From Video to Images: Contrastive Pretraining for Emotion Recognition from Single Image (Student Abstract)

Bhanu Garg, Kijun Kim, Sudhanshu Ranjan

University of California San Diego
bgarg@ucsd.edu, kik004@ucsd.edu, sranjan@ucsd.edu

Abstract

Emotion detection from face is an important problem and has received attention from industry and academia. In this paper, we try to use information from videos of human making emotions in a self-supervised manner to recognise emotion from a single image. More specifically, we leverage contrastive loss for pre-training the network on the videos to learn embeddings of the emotion from the image. Once the embeddings have been trained, we test them on a standard emotion classification task. Our method significantly improves the performance of the models and shows the efficacy of self supervision in emotion recognition.

Introduction

Facial expression is one of the main modalities used to help understand the emotional status of an individual, and is an useful contextual clue for social communication. However, this is not an easy task for a machine, because the facial expressions vary among individuals. Further, getting accurate labels is a hard (Kim et al. 2019).

The current state-of-the-art models on popular facial recognition datasets such as FER-2013 (Goodfellow et al. 2015) uses ensemble of convolution and hand designed features to perform classification. While the hand designed features, or ensemble learning do get the best results, we believe that there could be a way to improve the accuracy of learning emotion recognition in an end-to-end manner. Since the datasets such as CK+ provide video frames of emotion from neutral to the peak by varying the intensity of it, we aim to leverage frames at different intensities of emotions to generate triplets to be used in learning embeddings for facial expression recognition.

Furthermore, it is known that the applications of contrastive loss has had tremendous success in facial recognition tasks (Schroff, Kalenichenko, and Philbin 2015). In this paper, we propose to extend the use of contrastive loss to recognize facial expressions.

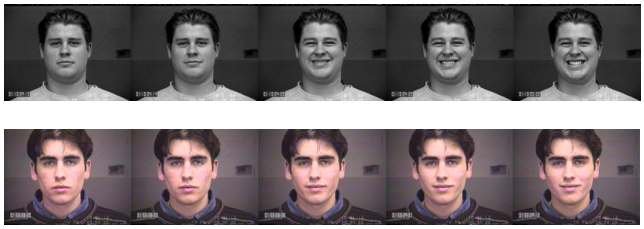

Figure 1: Sample frames of happy emotions from the CK+ dataset. Copyright: ©Jeffrey Cohn

Methodology

Contrastive Loss

Contrastive loss is widely-used in unsupervised and self-supervised learning. Following (Schroff, Kalenichenko, and Philbin 2015), the embedding is represented by $f(x) \in \mathbb{R}$ for an image x. We constrain the embedding to live on the d-dimensional hypersphere $||f(x)||_2 = 1$. This loss function operates on a triplet of data points (anchor, positive and negative) rather than an individual sample itself. We want to ensure the image x_i^a (anchor) with neutral intensity expression is closer to the positive image x_i^p (positive) of a medium intensity expression than the x_i^n (negative) of a high intensity expression. The contrastive loss is defined as:

$$L_{ct} = \left[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha\right]_+$$

where α is a margin that's enforced between positive and negative pairs.

Sampling

We want to generate triplets that are hard i.e. results in a positive loss for the model continually, else the model's learning will be saturated. In the CK+ data, some of the video sequences are labeled and we use this information. Given this, we use the following two sampling processes.

Method 1: Anchor and positive are chosen randomly from a video sequence, such that the anchor is that of lower intensity and positive being higher intensity; negative is a randomly chosen from the video of same person's other emotion. Intuitively, the difference between anchor/positive and negative is only the emotion, the self-supervision contrastive

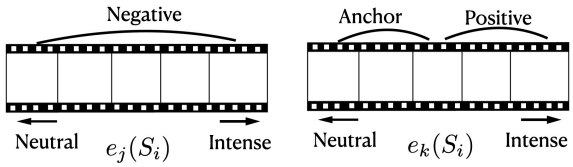

Figure 2: Choosing triplets for Method 1

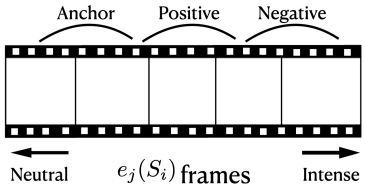

Figure 3: Choosing triplets for Method 2

loss learning should help the model learn emotion representations. The embeddings of similar emotion will be clustered together. Let the set of subjects be $S = \{s_i | i, 1 \leq i \leq n\}$, and the set of emotions under consideration are $E = \{e_i | i, 1 \leq i \leq m\}$. We define $e_j(s_i)$ as the video clip where subject i is expressing emotion j, and R_k be the random choice to pick k sorted numbers from the given range. The equation below and figure 2 below shows the sampling scheme for Method 1, where $k \neq i$.

$$\text{anchor, positive} = R_2(e_j(s_i)) \tag{1}$$

$$\text{negative} = R_1(e_k(s_i)) \tag{2}$$

Method 2: Anchor, positive, and negative frames are chosen randomly from a video sequence, such that the anchor is that of lowest intensity, positive being middle intensity; and negative is the most intense. The idea here is that the difference between anchor/positive and negative is only the intensity of emotion, the self-supervision contrastive loss learning should help the model learn emotion intensity representations, for which it needs to recognise the emotion. The equation below and figure 3 shows the sampling scheme for Method 2.

$$\text{anchor, positive, negative} = R_3(e_j(s_i)) \tag{3}$$

Experiments

We do extensive experiments with both of the above sampling methods, and also their combination to enhance the learning.

The results from the experiments are reported in table 1. We observe an improvement of $\sim 1\%$. We find that the best results are obtained when Method 2 is used after Method 1. The results follow the intuition, as Method 1 helps the model to distinguish between two emotions, while Method 2 brings the embeddings of the same emotion closer.

Conclusion

We explore the applicability of self supervised contrastive learning based methods using video data to learn meaningful representations for single image emotion recognition. Future works shall further experiment with better triplet selection, and newer architectures.

Method	Top1 acc
MAXIM MILAKOV	68.82%
RADU + MARIUS + CRISTI (Ionescu and Grozea 2013)	67.48%
InceptionV1-full_FT	67.73%
InceptionV1-VGGFace2-full_FT	67.92%
InceptionV1-VGGFace2-FC_only	48.86%
CT512_1-full_FT	68.76%
CT512_2-full_FT	**69.15%**
CT512_1-FC_only	41.96%
CT512_2-FC_only	42.21%

Table 1: The top block are the top entries of the Kaggle competetion (Goodfellow et al. 2013). The middle block shows the results of InceptionV1 (without contrastive learning). InceptionV1-VGGFace2 indicates that pre-trained model chosen. full_FT and FC_only indicate full finetuning and FC layer training respectively. The last block are the results of contrastive learning. CT512 indicated embedding size of 512 was used. CT512_1 indicates Method 1 sampling was used, while CT512_2 indicates Method 1 followed by Method 2 sampling was used. Other notations follow.

References

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; Ionescu, R.; Popescu, M.; Grozea, C.; Bergstra, J.; Xie, J.; Romaszko, L.; Xu, B.; Chuang, Z.; and Bengio, Y. 2013. Challenges in Representation Learning: A report on three machine learning contests. arXiv:1307.0414.

Goodfellow, I. J.; Erhan, D.; Luc Carrier, P.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; Ionescu, R.; Popescu, M.; Grozea, C.; Bergstra, J.; Xie, J.; Romaszko, L.; Xu, B.; Chuang, Z.; and Bengio, Y. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64: 59–63. Special Issue on "Deep Learning of Representations".

Ionescu, R. T.; and Grozea, C. 2013. Local Learning to Improve Bag of Visual Words Model for Facial Expression Recognition. ICML'13 Workshop on Representation Learning.

Kim, Y.; Yoo, B.; Kwak, Y.; Choi, C.; and Kim, J. 2019. Deep generative-contrastive networks for facial expression recognition. arXiv:1703.07140.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.