Facial Emotion Recognition Using Deep Learning

Csaba Bokányi (FUWQK1), Tamás Mészégető (H6LGMQ)

***Abstract*—**

Impact Statement -

***Index Terms*—Deep Learning, FER, Machine Vision**

I. Introduction

DEEP learning has conquered numerous areas of data processing, and is playing a key role in the image processing field. One of the main reasons, among others, is the high level of abstraction used by humans when understanding and analyzing scenes, which is difficult to model with hard-coding approaches. That is why we choose an image processing task for our project, which, despite being a traditional classification task, is challenging for humans and machines alike. Faces representing any kind of sentiment have many common features, which may be the reason for the relatively low average human accuracy of 65%. On the other hand, this performance leaves room for algorithmic solutions not only to replace human activity, but also to surpass it.

II. Literature review

A. FER overview

The survey from (Li and Deng 2020) summarizes the latest developments in the field: with the appearance of sufficiently large and diverse datasets, such as the FER2013[ ] used in this project, it became possible to train deep neural networks for emotion recognition tasks, surpassing the accuracy of earlier solutions. While dynamic approaches consider multiple subsequent images of the subject thus may utilize this additional information, static solutions solve the problem using a single frame. The possible additional variation in the data and such the needed greater data quantity and model complexity are considerable challenges in the case of the dynamic approach, and lead us to the decision to deal with the static case.

In the intensively researched field of static FER (Facial Emotion Recognition) we orientated with the help of surveys (Li and Deng 2020; Pramerdorfer and Kampel 2016. 12. 09) and also reviewed standalone studies that consider the FER2013 dataset **(Alizadeh and Fazel 2017. 04. 22; Agrawal and Mittal 2020; Boughrara et al. 2016) and Raghuvanshi** and several other works to capture the diversity of approaches in the field.

The referenced surveys summarize the mainstream workflow of a FER solution. The pre-processing approaches mainly try to compensate for the lack of sufficiently large datasets by either eliminating certain unwanted variations or by generating additional training samples. Face alignment using feature detectors or convolutional neural networks, and illumination and pose normalization are among the most widely used approaches to eliminate certain variances. Data augmentation efforts on the other hand, strive to solve the same problem by generating more samples. In this case the usual approaches used in many image classification solutions are applicable, such as the addition of certain types of noise, horizontal flipping and random cropping, modifications to the saturation, contrast or other image parameters depending on the representation and other image transforms which preserve the semantical contents of the image.

In the terms of network design, a great diversity is observable. The main approaches used in image processing tasks, such as convolutional neural networks (CNN), Deep Belief Networks (DBN), Deep Autoencoders, Recurrent Neural Networks (RNN) and General Adversarial Networks (GAN) have all been used to solve the emotion recognition task. (Pramerdorfer and Kampel 2016. 12. 09) mentions, that many convolutional networks that were developed for such purposes are significantly shallower than in related fields, still realizing near state-of-the art results. However, common deep convolutional architectures, such as VGG, Inception or ResNet, among others, have been utilized successfully for the emotion recognition tasks, thanks to the evolution of databases and pre-processing solutions. The authors also mention that deep networks provide the possibility of further improvements over earlier, shallower solutions.

Although going deeper is one of the main developments in the field, some studies show that models containing almost 2 magnitudes fewer parameters can also deliver good results (Agrawal and Mittal 2020). Such solutions may prove robust as fewer parameters limit the dangers of overfitting and thus force a higher level of abstraction.

(Fernandez et al. 2019. 06. 16. - 2019. 06. 17) build a network in their recent work with an integrated attention net. This approach is another way to eliminate variance by focusing only on the face, and allows accuracy levels over 80% on different FER datasets.

Despite various approaches to create additional feature vectors as inputs, the work of (Alizadeh and Fazel 2017. 04. 22) shows that well designed networks are capable of learning such features and such efforts are not necessarily needed.

In the terms of training approaches, diversity prevails: there are numerous end-to-end design and training solutions, for example (Agrawal and Mittal 2020; Zhang et al. 2019). Still, to counter the data quantity and quality issues, human face related pretrained networks (such as face detectors or FER models created for a wider variety of emotions (James Li et al 2019.)) are often utilized as a base for FER solutions. Alternatively, pre-trained networks are not only complemented, but the whole network is trained, and such the pre-training functions as an initialization approach. (Boughrara et al. 2016) uses an on-the-fly hyperparameter optimization: while training, if certain predefined accuracy levels are not reached, the number of neurons is changed (increased) in the hidden layers.

B. Neural Network ensembles

Beyond pre-processing, architecture design, initialization, hyperparameter optimization and training strategies, there remains on important step to improve the overall performance of FER solutions: the interpretation of the output of the network.

Although many FER solutions are not particularly deep, and there are many approaches to understand the inner functioning of artificial neural networks, deep learning is still a bit of a black box. There is no guaranteed reaction to any new data, and all the above-mentioned steps (pre-process, design, initialization, optimization) are going to influence this reaction. This sometimes results in huge output-variance: small changes to the input result in huge changes in the output, potentially resulting in false classification.

Network ensembles counter this sensitivity and uncertainty by utilizing multiple networks and a decision-making step based on the multiple outputs. It is important to note, that there are also feature-level ensembles, but in FER they are more widely used in the decision-making step.

To create an ensemble, one needs to gather different models: different architectures, data and data augmentation techniques, initialization and optimization strategies all yield different solutions. Provided that all models yield the output in the same format, there are multiple decision-making strategies: majority voting and averaging are simple solutions, that assign the same weight to all networks. However, one may want to consider the confidence in each solution. Many approaches are available, such as the composition of weighted averages based on accuracy or loss, or the exponentially-weighted decision function and hierarchical committees described by (Kim et al. 2016).

Using ensembles, significant improvements have been reached on many datasets, such as in the case of (Pramerdorfer and Kampel 2016. 12. 09), which motivates the further integration of newer and better solutions.

III. Concept of the project

When building and training neural networks, the ultimate goal is robustness: the net should be able to handle new data, so that it can be used in real-world applications, along with hard-coded software solutions. This is what motivated our team to experient with small ensembles, using slightly modified models described in the literature […amit hasznalunk..]

IV. Models and training

V. building an ensemble

VI. Conclusion

).

VII.

References

Challenges in representation learning: Facial expression recognition challenge: http://www.kaggle.com/c/challengesin-representation-learning-facial-expression-recognitionchallenge

VIII. Publication bibliography

Agrawal, Abhinav; Mittal, Namita (2020): Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. In *Vis Comput* 36 (2), pp. 405–412. DOI: 10.1007/s00371-019-01630-9.

Alizadeh, Shima; Fazel, Azar (2017. 04. 22): Convolutional Neural Networks for Facial Expression Recognition. Available online at http://arxiv.org/pdf/1704.06756v1.

Boughrara, Hayet; Chtourou, Mohamed; Ben Amar, Chokri; Chen, Liming (2016): Facial expression recognition based on a mlp neural network using constructive training algorithm. In *Multimed Tools Appl* 75 (2), pp. 709–731. DOI: 10.1007/s11042-014-2322-6.

Fernandez, Pedro D. Marrero; Pena, Fidel A. Guerrero; Ren, Tsang Ing; Cunha, Alexandre (2019. 06. 16. - 2019. 06. 17): FERAtt: Facial Expression Recognition With Attention Net. In : 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA, 2019. 06. 16. - 2019. 06. 17: IEEE, pp. 837–846.

Kim, Bo-Kyeong; Roh, Jihyeon; Dong, Suh-Yeon; Lee, Soo-Young (2016): Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. In *J Multimodal User Interfaces* 10 (2), pp. 173–189. DOI: 10.1007/s12193-015-0209-0.

Li, Shan; Deng, Weihong (2020): Deep Facial Expression Recognition: A Survey. In *IEEE Trans. Affective Comput.*, p. 1. DOI: 10.1109/TAFFC.2020.2981446.

Pramerdorfer, Christopher; Kampel, Martin (2016. 12. 09): Facial Expression Recognition using Convolutional Neural Networks: State of the Art. Available online at http://arxiv.org/pdf/1612.02903v1.

Zhang, Hongli; Jolfaei, Alireza; Alazab, Mamoun (2019): A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing. In *IEEE Access* 7, pp. 159081–159089. DOI: 10.1109/ACCESS.2019.2949741.