Recognizing Emotion from Static Images

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

I. Abstract

The effect of SVMs and CNNs on emotion detection in static images are analyzed in this paper. The investigation into the SVMs included testing various parameters and kernels as well as making multiple hypotheses concerning proper feature vectors: that emotions depend simply on the locations of facial landmarks and that emotions are dependent on the relationship between landmarks. The investigation into CNNs included parameter fine-tuning and extracting the final layer as a feature vector for the SVM. The CNN far outperformed the SVM with a 91.3% accuracy.

II. 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. Because emotion recognition is so important to everyday life, we want to train an algorithm for this task. 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. During the course of our project, we compared the performance of SVMs and CNNs on emotion classification.

III. Model Selection

We considered both generative and deterministic models for our work. While some real-world scenarios may be accurately modeled with certain emotional probability distributions (i.e. people are more likely to be happy than angry at social events), datasets tend to contain a more equal distribution of emotion. Because we determined that there was no clearly discernable or evidently meaningful a priori probability distribution for general emotion detection in a curated set of static images, we decided to focus on more robust discriminative models for this project.

We chose to use an SVM because it has the ability to capture complex relationships in both regression and classification problems. The SVM accounts for nonlinearity in features and can be tuned to balance bias and variance effectively. In previous research conducted on this topic, SVMs were also shown to provide a baseline for understanding what kinds of features are relevant to this problem. Given the wide number of feature possibilities on a given face, the SVM was best suited to our task.

Additionally, we decided to investigate Convolutional Neural Networks. CNNs have been shown to do well on image recognition tasks and provide versatility in feature learning, and they are often considered state of the art for image learning tasks.

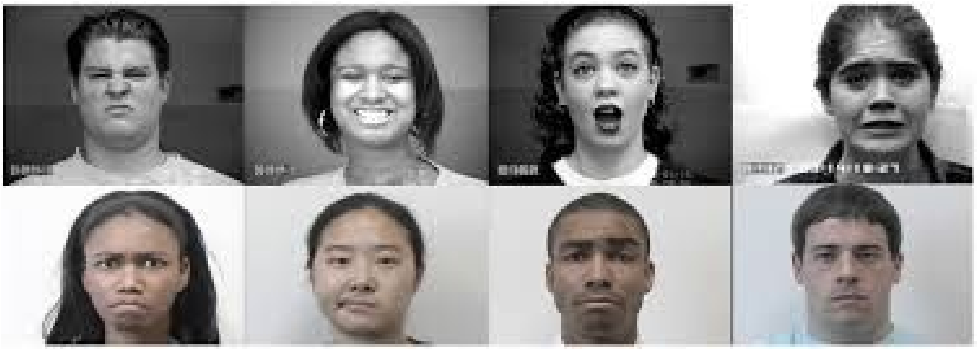
III. Related Work

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.

IV. Methodology

We are using the Extended Cohn-Kanade Dataset, a dataset specifically released for the purpose of encouraging emotion detection research. The faces show a range of 8 emotions. [5][6] 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 probability estimates; however, for the sake of self-discovery, we have chosen not to use this function.

Figure 1: Sample from CK+ Dataset



For our SVM, we began by cropping the pictures so that all the faces were of the same scale and size. Then, we took the Google API and extracted the locations of facial features and facial landmarks (i.e. Corner-of-Eye, Center-of-Mouth). Using these features, we found the geometric 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 did not occur, we applied 5-fold cross validation to our dataset in training and testing.

Additionally, images were classified using a pre-trained CNN. The first and second fully connected layers of the Caffe ImageNet CNN were extracted and used as input for the SVM.

V. SVM

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, such as the 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. However, we hypothesized that when humans interpret facial features, they look at how the features interact with each other to draw a conclusion. 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.

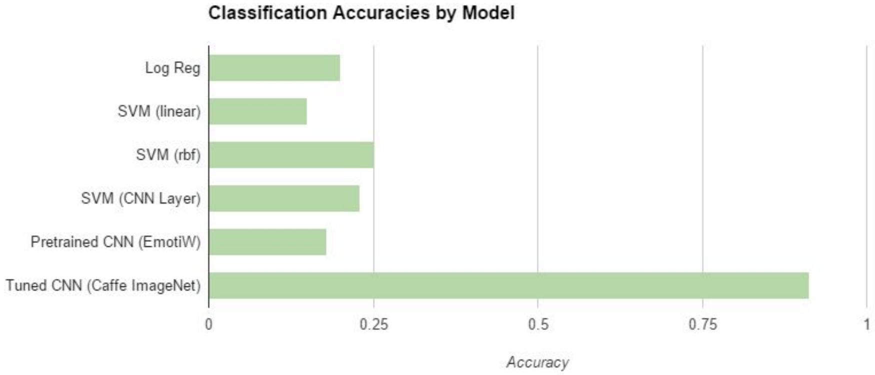
VI. CNN

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. Two different CNNs were used in this project.

The CNN used for direct 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. While this architecture did not win the challenge, it was chosen due to its accessibility, ease of implementation, and relevance to our small dataset. It was used “out of the box” with no fine tuning.

The CNN used for SVM feature extraction was the Caffe architecture, pretrained on ImageNet [8] Fc6 and fc7 of this network were used as features in the SVM for our dataset, in combinations with our manually crafted features. This CNN was also fine-tuned on our dataset and used to generate predictions. The summary of these results is reported below.

Figure 2: Results Summary



Classification accuracies several 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. However, the fine-tuned CNN far outperformed other methods of emotion classification, correctly classifying 91.3% of the test set. The superior performance of the CNN aligns with the results of previous work.

VII. Interpretation of Results

Our SVM did not perform as well as we hoped. 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, the emotion surprise) even with parameter tuning, limiting its usefulness in practical applications. This bias is likely because there was not enough data to allow for meaningful separability given the complexity of our feature set. Because SVM performance is highly dependent on feature sets, we believe that the SVM can be optimized with more directed features. Given that the SVM run solely on features of angles between facial landmarks performed better than the pure landmarks, we can hypothesize that features that help an algorithm better understand the interactions between facial landmarks are better; however, there are still improvements that can be made.

Presently, our features fall on a continuous range of values; yet, this leaves much of the discretization of the data to the algorithm. We believe that using a feature set of indicator functions (e.g. an indicator for a smile or for furrowed eyebrows) would boost SVM performance, as the plane of separability would theoretically become more apparent.

What is interesting to note is the the features extracted from the CNN seemed to contribute to SVM performance gains. Despite the pre-trained 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.

The restrictions of our small dataset also prevented us from attempting to fully train a CNN because of the large sample sizes these models require. The low performance of the emotion recognition CNN from the EmotiW challenge is likely due to nuanced differences between our data and that data used for training. Future studies would separate color pictures from black and white pictures, as well as ensure image size and proportion compatibility to the EmotiW training data before retesting this model. The extremely high accuracy of the tuned Caffe CNN (pretrained on ImageNet) is likely due to some level of overfitting. Because the dataset is so small, the neural net is likely learning features specific to this dataset rather than features relevant to emotion recognition as a whole.

VIII. Future Work

We plan to investigate how fine tuning other emotion-specific CNNs 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. Additionally, fine-tuning on larger datasets could also decrease the effect of overfitting. Future studies could investigate combinations of datasets to add to variation in training and testing data.

Lastly, to further optimize the SVM, we would implement indicator function features as previously mentioned. Related work also suggests that Histogram of Oriented Gradients (HOG) features instead of geometric ones could lead to significant performance gains. [9]

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 &#039;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.
[9] Dahmane, Mohamed, and Jean Meunier. "Emotion recognition using dynamic grid-based hog features." *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011.