Deciphering emotions using convolutional neural networks on video data

Elizabeth Tran, Michael B. Mayhew, Alan D. Kaplan

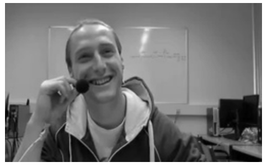
Lawrence Livermore National Laboratory

BACKGROUND

In recent years, there has been a push for automatic facial recognition and to interpret facial expressions using deep learning in computer vision tasks. Here, we are leveraging these methods in order to create continuous emotion predictions.

Emotion Recognition

Emotional recognition has been focused on using categorical models to group emotions into discrete categories, but that method does not capture all expressible human emotions, especially with microexpressions. Microexpressions are brief facial expressions that occur when a person is deliberately or unconsciously concealing an emotion.



Time	FM1	FM2	FM3	FW1	FW2	FW3
2.56	0.00	0.14	-0.01	0.00	0.02	0.00

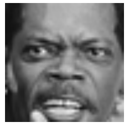
Time	FM1	FM2	FM3	FW1	FW2	FW3
269.64	-0.10	0.19	-0.44	0.00	0.08	0.10

Table 1. Six raters (3 males, 3 females) rated each participant's arousal and valance ranging from [-1, 1]. The above scores show only the participant's valance score.

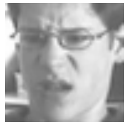
Universal Facial Emotions

There's strong evidence for universal facial expression of seven emotions – anger, disgust, fear, happy, sad, surprise, and neutral.

Anger



Disgust

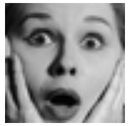


Fear

Happy

Sad

Surprise



Neutral

More than meets the eye

Different facial expressions of emotions have diverse uses in human behavior and cognition and may be associated to multiple emotional categories. This contradicts with continuous models in cognitive science and the multidimensional approaches typically employed in computer vision.

METHODS

Categorical emotion prediction to continuous emotion prediction

Using Kaggle's Facial Expression Recognition Challenge dataset, we trained a convolutional neural network to classify human faces into discrete emotion categories. With four categories (angry, happy, sad, neutral), we were able to achieve a test accuracy of 68.4%. With Kaggle's weights, we were able to continue to experiment with the model using the RECOLA's dataset (a multimodal dataset combining ECG, EDA, audio, video recordings, and annotations ranking multiple emotional characteristic)[5]. Through different architectures and hyperparameters, such as quadrant pooling and fine-tuning, we were able to asses the overall performance of neural networks to recognize emotion from videos.

CNN Architecture

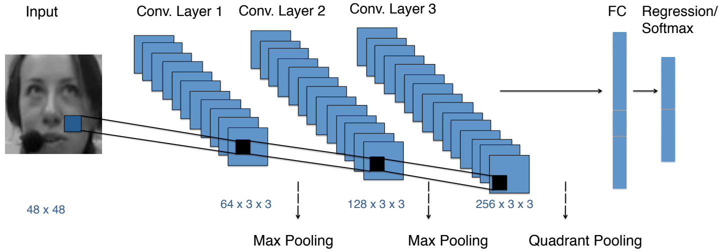


Figure 1. Our network consists for three convolutional layers with 64, 128, 256 filters, and each layer was followed by a 3x3 ReLu (Rectified Linear Unit) activation. Following the first two convolution layers, we inserted a 2x2 max pooling layer and a quadrant pooling after the third layer. The three convolutional layers are then followed by a fully connected layer with 200 hidden units and a linear regression layer that approximates the valence score.

RESULTS

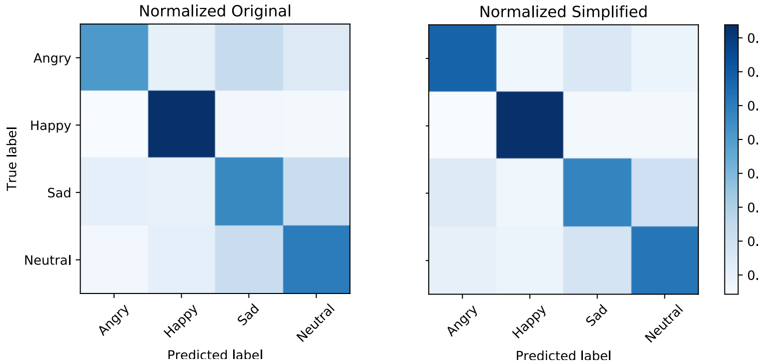


Figure 2. The test accuracy between our Kaggle's inspired by EmoNet model [6] and our model shows that a smaller network can perform emotion recognition with great accuracy. With four classes our simplified model scored 69.69% test accuracy while our original model scored 68.45% test accuracy.

Method	RMSE	Pearson Correlation Coefficient	Concordance Correlation Coefficient
LSTM [1]	0.117	0.358	0.273
LGBP-TOP +LSTM [2]	0.114	0.430	0.354
LGBP-TOP + Deep Bi-Dir. LSTM [3]	0.105	0.501	0.346
Single Frame CNN +D [4]	0.114	0.468	0.326
CNN + Fine Tuning Kaggle's EmoNets – Ours	0.189	Rater 1: .0292 Rater 2: 0.193 Rater 3: 0.325 Rater 4: 0.201 Rater 5: 0.048 Rater 6: 0.152	Rater 1: 0.221 Rater 2: 0.146 Rater 3: 0.246 Rater 4: 0.152 Rater 5: 0.036 Rater 6: 0.115
CNN + Fine Tuning – Ours	0.187	Rater 1: 0.273 Rater 2: 0.169 Rater 3: 0.265 Rater 4: 0.169 Rater 5: 0.154 Rater 6: 0.082	Rater 1: 0.220 Rater 2: 0.136 Rater 3: 0.213 Rater 4: 0.136 Rater 5: 0.124 Rater 6: 0.066

Table 2. Performance comparison between our model and other methods. In our experiment, we focused on predicting the valance score using the video modality of the RECOLA dataset. Unlike the other methods, where they averaged the six rater's scores, we decided to keep each rater separate in order to see the variability.

CONCLUSION & FUTURE WORK

Not a big network, like our original Kaggle model, is required to perform such accuracy for emotion recognition. However, our model finds certain emotions, such as angry, neutral, and sad, similar and tends to mistake them for one another.

Human perception is extremely tuned to small configurations and shape changes. We hope to improve our algorithm to emulate this capacity of precise detection of faces and facial features in order to bridge the gap between categorical and continuously emotion recognition. Having emotionally aware algorithms will improve our understanding of human cognition and behavior.

References

[1] Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., ... & Pantic, M. (2015, October). Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 3-8). ACM.

[2] Chen, S., & Jin, Q. (2015, October). Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 49-56). ACM.

[3] He, L., Jiang, D., Yang, L., Pei, E., Wu, P., & Sahli, H. (2015, October). Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 73-80). ACM.

[4] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, T. S. Huang: "How deep neural networks can improve emotion recognition on video data", in arXiv preprint arXiv:1602.07377, February 2016.

[5] Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on (pp. 1-8). IEEE.

[6] Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., ... & Ferrari, R. C. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, 10(2), 99-111.

LLNL-PRES-731426 This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC