Facial Expression Recognition

Subjects: Engineering, Electrical & Electronic

Contributor: Seokhoon Yoon

Facial expression recognition (FER) is a challenging problem in the fields of pattern recognition and computer vision. The recent success of convolutional neural networks (CNNs) in object detection and object segmentation tasks has shown promise in building an automatic deep CNN-based FER model. However, in real-world scenarios, performance degrades dramatically owing to the great diversity of factors unrelated to facial expressions, and due to a lack of training data plus an intrinsic imbalance in the existing facial emotion datasets. To tackle these problems, this paper not only applies deep transfer learning techniques but also proposes a novel loss function called weighted-cluster loss, which is used during the fine-tuning phase. Specifically, the weighted-cluster loss function simultaneously improves the intra-class compactness and the inter-class separability by learning a class center for each emotion class. It also takes the imbalance in a facial expression dataset into account by giving each emotion class a weight based on its proportion of the total number of images. In addition, a recent, successful deep CNN architecture, pre-trained in the task of face identification with the VGGFace2 database from the Visual Geometry Group at Oxford University, is employed and fine-tuned using the proposed loss function to recognize eight basic facial emotions from the AffectNet database of facial expression, valence, and arousal computing in the wild. Experiments on an AffectNet real-world facial dataset demonstrate that our method outperforms the baseline CNN models that use either weighted-softmax loss or center loss.

Keywords: facial expression recognition ; deep convolutional neural network ; transfer learning ; auxiliary loss ; weighted loss ; class center

1. Introduction

Facial expressions are undoubtedly a dominant, natural, and effective channel used by people to convey their emotions and intentions during communication. Over the last few decades, automatic facial expression analysis has attracted significant attention and has become one of the most challenging problems in computer vision and artificial intelligence fields. Numerous studies have been conducted on developing reliable automated facial expression recognition (FER) systems for use over a wide range of applications, such as human-computer interaction, social robotics, medical treatment, virtual reality, augmented reality, and games [1][2][3][4]. In this work, we constructed deep convolutional neural network (CNN)-based FER models to recognize eight common facial expressions (happy, sad, surprise, fear, contempt, anger, disgust, and neutral) from the AffectNet database of facial expression, valence, and arousal computing in the wild [5]. This was motivated by an upcoming challenge on AffectNet's website [6].

The FER model is usually composed of three main stages: face detection, feature extraction, and emotion classification. First, the face and its components (e.g., eyes, mouth, nose, and eyebrows) are detected from images or video sequences. Then, features that are the most effective at distinguishing one expression from another are extracted from the face region. Finally, a classifier is constructed, given the extracted feature set for each target facial expression. The literature is rich with handcrafted face detection and feature extraction methods for FER that have achieved satisfactory results in laboratory-controlled settings [7][8][9][10][11]. However, these traditional methods have been reported to be incapable of discriminating a great diversity of unrelated factors in facial expressions (e.g., subtle facial appearances, head poses, illumination intensity, and occlusions) with FER tasks for in-the-wild settings [12][13].

Recently, the success of convolutional neural networks in both computer vision and pattern recognition has promoted a transition in FER from using handcrafted feature-learning methods to using deep learning technologies. A deep learning-based FER system commonly uses a CNN model to extract and learn high-level features directly from input images. Then, an output layer (which usually uses softmax as an activation function) is attached on top of the CNN model to distinguish the emotion to be detected. This allows a faster emotion recognition system with higher accuracy in challenging, real-world environments [14][15][16].

2. Description

First, deep learning-based FER models require a large amount of data for training to acquire suitable values for model parameters. Directly training the FER model on small-scale datasets is prone to overfitting [17], which leads the model to be less generalized and incapable of handling FER tasks in real-world environments. Although a great effort has been made to collect facial expression training datasets, large-scale, annotated, facial expression databases are still limited [13]. Therefore, overfitting caused by a shortage of data remains a challenging issue for most FER systems.

Second, imbalances in the distribution of facial expression samples from real-world FER datasets may degrade the overall performance of the system [18]. Due to the nature of emotions, the number of collected facial images for the major classes (e.g., happiness, sadness, and anger) is much larger than for the minor classes (e.g., contempt, disgust, and fear). In the AffectNet dataset, the happy category comprises about 47% of all the images, whereas the contempt category comprises only 1.2%. FER systems being trained on an imbalanced dataset may perform well on dominant emotions, but they perform poorly on the under-represented ones. Usually, the weighted-softmax loss approach [5] is used to handle this problem by weighting the loss term for each emotion class based on its relative proportion in the training set. However, this weighted-loss approach is based on the softmax loss function, which is reported to simply force features of different classes to remain apart without paying attention to intra-class compactness. One effective strategy to address the problem of softmax loss is to use an auxiliary loss to train the neural network. For instance, triplet loss [19] and center loss [20] introduce multi-loss learning to enhance the discriminating ability of CNN models. Although these loss functions do boost the discriminative ability of the conventional softmax loss, they usually come with limitations. Triplet loss requires a comprehensive process of choosing image pairs or triplet samples, which is impractical and extremely time-consuming owing to the huge number of pairs and samples in the training phase. Center loss does not consider inter-class similarity, which may lead to poor performance by the FER system. In addition, none of these auxiliary loss functions is able to address data-imbalance problems.

To address the first problem (the shortage of data), in this work, the transfer learning technique is applied to build the FER system. Transfer learning is a machine learning technique by which a model trained on one task is repurposed for another related task. It not only helps to handle the shortage of data but also speeds up training and improves the performance of the prediction model. In this paper, we take a transfer learning approach by employing two recent CNN architectures in two-stage, supervised pre-training and fine-tuning. Specifically, a squeeze-and-excitation network (SENet) model [21] which is pre-trained for the face identification task on the VGGFace2 [22] database from the Visual Geometry Group at Oxford University, was fine-tuned on the AffectNet dataset [5] to recognize the above-mentioned eight common facial expressions.

Tackling the second problem of imbalanced data distribution in existing FER datasets, we propose a new loss function called the weighted-cluster loss, which integrates the advantages of the weighted-softmax approach and the auxiliary loss approach. First, weighted-cluster loss learns a class center for each emotion, which simultaneously reduces the intra-class variations and increases the inter-class differences. Next, the proposed loss gives weights to each emotion class's loss terms based on their relative proportion of the total number of samples in the training dataset. In other words, weighted-cluster loss penalizes networks more for misclassifying samples from minor classes while penalizing those networks less for misclassifying examples from major classes. Furthermore, the training process is simple because weighted-cluster loss does not require preselected sample pairs or triples.

References

1. Hickson, S.; Dufour, N.; Sud, A.; Kwatra, V.; Essa, I. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1626–1635.
2. Chen, C.H.; Lee, I.J.; Lin, L.Y. Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. Res. Dev. Disabilities 2015, 36, 396–403.
3. Dornaika, F.; Raducanu, B. Efficient facial expression recognition for human robot interaction. In Proceedings of the International Work-Conference on Artificial Neural Networks, San Sebastián, Spain, 20–22 June 2007; pp. 700–708.
4. Zhan, C.; Li, W.; Ogunbona, P.; Safaei, F. A real-time facial expression recognition system for online games. Int. J. Comput. Games Technol. 2008, 2008, 10.
5. Mollahosseini, A.; Hasani, B.; Mahoor, M. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Trans. Affective Comput. 2017.

6. AffectNet Database. Available online: http://mohammadmahoor.com/affectnet/ (accessed on 15 August 2019).
7. Tian, Y.l.; Kanade, T.; Cohn, J.F. Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In Proceedings of the Fifth IEEE International Conference on Automatic Face Gesture Recognition, Washington, DC, USA, 21 May 2002; pp. 229–234.
8. Whitehill, J.; Omlin, C.W. Haar features for facs au recognition. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006.
9. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Comput. 2009, 27, 803–816.
10. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 915–928.
11. Dahmane, M.; Meunier, J. Emotion recognition using dynamic grid-based HoG features. In Proceedings of the Face and Gesture 2011, Santa Barbara, CA, USA, 21–25 March 2011.
12. Li, S.; Deng, W. Deep facial expression recognition: A survey. arXiv 2018, arXiv:1804.08348.
13. Ko, B. A brief review of facial emotion recognition based on visual information. Sensors 2018, 18, 401.
14. Ebrahimi Kahou, S.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent neural networks for emotion recognition in video. In Proceedings of the 17th ACM International Conference on Multimodal Interaction, Seattle, DC, USA, 9–13 November 2015; pp. 467–474.
15. Walecki, R.; Rudovic, O.; Pavlovic, V.; Schuller, B.; Pantic, M. Deep structured learning for facial expression intensity estimation. Image Vis. Comput. 2017, 259, 143–154.
16. Kim, D.H.; Baddar, W.J.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affective Comput. 2017, 10, 223–236.
17. Hawkins, D.M. The problem of overfitting. J. Chem. Inf. Comput. Sci. 2004, 44, 1–12.
18. Visa, S.; Ralescu, A. Issues in mining imbalanced data sets-a review paper. In Proceedings of the Sixteen midwest artificial intelligence and cognitive science conference, Dayton, OH, USA, 22 February 2005.
19. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 8–10 June 2015; pp. 815–823.
20. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the European conference on computer vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 499–515.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018; pp. 7132–7141.
22. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74.
