.

2.1 Face detector

The first attempt to solve the object detection problem using CNN was [6]. Inspired by CNN's large success in image classification problems, he proposed RCNN. It uses the external region proposal method (e.g. edgebox [15]) to extract the area where the object exists. Then, with cropping and resizing these regions, it performs classification by putting them into the network as input. However, this method has the disadvantage that it is very expensive and slow. This is because it has to feed forward so many proposal areas separately and there is a lot of computation for overlapping areas. In order to alleviate this problem, Fast-RCNN [5] has appeared. In this network, the original image is passed through CNN only once and the proposal regions are cropped on the generated feature maps. Therefore, it can share the computation load of feature extraction and consequently reduce the amount of computation. Both RCNN and Fast-RCNN, however, still use the external region proposal method, which acts as a bottleneck of computation cost. In order to overcome this problem, Faster-RCNN [4] introduces a new layer called the Region Proposal Network (RPN). RPN is a layer in which proposal areas can be generated with anchor box by sliding window in a network. It takes into account different spatial locations, scales, and aspect ratios on feature map. The class prediction for each anchor box and the bounding box coordinates regression are performed simultaneously in the network. This allows the network to achieve much faster speed than previous external region proposal based method.

In our method, it is very important to detect the face and to find the box coordinate exactly in image. It is because happiness intensities from detected faces act as crucial cues to determine the group-level happiness intensity. Inspired by [8], we used Faster R-CNN architecture to make a face detector. To find the faces effectively in the wild images, we trained WIDER Face dataset [7] on this network. It is a face detection benchmark where images are selected from publicly available WIDER dataset. This is composed of 32,203 images and 393,703 annotated faces with 60 events classes, which have a high variability in scale, pose and occlusion.

2.2 Happiness intensity on each face

In HAPPEI (HAPpy PEople Images) dataset [2], the happiness intensity of face is annotated by six labels, which indicate neutral, small smile, large smile, small laugh, large laugh and thrilled respectively. The prediction process of happiness intensity of faces can be regarded as an image classification problem for the detected and cropped faces.

We used Convolutional Neural Network (CNN) with VGG16-layer architecture [9] to classify six emotion intensity labels from each detected face. To train the CNN model, we made additional data, which is consist of only cropped faces and intensity labels from HAPPEI dataset.

Figure 2. Examples for happiness intensity of faces in HAPPEI dataset.

We used the face detector that we made in previous step to extract precise face coordinate from the annotated face bounding box in HAPPIE. Then, we resized them to 224 by 224 to match the input size of the network. The data was divided into 5 equal parts, and 1/5 was used as validation data and the rest was used as train data.

Commonly, it is well known that starting training from well-initialized weights shows much better performance than from random initialized weights. For this reason, before training the face happiness intensity on HAPPEI dataset directly, we initialized VGG16-layers weights by VGG Face description [3]. It is a deep neural network model for face recognition trained on 2.6 Million images of celebrities collected from the web. Both face recognition problem in VGG Face and this emotion intensity classification problem are considered to use common facial features. Therefore, if all the convolutional filters in the VGG Face model were well-trained to extract facial features, it would be better to use this pre-trained weight to our classification problem. Experimentally, the accuracy of the model with random initialize weight on validation data was about 45%, while the accuracy of the model with pre-trained VGG Face model was about 68% on validation data.

2.3 Group happiness intensity analysis

To predict the group-level emotion intensity in the whole image, we need a model that can generate one output from the happiness intensities of all faces in the image. Since each image contains a different number of faces, the model must be able to handle a different number of inputs. In addition, the model should capture the relationship between happiness intensities of faces and the group-level emotion intensity.

The Long Short Term Memory (LSTM) is a kind of Recurrent Neural Network (RNN) architecture proposed by [10]. It resolved the issue of gradient vanishing problem when sequence length is long enough by using a summation of memory status instead of multiplicative

status of vanilla RNN. It was achieved by introducing several gates that decide what information it will throw away from the previous cell state and what information it will store in the cell state. For example, a simple LSTM block includes gates, input gate, output gate and forget gate.

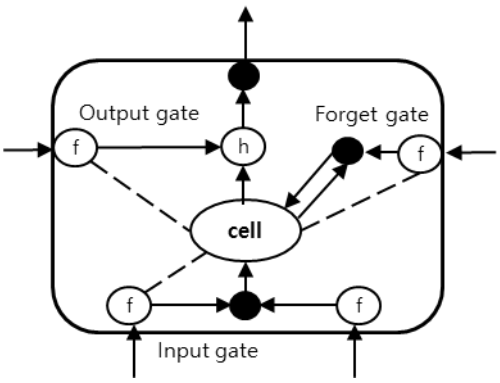

Figure 3. The illustration of simple LSTM unit.

In our method, the LSTM unit is used to predict group-level happiness intensity from each happiness intensity of all detected face. It takes CNN's output vectors as an input. One of the reasons why LSTM is suitable for this task is that it can handle a problem of varying the number of faces in different images. Practically, when we train the model, the max sequence length was set to 10. If the input length is less than 10, the remaining sequences are padded with zeros. The output label is obtained from the last sequence containing the actual value, which is not padded value.

We need to determine the input order of the predicted happiness intensity vector of each face. This order is decided by specific factor, which is calculated by size of the face and Euclidean distance from image center. This factor increases when the relative face size occupied in the image is large and when the face is closer to the image center. Considering this factor can be regarded as mimicking the human perception process. It was motivated by the survey to understand the attributes affecting the perception of the group mood conducted by [2]. According to this survey, when people choose which image looks happier for the two different images, the attributes affecting the selection are large number of people smiling, large smiles of people in the center of the group, and the large size of smiles in the image etc.

3. Experiments and results

3.1 Dataset

Happy People Images (HAPPEI) dataset is composed of images that were collected on Flickr with happy-property keyword based on group of people and events (e.g. 'party + people', 'marriage', 'graduation + ceremony', 'bar', 'convocation', etc.). They were annotated with group level mood intensity. The moods are represented by six stage of happiness: neutral, small smile, large smile, small laugh, large laugh, and thrilled. In addition, for the collected group images, a Viola-Jones face detector [16] was executed on the images and detected faces were also annotated with happiness intensities.

3.2 Sequential Order for LSTM input

To determine what order to put each predicted emotion vector into LSTM model, we consider a factor that reflects how much each face will effect on group emotion intensity in the image. Let W, H is a width and height of image and w, h is a width and height of bounding box of detected face. Cx, Cy is center coordinate of image and cx, cy is center coordinate of face bounding box. The detailed description of these parameters can be found in figure 4.

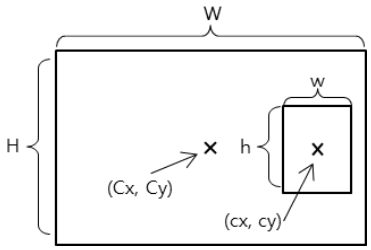

Figure 4. Parameter description for calculating the factor of each detected face.

Then, we can calculate the factor for each bounding box by equation (1).

$$factor = \frac{\frac{w \times h}{W \times H}}{\sqrt{(\frac{Cx - cx}{W})^2 + (\frac{Cy - cy}{H})^2}} \qquad (1)$$

We considered two sequential order setting of LSTM input. In order 1, the emotion intensity vectors of faces are sorted by factor. Specifically, the vector which has the largest factor value is the first input. In order 2, on the other hand, the vectors have random input order. The result is in Table 1. The case of order 1 and order 2 shows similar accuracy performance on the validation set. Order 1, however, shows better RMSE performance over random sequential order for the validation set. This means that if the vectors of faces do not randomly enter the input of LSTM but are sorted by factor, the performance is better.

Table 1. Performance comparison on group-level happiness intensity prediction with input order sequence in LSTM.

	Val RMSE	Val Accuracy
Order 1	0.617	69.071
Order 2	0.655	68.811

3.3 RMSE on test dataset

We achieved 0.965 RMSE using single LSTM model on the HAPPEI test dataset. To boost the performance, we additionally used multiple LSTM model and merged them by majority voting. We changed the hidden unit of the LSTM to 64, 128, and 256, and trained them three times

each to make additional 9 models. This ensemble method improves about 0.02 RMSE on the validation set and about 0.05 RMSE on the test set. Consequently, our final achievement is 0.918 RMSE and this shows largely improvement over baseline RMSE 1.30. Table 2 shows our performance result on HAPPEI test dataset.

Table 2. Performance result on HAPPEI test dataset.

	Test RMSE	Val RMSE
Baseline	1.30	0.78
Our method	0.965	0.617
Our method + Ensemble	0.918	0.594

4. Conclusions

We described our method to predict group-level emotion intensity in the wild images. We combined CNN model that is used to predict happiness intensity of detected face and LSTM model that is used to predict group-level happiness intensity with CNN outputs.

In the first stage, we detected multiple face bounding boxes on image using Faster-RCNN architecture, which is trained on WIDER Face dataset. Second, the detected faces are feed to CNN model, which predict each face happiness intensity. Next, predicted happiness intensity vectors of faces are feed to LSTM model. The input sequential order is determined by factor, which reflects how much each face effect on global emotion intensity in image. Finally, in the ensemble stage, the outputs of several LSTM model select the final group-level happiness intensity by majority voting.

The final achievement we have made is 0.918 RMSE on the HAPPEI test dataset, which shows large improvement over the baseline 1.30 RMSE on test dataset.

Acknowledgements

This work was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program 2016. We are thankful to EmotiW for offering HAPPEI dataset.

References

[1] Dhall, Abhinav, et al. "Emotiw 2016: Video and group-level emotion recognition challenges." *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016.

[2] Dhall, Abhinav, Roland Goecke, and Tom Gedeon. "Automatic group happiness intensity analysis." *IEEE Transactions on Affective Computing* 6.1 (2015): 13-26.

[3] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." *British Machine Vision Conference*. Vol. 1. No. 3. 2015.

[4] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.

[5] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[6] Girshick, Ross, et al. "Region-based convolutional networks for accurate object detection and segmentation." *IEEE transactions on pattern analysis and machine intelligence* 38.1 (2016): 142-158.

[7] Yang, Shuo, et al. "WIDER FACE: A Face Detection Benchmark." *arXiv preprint arXiv:1511.06523* (2015).

[8] Jiang, Huaizu, and Erik Learned-Miller. "Face detection with the faster R-CNN." *arXiv preprint arXiv:1606.03473* (2016).

[9] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[10] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[11] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010..

[12] Wallhoff, Frank. "Facial expressions and emotion database." *Technische Universität München* (2006).

[13] Gross, Ralph, et al. "Multi-pie." *Image and Vision Computing* 28.5 (2010): 807-813.

[14] Dhall, Abhinav. "Collecting large, richly annotated facial-expression databases from movies." (2012).

[15] Zitnick, C. Lawrence, and Piotr Dollár. "Edge boxes: Locating object proposals from edges." *European Conference on Computer Vision*. Springer International Publishing, 2014.

[16] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE, 2001.