Facial Emotion Recognition using Facial Features as an Attention Mechanism

Halvor Kvernes MEEN[1,a) Michifumi Yoshioka[1,b) Katsufumi Inoue[1,c)

Abstract

There has been extensive research on facial emotion recognition, yet it continues to be a difficult task. In this research the goal is to achieve state-of-the-art accuracy on datasets not made in controlled environments by utilizing facial features[*1] as attention maps, as the current state-of-the-art struggles to achieve high accuracy ([8]). The use of facial features is further motivated by the lack of big publicly available datasets. The most popular datasets are small compared to datasets in other classification tasks. The use of attention maps can compensate for the limited data by enhancing the information in each input such as to improve the learning quality. The most successful research have been utilizing different types of Convolutional Neural Networks (CNNs) for both feature extraction and classification as they outperform traditional methods. For the same reason, also in this research CNNs are used, but with the addition of facial features extracted as an attention mechanism. The model consists of two independent networks trained on separate datasets; A face classification network for creating the attention maps, and an emotion recognition network for classification of the emotions. In this research the effectiveness of these attention maps will be investigated.

1. Introduction

There have been many good results on datasets made in controlled environments. However, there has been a transition towards “facial emotion recognition in the wild” which is a much more challenging task. The natural conditions make the task a lot more challenging as it introduces a lot of variables such as variations in head pose, illumination, different angles, and maybe some parts of the face are covered or not in the picture at all. In addition, the emotions in pictures taken in controlled environments usually have a stronger intensity compared to natural ones as the person getting photographed consciously tries to show the emotion. Natural emotions are often expressed in a more subtle way which makes them harder to recognize. Recent work ([9], [10], [8], [11], [7]) have mostly focused on designing deep network architectures in order to achieve state-of-the-art performance. In this research the focus is on improving the accuracy by using attention maps created from facial features. Different CNN architectures such as VGG[18], ResNet[20] and DenseNet[21] have been experimented with. The network trained for face classification uses the same architecture as the emotion classification network. As datasets made for face classification were not found, a new dataset was created using various existing datasets. The dataset used for facial emotion recognition is FER2013[1] which was a widely used dataset for emotion recognition. It has a size of 35,887 images which is just half of MNIST[19] (dataset for classifying handwritten digits) which is used for a much less complex task than emotion recognition. FER2013’s data size might not be sufficient, at least considering there is some mislabeling as well. By the use of attention maps the network might be able to learn more from each image as there is more focus on the important areas. This research will evaluate the effectiveness of these attention maps, and whether or not they contribute additional valuable information to the task of emotion recognition.

2. Related work

The use of facial features as an attention mechanic was inspired from CAM [2], Grad-Cam [3] and attention branch network [4]. The attention branch network was an extension of CAM, and had the added feature of being able to use attention maps in feed forward propagation, without having the need to backpropagate first. CAM was also not able to generate attention maps during the training process as it was dependent on the feature map and weights after training, whereas the attention branch network used the internal features as attention. These were used to enhance the feature maps inside the network. In [6] they collected features from early, mid, end layers and used them as attention in a weighted combination. A similar approach is attempted in this research, with the main difference being that they used features from a network to enhance the same network, while in this research features are extracted from a separate network. In [5] they created a new approach for classifying global emotion of images with multiple persons. They calculated the global emotion probability from the individual emotions from each person in the image. The new approach was to score and weight the individual emotions. This differ

[1] Osaka Prefecture University
[a) nc101042@edu.osakafu-u.ac.jp
[b) yoshioka@cs.osakafu-u.ac.jp
[c) inoue@cs.osakafu-u.ac.jp

[*1] Features extracted from a network trained on face classification

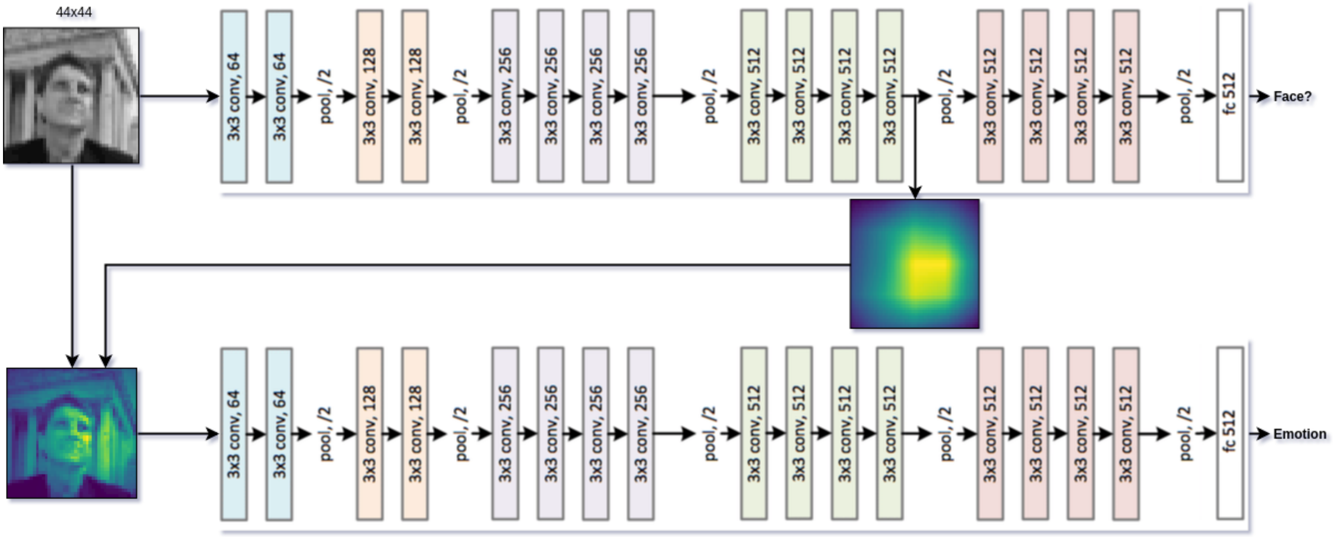

Fig. 1 The architecture of Model 1 and Model 2.

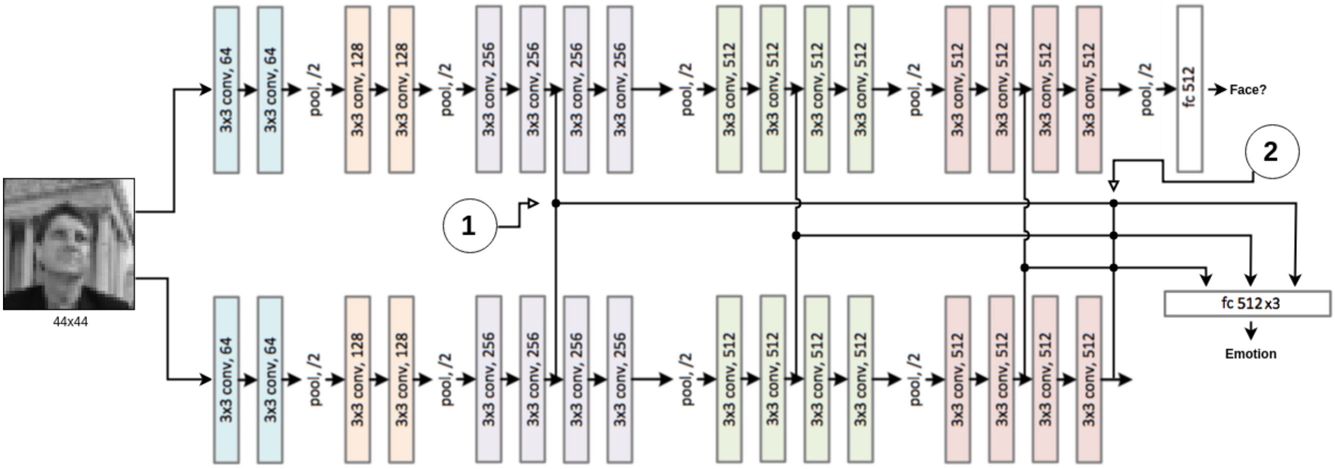

Fig. 2 The architecture of Model 3 (2nd variation). (1): Features from both networks combined. (2) Weighted features applied to the last layer output.

from this research as the goal is to classify the emotion from a single person. However, similar scoring- and weighting functions are utilized. In [7] they achieved state-of-the-art face emotion recognition accuracy by creating a deep CNN ensemble. This differ from this research as the focus is on attention maps, and the architecture of the CNN is not of high importance.

3. Approach

To compensate for the limited sized dataset, the proposal is to utilize facial features as an attention mechanism as to provide the network with images that are easier to classify. This is achieved by highlighting the most important areas of the input image by the use of attention maps. Three different approaches have been attempted. These approaches will be referred to as Model 1, 2 and 3. In all of them, the input image is first sent through the face classification network in order to get the features. The features are then, together with the original input image, sent to one of the three models.

3.1 Model 1

Model 1 [Fig.1] is the simplest of the three models. In this model, the facial features are first applied to the input image to create attention maps, that further are processed and classified by the model. The facial features are represented by a (512, 5, 5) feature tensor where 512 is the depth (number of features), and 5, 5 are the width and height respectively. Some different variations are used; using all the 512 features, averaging all the features into a (1, 5, 5) tensor, and sending the features through a convolutional layer with a (1, 5, 5) tensor as the output. The resulting tensor is then interpolated to the same dimensions as the input image, and then either concatenated with the input image or the dot-product is taken. This is then classified by the emotion recognition network. When taking the dot-product between the facial features and the input image, two variations are used; (Eq. 1): simply taking the dot product

$$inp_1 = ffeats \cdot img \tag{1}$$

$$inp_2 = (1 + ffeats) \cdot img \tag{2}$$

where inp_1 and inp_2 are the input tensors (the resulting attention map shown in Fig.1), *ffeats* is the facial features and *img* is the input image. (Eq. 2): adding weight to the input image (as done in [4]). In the first variation, areas of the attention maps will get a zero value if not contained in the facial features (See Fig3. This can be a good approach if it is guaranteed that the all areas of the face are contained in at least one feature, and that the features contain big enough areas of the face such that it is possible to learn the relation between attention maps and the emotions. The second variation prevents zero values by adding the input image as a base value, and thus guarantees that all areas of the image will have a value.

3.2 Model 2

Model 2 [Fig.1] is the same as Model 1 with the addition of the facial features being sent through a softmax layer and weighted, before being concatenated with the input image or the dot-product between them being taken. In Model 1 all the features have equal weight by default, but some features might have a higher importance than the others. Consequently, in Model 2 the facial features are all weighted, so that the more important features get to contribute more to the attention map, resulting in a presumably more accurate attention map.

3.3 Model 3

Model 3 [Fig.2] uses both the facial features, as well as features from the emotion network as attention. The facial features in this model are extracted from three different layers in the face classification network. By extracting features from three different stages in the network, a bigger variety of attention maps are created. The early layers capture smaller, local features, while the mid and later layers capture bigger, more global features. Some different variations are used; using only the facial features, concatenating or taking the dot-product between the facial features and the features from the emotion recognition network, and applying the facial features to the input image as attention similar to Model 1, before extracting and applying the emotion features to create the final attention map. In the second variation a combination of features from both networks are experimented with. The motivation is that attentions maps that are based the combination of facial- and emotion features might highlight more important areas than either of them would individually. The third variation is similar to Model 1, but has attentions maps from different layers instead of just one, and uses internal features as additional attention. This is to utilize the internal features of the emotion recognition network as attention to further improve the attention maps gotten from the face classification network. As in Model 2, all the features are weighted.

4. Experiments

4.1 Dataset

All the models are evaluated against FER2013. The 35,887 images are divided into a training set of 28,798 images, a validation set of 3,589 images and a finally a test set of 3,589 images. All the images have 48 x 48 pixels. There are 7 different classifications: anger, disgust, fear, happiness, neutral, sadness, and surprise. The face classification network uses a custom dataset created from existing datasets. For the positives, different face detection/recognition datasets ([13], [17], [15], [16]) were used, and for negatives, multiple object and scenery classification datasets ([12], [13], [14]) were used.

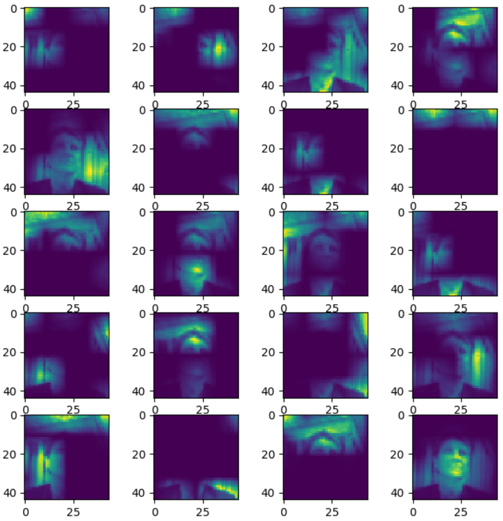

Fig. 3 Attention maps given by the dot-product of the facial features and the input image (Eq. 1)

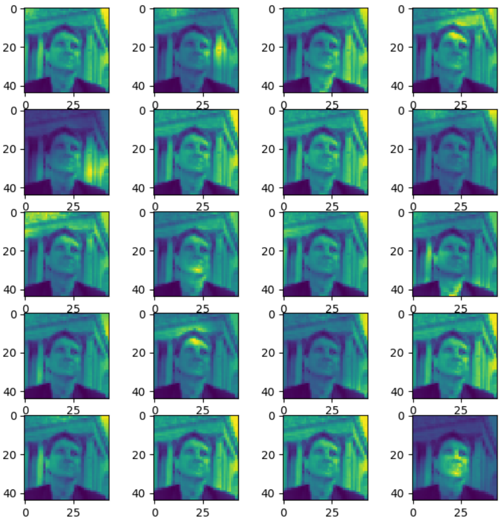

Fig. 4 Attention maps given by the weighted dot-product of the facial features and the input image (Eq. 2)

4.2 Implementation

The images are first randomly cropped into 44 x 44 sized

Table 1 Results on FER2013.

	Variation	Accuracy
VGG19		72.61%
Model 1	All features	72.9%
	Averaging	73.1%
Model 2	All features	73.08%
	Averaging	72.69%
Model 3	Facial features	73.35%
	Emotion features	73.08%
	Combination	72.97%
	Separately	73.59%

images, then randomly horizontally flipped. For model 1 and 2 the batch size is set to 128, but for model 3 it is reduced to 16 because of memory limitations. The total number of iterations range from 150-300 epochs depending on the model and variation. The initial learning rate is set to 0.01 and is further divided by 10 after 50%, 75% and 90% of the total number of epochs. All the models use stochastic gradient descent (SGD) as the optimizer with 0.9 as momentum. Weight decay is set to 5e-4. The criterion used is cross entropy loss. All models are based on VGG19[18], and all models are trained from scratch.

4.3 Results

Table 4.3 presents the results. In the table *All features* refers to using all features extracted from the network, *Averaging* refers to using a single averaged feature (as explained in Model 1). For Model 1 and 2, *features* refers to features from the face classification network. For Model 3, *Facial features* refers to features from the face classification network, *Emotion features* refers to features from the emotion recognition network, *Combinations* refers to the combination of both (as explained in the second variation of Model 3). Finally *Separately* refers to first applying the facial features to the input image, then using the emotion features (as explained in the third variation of Model 3). For each model, only the variations with best results will be presented. For Model 1, this is the variation with weighted dot-product (Eq. 2). For Model 2 using concatenation with all features, and using weighted dot-product with averaging, gave the best results. For Model 3, using simple dot-product (Eq. 1) gave the best results. All models achieved better accuracy compared to the standard VGG19 architecture. Model 2 did not achieve better accuracy than Model 1, even with the weighted features. This might imply that the weights given by the facial features are not relevant in the task of emotion recognition. Model 3 outperforms both other models. This is expected as it not only utilizes more diverse attention maps, but also features from both networks.

5. Conclusion

The presented models did not improve the accuracy by a significant amount, at best 73.59% which is just 0.98% above the baseline. State-of-the-art performance is not achieved, as the current best achieves 75.2% which is 1.61% better. This might be because the facial features used as attention maps do not add any extra valuable information or that there are better ways to utilize them. For future work other types of combinations can be attempted and the models can be evaluated on other datasets.

References

[1] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., "Challenges in representation learning: A report on three machine learning contests, in International Conference on Neural Information Processing.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," In Proc. of CVPR, 2016.

[3] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," arXiv:1710.11063, 2017.

[4] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. "Attention branch network: Learning of attention mechanism for visual explanation," arXiv:1812.10025, 2018.

[5] A. Gupta, D. Agrawal, H. Chauhan, J. Dolz, and M. Pedersoli, "An Attention Model for group-level emotion recognition," EmotiW 2018 group-level emotion recognition challenge, 2018.

[6] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention." arXiv:1804.02391, 2018.

[7] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art, arXiv:1612.02903, 2016.

[8] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," arXiv:1804.08348, 2018.

[9] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," arXiv:1705.07871, 2017.

[10] J. Zeng, S. Shan, and X. Chen, "Facial Expression Recognition with Inconsistently Annotated Datasets," In Proc. of ECCV, pp. 222-237, 2018.

[11] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," arXiv:1608.01041, 2016.

[12] L. Fei-Fei, R. Fergus and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," In Proc. of CVPRW, 2004.

[13] G. Griffin, AD. Holub, and P. Perona, "The Caltech 256," Caltech Technical Report, 2006.

[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. "Learning Deep Features for Scene Recognition using Places Database. In Proc. of NIPS, 2014.

[15] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," In Proc. of ICAFGR, 46-53, 2000.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression,". In Proc. of CVPRW, 94-101, 2010.

[17] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising face across pose and age," In Proc. of ICAFGR, 2018.

[18] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556v6, 2014.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition." In Proc. of IEEE, 2278-2324, 1998.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv:1512.03385v1, 2015.

[21] G. Huang, Z. Liu, and, K. Q. Weinberger, "Densely Connected Convolutional Networks," arXiv:1608.06993, 2016.