Lian Duan
Nimit Dilsukhbhai Hingrajia
Dhruvil Niraj Shah

School of Computer Science

University of Windsor

Lian Duan	1.) Facial Expression Recognition with Convolutional Neural Networks 2.) Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network
Dhruvil Niraj Shah	3.) Facial Expression Recognition Based on Transfer Learning from Deep Convolutional Networks 4.) Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning
Nimit Dilsukhbhai Hingrajia	5.) TransFER: Learning Relation-aware Facial Expression Representations with Transformers 6.) Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism

Proposal

Problem Statement

Facial expressions are critical for human communication because they assist us in interpreting the intentions of others. Humans use facial expressions and vocal tones to infer emotions such as joy, sadness, and anger from others. According to different surveys, verbal components reflect one-third of human communication, whereas nonverbal components express two-thirds (Mehrabian, 2008; Kaulard et al., 2012). Facial expressions are one of the most significant non-verbal components of interpersonal communication because they carry emotional content. As a result, it’s no wonder that research into facial expression recognition has gained a lot of momentum in recent decades, with applications spanning from perceptual and cognitive sciences to affective computing and computer animations (Kaulard et al., 2012)

Project Scope & Methodology

In this project, we would like to build a real-time prediction model including age, gender, and emotion from a webcam using several deep learning approaches. More specifically, we propose to implement transfer learning as well as Spatial-Transform CNN. The models will be trained from UTKface and FER2013 datasets. Image segmentation and augmentation techniques will also be applied to datasets as the step of preprocessing.

Goals

Finally, this project will provide a solution to detect the emotions and expressions of particular human beings in a precise form. Accurate emotion predictions can also help the automatic systems to provide detailed descriptions about what will be the next action of human beings or how a particular human being is feeling. This project can also help the government authorities to recognize the criminal's face and their emotions to understand them precisely.

Literature Review

1. Facial Expression Recognition with Convolutional Neural Networks

Overview

Singh and Nasoz (2020) described the recent implementation of Facial Expression Recognition using Convolutional Neural Networks (CNN). They demonstrate the process of classification by plugging static images into CNN architecture. They also provided some techniques to improve the accuracy of CNN prediction using pre-processing and feature extraction methods. After tunning and training the CNN model, they received a test accuracy of 61.7% on FER2013 dataset.

Methodology

Singh and Nasoz (2020) illustrated how pre-processing can be used to improve the performance of Facial Expression Recognition. Data pre-processing includes detection, correction of illumination, data augmentation, etc. One of the examples Singh and Nasoz (2020) gave is to use Haar Cascade classifier to identify human faces. Since the illumination of each image when captured varies, utilizing the illumination correction technique can somehow improve the accuracy of feature extraction in order to improve the overall classification performance. As for feature extraction, it can help researchers to reduce the size of data to accelerate model training. Singh and Nasoz (2020) applied dlib facial landmark detector that is trained on iBUG 300-W dataset for feature extraction, and they extracted the eight most important parts of the face. In the section on CNN implementation, they displayed the optimized architecture of their CNN model, which contains 6 convolutional layers, 3 max-pooling layers, 2 drop-out layers, and 1 flattened layer with 2 dense layers.

Adaptation

Even though the CNN model proposed by Singh and Nasoz (2020) has achieved a moderate accuracy score, some enhancements can be added to this CNN model in the future to receive a higher accuracy. For example, they pointed out that pre-processing and feature extraction techniques can be included in the later implementation of the pipeline to achieve better accuracy than the current CNN model. They also stated that overfitting is an outstanding issue they need to fix by using data augmentation.

2. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network

Overview

Minaee et al. (2021) proposed a new deep learning approach for facial expression recognition by using an Attentional Convolutional Network. Compared with other current deep learning approaches, such as CNN, they stated that the attentional approach is able to make a significant improvement in face recognition by focusing on important parts of the face. They also mentioned that they utilized visualization techniques to highlight the salient regions of face images, which are the most pivotal parts for distinguishing different facial expressions.

Methodology

Minaee et al. (2021) revealed that in order to improve the performance of a deep neural network, adding more layers/neurons, facilitating gradient in network, and applying better regularizations are the typical methods. However, due to the limited number of classes for facial expression recognition, they introduced a convolutional network with less than 10 layers as well as attention to present better results than state-of-the-art approaches. As to the architecture of model, it consists of two components: localization network and convolutional network. For the localization network, it utilizes a spatial transformer that contains two convolution layers and two fully connected layers. The output of transformation is a sampling grid, which is used to be combined with the architecture of the convolutional network.

Adaptation

Minaee et al. (2021) concluded that the framework they proposed improves accuracy, even in some datasets, is higher than the state-of-the-art models. They suggested that with the help of attention to special regions, the neural network can be built with less than 10 layers to achieve a high accuracy rate.

3. Facial Expression Recognition Based on Transfer Learning from Deep Convolutional Networks

Overview

In this paper, Xu et al. (2015) have proposed an efficient facial expression model dependent on transfer features from ConvNets. These deep ConvNets were trained on the MSRACFW database using 1580-class face identification. Finally, Xu et al. (2015) transferred high-level features from a trained deep model to recognize expression. This model was trained and tested at the large scope using their own build database of 7 basic emotion states and 2062 imbalanced samples depending on 4 facial expression databases. Overall, Xu et al. (2015) achieved 50.65% recognition of Gabor features with 7-class SVM2 and 78.84% recognition of the self-built facial expression database with 7-class SVM.

In conclusion, Xu et al. (2015) tested the model on occluded conditions as well and showed the ability of the model classification in small occlusion cases, Xu et al. (2015) also modified the model for improving the facial recognition on a self-build database and achieved 81.50 average accuracy.

Methodology

Xu et al. (2015) proposed four convolutional layers with max pooling for features extraction on every layer, the fully connected layer will have high-level features layers with a SoftMax output layer for class identification through prediction. Inputs of 39*39 gray facial patches extracted from the MSRA-CFW database using the viola-jones algorithm. On every step number of features was decreased and finally, the high-level features layer is fixed to 120 features for face information extraction. The last SoftMax output layer totally connected to the high-level features predicted for 1580 identity classes. Xu et al. (2015) distributed the weights locally within every layer of our deep convNets to learn regional features. Now the high-level features are fully converted to the 4 cnovNets layers after ReLU. The model of Xu et al. (2015) was able to predict the 1580-way probability distribution of input facial path to identify the face. This whole ConvNets was developed using backpropagation and stochastic gradient descent method by Xu et al. (2015).

On adequate training of deep ConvNets Xu et al. (2015), adopted multiclass SVM and 120 high-level dimensional features which were transferred from trained deep ConvNets to predict and classify 7 emotion states. Based on the testing and experiments Xu et al. (2015) stated that their kernel function selects RBF (Radial Basis Function), and their model was able to recognize facial emotion after the 7-class SVM model learned the model that was developed by Xu et al. (2015).

Xu et al. (2015) define a problem related to occlusion; they provided a deep robust model whose features were able to overcome occlusion, but they failed when there was a slight increase in occlusion. To solve this problem and to increase the robustness of occlusion, Xu et al. (2015) merged the high-level features of two trained deep ConvNets models of the same structure and created a new modified model. This new modified model used the same 7-class SVM classifier with 240-dimensional high-level merged features. Two deep ConvNets and deep ConvNets are the same in structure for face detection and performing the task at 1580-class face identification. These two models were trained on MSRACFW and MSRA-CFW databases with additive occluded samples whereas the modified model was developed on a self-built facial recognition database.

Adaptation

Xu et al. (2015) and other authors concluded that this model can be explored more in detail to adapt the other recognition like different facial poses and real-time recognition.

4. Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning

Overview

In this paper, Ng et al. (2015) classified the emotions in static images of the human extracted from the movies using a transfer learning approach for deep CNN architecture. Using a pre-trained model on the ImageNet dataset Ng et al. (2015) performed supervised fine-tuning on the networks in 2 stages. Firstly, on dataset relevant to a facial expression which is followed by the given dataset.

Ng et al. (2015) stated that their experimental results show that cascading fie-tuning approach has achieved better results as compared to single-stage fine-tuning with the combined datasets with an overall accuracy of 48.5% and 55.6% in validation and testing dataset respectively.

Methodology

Ng et al. (2015), used OpenCV's Viola & Jones face detector initially for faces detection for the EmotiW dataset, with the help of intraface library Ng et al. (2015) discarded the false positive faces. Ng et al. (2015) choose the ILSVRC-2012 (AlexNet) and CNN-M-2048 from (VGG-CNN-M-2048) transfer learning CNN based on two different architectures. Here Ng et al. (2015) used the FER-2013 as an auxiliary dataset instead of the EmotiW dataset because of its larger size and fine-tuned the data of ILSVRC-2012 (AlexNet) to find the results of the model.

Ng et al. (2015) divided the dataset into 3 parts, FER28, FER32, and FER32 + EmotiW. Here FER28 consists of 28709 images for training and 3589 images for validating the fine-tuned models whereas FER32 consists of (28709+3589=32298) total images for training and 3589 images to validate the fine-tuned models. On the other hand, FER32 + EmotiW consist of a combination of both the FER32 and EmotiW training dataset for training and only the EmotiW validation dataset for validating the fine-tuned models. Ng et al. (2015) stated that they used different schemes for fine-tuning the base pre-trained CNN model using different datasets in combination with EmotiW training data, this ultimately provided fine-tuning directly to the ILSVRC-2012 with stage tuning.

This overall process of fine-tuning was first done using the FER-2013 dataset and then the fine-tuning process was done using a targeted dataset of EmotiW. These CNNs were trained using stochastic gradient descent with hyperparameters with a learning rate of 0.001. Ng et al. (2015) didn't change the weights but did some delays which helped them to overcome overfitting problems to achieve the desired accuracy by dropping the learning rate by a factor of 10 every 10 epochs.

Adaptation

Ng et al. (2015) state that transfer Learning methods which are the deep CNN trained on sufficiently large auxiliary face expression datasets can alone easily be used to obtain very effective results than the baseline without using any data from the EmotiW datasets. This can be further extended for model improvement using the EmotiW dataset with a sufficiently larger face dataset such as FER-2013. This can be either done by adding it to the auxiliary dataset or by adding a new round of fine-tuning to achieve the marginal owing to its small size.

This provides a suggestion to Ng et al. (2015) that if deep CNN is exploiting facial expression recognition to achieve significant accuracy and gains than a big dataset is required, and it is very crucial. Lastly, Ng et al. (2015) also faced the problem related to increasing the model accuracy due to inherent difficulty that they faced while labeling the faces to eliminate the nuanced emotions. This problem made Ng et al. (2015) to develop the model with transferring samples from one class to another instead of a binary accuracy measure for cost-sensitive performance.

5. TransFER: Learning Relation-aware Facial Expression Representations with Transformers

Overview

Xue et al.(2021) propose the Trans-FER model which can learn rich relation-aware local representations. It mainly comprises three components: MultiAttention Dropping (MAD), ViT-FER, and Multi-head SelfAttention Dropping (MSAD). MAD is used to overcome the problem of overfitting the model. While ViT-FER is used for building relations between the different local patches. To address the problem of multiple self attention that may extract similar relations the MSAD is proposed to randomly drop one self-attention module, increasing the diversity in relations.

Methodology

Datasets used are RAF-DB and FERPlus. All images are aligned and resized to 112 × 112 pixels. Pre-trained on Ms-Celeb-1M, the IR-50 is used as the stem CNN where only the first three stages in IR-50 are used. Pre-trained on ImageNet1, ViT with eight self-attention heads and a stack of M = 8 identical encoder layers are adopted as the Transformer Encoder.

Due to the issue of class imbalance, unsampling of the training dataset is used to balance the class distribution. The drop rates of MAD in local CNNs (p1) and MSAD (p2) are set to 0.6 and 0.3 for RAF-DB and FERPlus, and 0.2 and 0.6 for AffectNet, respectively. To minimize the cross-entropy loss Xue et al.(2021) trained TransFER with the SGD optimizer. At test time, Xue et al.(2021) only resize the original image to 112 × 112 pixels and feed it to the model directly. For RAF-DB and FERPlus, Xue et al.(2021) train 40 epochs with an initial learning rate of 1e-3 decayed by a factor of 10 at the 15 and 30 epochs. For AffectNet, due to its large number of sample data, Xue et al.(2021) train 20K iterations with an initial learning rate of 3e-4 decayed by a factor of 10 at 9.6K and 19.2K iterations. Xue et al.(2021) used two NVIDIA V100 GPUs with 32GB RAM to train their model.

Adaptation

Xue et al.(2021) have proposed a new architecture based on the Transformer for the FER task, called TransFER, which can learn rich, diverse relation-aware local representations. Firstly, a Multi-Attention Dropping (MAD) has been proposed to guide local CNNs to generate diverse local patches, making models robust to pose variations or occlusions. Secondly, the ViT-FER is applied to build rich connections upon multiple local patches where important facial parts are assigned with higher weights and useless ones are assigned smaller weights. Thirdly, the MSAD has been proposed to explore more rich relations among diverse facial parts. To the best of Xue et al.(2021) knowledge, this is the first work to utilize the Transformers for the FER task. Extensive experiments on three public FER datasets demonstrated that Xue et al.(2021) approach outperforms the state-of-the-art methods.

6. Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism

Overview

FER in practical use poses unconstrained problems like detecting expression on partially occluded faces. In this paper, Li et al. (2018) propose a convolutional neural network (CNN) with attention mechanism (ACNN) that can perceive the occlusion regions of the face and focus on the most discriminative un-occluded regions. ACNN combines multiple regions from facial regions of interest (ROIs). Each representation is weighed via a proposed gate unit that computes an adaptive weight from the region itself according to the unobstructedness and importance. Considering different RoIs, Li et al. (2018) introduce two versions of ACNN: patch-based ACNN (pACNN) and global–local-based ACNN (gACNN). pACNN only pays attention to local facial patches. gACNN integrates local representations at patch-level with global representation at imagelevel.

Methodology

Li et al. (2018) conducted face detection, alignment, cropping for each image in FED-RO, RAF-DB, AffectNet dataset. Li et al. (2018) extracted features from a VGG16 network (pretrained on RAF-DB and AffectNet datasets) for each facial image to get a 4096 dimensional feature. Li et al. (2018) calculated cosine similarity score for each image pair (i.e., one image comes from FED-RO, the other image comes from RAF-DB or AffectNet dataset). If the similarity score was larger than a predefined threshold, the image pair would be checked by humans. If the two images were the same, Li et al. (2018) would drop the corresponding image in FED-RO. Li et al. (2018) set the threshold as 0.01. Finally, FED-RO contains 400 images in total. The images are categorized into seven basic expressions (i.e., neutral, anger, disgust, fear, happy, sad, surprise). Occlusion patterns in FED-RO are diverse in color, shape, position and occlusion ratio.

Adaptation

The Gate Unit in ACNN enables the model to shift attention from the occluded patches to other unobstructed as well as distinctive facial regions. Considering that facial expression is distinguished in specific facial regions, Li et al. (2018) designed a patch based pACNN that incorporates region decomposition to find typical facial parts that are related to expression. Li et al. (2018) also developed an efficient gACNN to supplement global facial information for FER in the presence of occlusions. Experiments under intra and cross dataset evaluation protocols demonstrated ACNNs outperform other state-of-the-art methods. Ablation analyses show ACNNs are capable of shifting attention from occluded patches to other related ones.

Summary

In this proposal and literature review, we summarized six papers from different domains of facial expression recognition using different techniques. From there, we can easily find out the mainstream and state-of-the-art approaches to solving facial expression recognition problems.It will also be helpful for us to decide the approaches we can use in order to receive a remarkable result.

References

Mehrabian, A. (2008). Communication without words. Communication theory, 6, 193-200.

Kaulard, K., Cunningham, D., Bulthoff, H., Wallraven, C. (2012). ¨ The MPI Facial Expression Database — A Validated Database of Emotional and Conversational Facial Expressions. Plos ONE, 7(3), e32321. https://doi.org/10.1371/journal.pone.0032321

Singh, S., & Nasoz, F. (2020). Facial expression recognition with convolutional neural networks. 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). https://doi.org/10.1109/ccwc47524.2020.9031283

Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046. https://doi.org/10.3390/s21093046

Xu, M., Cheng, W., Zhao, Q., Ma, L., & Xu, F. (2015). Facial expression recognition based on transfer learning from deep convolutional networks. 2015 11th International Conference on Natural Computation (ICNC). https://doi.org/10.1109/icnc.2015.7378076

Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015, November). Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 443-449).

Xue, F., Wang, Q., & Guo, G. (2021). Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3601-3610).

Li, Y., Zeng, J., Shan, S., & Chen, X. (2018). Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28(5), 2439-2450.