Recognizing Emotions from Facial Features

Jason Chen (cheson), Theodora Chu (theodora), Priyanka Sekhar (psekhar)

Problem Overview

Emotions inform perception and action. Humans are socialized to learn how to act and react based on their understanding of the emotions of those around them. Being able to recognize emotions is something that children learn starting from birth. Because emotion recognition is so important to everyday life, we want to train an algorithm to do this. Highly accurate emotion recognition systems could lead to advances in psychology and sociology, which would lead to an increased understanding of decision-making and consumer preferences, among other things. Specifically, we looked at SVMs and CNNs to compare their performance on emotion classification. We chose to use an SVM because it is a flexible model that tends to outperform other deterministic models. We chose to use a CNN because CNNs have been shown to do well on image recognition tasks and provide versatility in feature learning.

Previous Attempts

In “Image based Static Facial Expression Recognition with Multiple Deep Network Learning,” Yu and Zhang use deep CNNs to determine facial emotions. Their baseline was around 35% accuracy, and they were able to optimize this to hit about 55% accuracy. [1] However, their CNN was also fine-tuned to the dataset they were working with. Due to the processing complexity of CNNs, training on the training set rather than using a pre-trained CNN would take on the order of weeks to run. Other attempts have focused on the escalation of emotion through video frames [2], finding the locations of facial features to inform changes in emotion [3], and using Hidden Markov Models to combine video and audio in emotion detection. [4] The one thing missing from this research is consensus over what features are most relevant to this problem. Previous attempts show that there is a lot of potential for using CNNs on pure image emotion detection. However, we were curious how this would compare if we used a pre-trained CNN. Specifically, we wanted to compare CNNs and SVMs in order to gain a better understanding of what features are extracted and what features are relevant to this problem.

Methodology

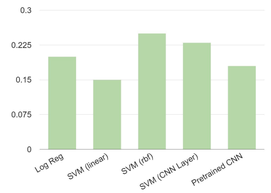
We are using the Extended Cohn-Kanade Dataset, a dataset specifically released for the purpose of encouraging research into emotion detection in images. The faces show a range of 7 emotions. [5][6] Example images in this dataset are shown below. To extract relevant data from these images, we use Google's Cloud Vision API, which has the ability to pinpoint faces and facial features, including their locations. It also has the ability to make emotion label predictions; however, for the sake of self-discovery, we have chosen not to use this function.

For our SVM, we began by cropping the pictures so that the faces were of the same scale and size. Then, we took the Google API and extracted the locations of facial features. Using this, we found the angles between the various facial features. These angles made up our feature vector. We then tuned the C and gamma values, in addition to testing various types of kernels. To ensure random sampling of data and to make sure overfitting does not occur, we applied 5-fold cross validation to our dataset in training and testing. Results of this are shown to the right.

Additionally, a pre trained CNN for emotion classification was used to categorize each image. The first and second fully connected layers of the Caffe ImageNet CNN were extracted and used as input into our SVM.

Key Results

Classification accuracies for each of the models were similar and lower than expected. Simple models somewhat outperformed more sophisticated ones on this dataset. The combination of fc7 features and our manually selected features had the highest performance of the feature combinations.



Interpretation of Results

Much of our difficulty with obtaining performance gains stemmed from the small size of our dataset. The rbf kernel tended to disproportionately choose the most common category (class 7, surprise) even with parameter tuning, limiting its usefulness in practical applications. This bias is likely because not enough data was present to prevent overfitting. Likewise, our CNN could not be trained on our dataset because of large sample size requirements.

What is interesting to note is the the features extracted from the CNN seemed to contribute to SVM performance gains. Despite the CNN’s relatively poor performance, the more generalized fully connected layers served as useful representations. The combination slightly outperformed both the pre trained CNN and the SVM, indicating that transfer learning would be worth further investigation.

SVM

With proper feature selection, SVMs have shown moderate performance on this emotion classification task. To implement the SVM, we used the python sklearn toolkit for training, fitting, and testing. Since we chose to classify a range of 7 emotions, we needed a multiclass SVM, which is provided by sklearn in both the one-vs-one and one-vs-all variations.

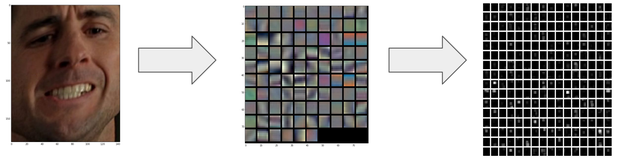
During the training, fitting, and testing process we found that the one-vs-one implementation often performed better than one-vs-rest, which is expected given a unique classifier is constructed for each pair of classes. Normally, the concern for one-vs-one SVMs is expensive computation, but since our datasets of images were relatively small, one-vs-one training was able to run quickly enough. To locate the SVM with the best results, we replicated the experiments with various kernels, mainly including linear, rbf, and sigmoid kernels. For each of these kernels, we also performed parameter tuning by trying the experiments with combinations of C and gamma values across the ranges [0.1, 1, 10, 100, 1000] and [2^-7, 2^-6, ... , 2^-1], respectively.

For feature selection, our naive implementation for SVM used the direct normalized landmark coordinates received from the Google API, which meant for each image the feature vector consisted of 34 tuples of (x, y, z) coordinates, each represented as a float. We then added more sophisticated features by calculating angles between each of the landmarks, which was done by choosing three coordinates, setting one as the vertex, and using the other two to form the left and right legs of the angle. Since this resulted in a computationally expensive process given all the permutations possible, we hand-picked certain facial landmarks that had the potential to change the most to include in our feature vectors. For example, the angle from bottom lip to left and right mouth corners were experimentally determined significant while the tip of nose landmark was removed completely.

CNN

Several Convolutional Neural Networks have been trained on the generic ImageNet dataset. These models are optimized for coarse labels (i.e. ‘window’ or ‘cat’), but the final connected layers of these nets are often fine tuned for more specific tasks on new datasets with different labels.

Visualization of conv1 layer and pool5 layer:



The CNN used for comparison in this project was the transfer learning model submitted to 2015 Emotion Recognition in the Wild contest by H. Win et al. [7] The cascading fine-tuning resulted in 48.5% in the EmotiW validation set and 55.6% in the EmotiW test set. This net was chosen due to its accessibility and relevance to our small dataset.

The Caffe architecture was used to extract the connected layers from the CNN pretrained bvlc reference model. [8] Fc6 and fc7 were used as features in the SVM for our dataset, in combinations with our manually crafted features.

Next Steps

We plan to investigate how fine tuning the CNN on our dataset affects performance. A larger dataset is essential for this task, and thus future work can train and test on this dataset in combination with datasets such as JAFFE. Furthermore, fine-tuning a CNN with higher accuracy on the EmotiW Challenge sets could produce more useful intermediate layers for input into the SVM. Previous work suggests these steps could result in marked increase in performance. More fine-grained parameter tuning will also likely decrease the tendency of the rbf kernel to produce trial results.

Given that the SVM run solely on features of angles between facial landmarks performed the best, we can hypothesize that features that help an algorithm better understand the interactions between facial landmarks are good. However, because the difference between the SVM performance and that of the other methods is statistically insignificant, we would next want to train a complete CNN if we were to continue this project.

Lastly, because we believe that facial landmarks are important features for these kinds of problems, we could investigate explicit indicator functions (i.e. smile exists, furrowed eyebrows, etc.) as additional features to an SVM.

References

[1]Z. Yu and C. Zhang, "Image based Static Facial Expression Recognition with Multiple Deep Network Learning", 2016. [Online]. Available: http://research.microsoft.com/pubs/258194/icmi2015_ChaZhang.pdf. [Accessed: 02- Jun- 2016].
[2]G. Littlewort, M. Bartlett, I. Fasel, J. Susskind and J. Movellan, "Dynamics of Facial Expression Extracted Automatically from Video", *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pp. 80-80, 2004.
[3]L. Luoh, C. Huang and H. Liu, "Image processing based emotion recognition", *2010 International Conference on System Science and Engineering*, 2010.
[4]P. Ng and L. De Silva, "Bimodal Emotion Recognition", 2016. [Online]. Available: https://www.researchgate.net/profile/Liyanage_De_Silva/publication/3845464_Bimodal_emotion_recognition/links/0deec53a2e358f40f3000000.pdf. [Accessed: 02- Jun- 2016].
[5] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.
[6] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101.
[7] Deep learning for emotion recognition on small datasets using transfer learning.
Proc. 17th ACM International Conference on Multimodal Interaction (ICMI), Emotion Recognition in the Wild Challenge, Seattle, WA, Nov. 9-13, 2015.
[8]Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.