Facial Expression Recognition for Wild Images with Analysis from Saliency Maps

Priyanka Rao
Stanford University
prao96@stanford.edu

Ling Li
Stanford University
lingli6@stanford.edu

Abstract

In this project, we explore the problem of facial emotion recognition for "wild" images – images that are spontaneous, unplanned, and have a variety of angles, noise, occlusion, and illumination levels. Our aim is to develop an understanding of how different CNN architectures perform on "wild" images when trained on grayscale and pre-processed images. To achieve this, we implement and examine the performance of both shallow CNN networks and existing architectures such as AlexNet and VGG-16. Through the use of saliency maps, we also investigate what CNNs look for when classifying an image.

1. Introduction

Neural networks and deep learning combined with the proliferation of big data present powerful approaches to problems in many fields, particularly in computer vision. One promiment use of neural networks is facial emotion recognition, which has a myriad of applications in video security, surveillance, advertising, and robotics. Especially for human-robot interaction, facial expressions play a key role in communication and understanding human behavior. In security and surveillance, the exponentially increasing amount of social media images makes it increasingly difficult to screen violent and dangerous images and videos. Facial emotion recognition simplifies and scales this task. In advertising as well, algorithms can present more tailored ads based on emotion recognition.

Fast and accurate facial expression recognition is hence crucial to these applications, particularly for images taken in uncontrolled conditions and crowds. These "wild" images are spontaneous and often contain various angles, poses, illumination conditions, occlusion levels, etc. Emotion recognition in the "wild" is an active and challenging research area, due to the lack of large amounts of labeled training data and the variety of conditions each image is taken in. Existing facial emotion datasets tend to have fixed numbers of human subjects, or a certain number of images per expression with little variation between sets. Rarely does a dataset have a broad variety of environmental conditions, illumination conditions, and subjects like those found in "wild" images.

Traditional machine learning approaches such as support vector machines and Bayesian classifiers have made some progress towards emotion recognition, but in mostly controlled environments. As these approaches are limited to expressions similar to those found in their training data, they struggle to classify images captured in the "wild" or sampled from real-time videos. However, the recent rise of deep learning with neural networks has led to some promising progress in this field [1]. The capability of neural networks to extract undefined features enables them to generalize better to unknown scenarios. A deep neural network has been successfully applied to classify images from a different dataset than the one it was trained on [2].

For facial expression recognition, there are typically three main steps - registration, feature extraction and classification. In the first step, registration, faces are first detected and then geometrically normalized to match some template. During feature extraction, a numerical feature vector is generated from the resulting registered image. These features are either geometric features such as facial landmarks [3], appearance features Local Binary Patterns (LBP) [4], or motion features such as optical flow [5]. Finally, in the final step, classification, a machine learning algorithm is typically used to classify the given face as portraying one of seven basic emotions.

2. Problem Definition

Our problem is two-fold. First, we seek to understand how and why CNNs misclassify certain images. Second, we seek to understand how CNNs transfer their classification ability to "wild" images when trained on grayscale pre-processed images. We aim to implement and compare the performance of different CNNs with varying architectures and depth, including shallow CNN networks and existing architectures such as AlexNet and VGG-16. We also seek to benchmark the performance of our models against the performance of pre-trained models on Imagenet and examine the reasons behind any differences in performance.

While we do try to optimize performance for our models, we realize that after reaching 55-60%, increases in accuracy for facial emotion recognition boil down to tuning hyperparameters. Furthermore, even for humans, accuracies for facial emotion recognition tend to be around 53%, indicating that emotion recognition is a subjective and ambiguous task [17]. This leads us to think that it is not as important to gain incremental increases in accuracy as it is to gain a deeper understanding of why neural networks classify images the way they do. Hence, our focus is less on pure optimization of performance and more on comparing and understanding the performances of different models.

3. Related Work

Compared to existing work that has mainly used traditional machine learning techniques such as SVMs and Bayesian classifiers, we focus on using convolutional neural networks. Some other groups have attempted this in the past. For example, in 2016, A. Mollahosseini *et al* used two convolutional layers followed by max pooling with four inception layers and trained their model on seven publically available facial expression databases including MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013. They achieved a better accuracy and training time than state-of-the-art methods and traditional convolutional neural networks [2]. In 2015, P. Barros *et al* implemented a multimodal emotion recognition system based on spatial-temporal hierarchical features to tackle challenges with illumination, pre-processing, and positioning of the face in images [6].

One of the winners of the EmotiW 2016 challenge applied a hybrid network that combines a recurrent neural network (RNN) with 3D convolutional networks (C3D) in a late-fusion fashion [9]. Another team used pipelined modules, consisting of face detection, image preprocessing, deep feature extraction, feature encoding and an SVM classifier [10]. In 2016, an architecture called HoloNet was proposed, which uses a modified Concatenated Rectified Linear Unit (CReLU) along with increased network depth and residual structure in the middle layers and an inception-residual structure in the top layers [11]. Another group pre-trained their network on ImageNet and then fine-tuned their model on the FER2013 dataset for further feature extraction [12].

The papers we investigated report accuracies ranging from 30% to 68%, compared to reported baselines of 30% to 40%. The highest report accuracy on the Kaggle leaderboard for our dataset is 71% [7].

4. Methods

For our project, we use several different architectures – a shallow 3-layer CNN, AlexNet, VGG-16, Inception, and Inception-Resnet. For the existing architectures, we train them on the Kaggle dataset and benchmark their performance against their performance when pre-trained on ImageNet. The architecture for the three-layer CNN is [conv - relu - 2x2 max pool] – [affine – relu] – [affine] [15]. For our loss, we are using the softmax function because we want our model to try to increase its output probability of the right emotion and decrease its output probabilities for the wrong emotions. One key aspect of our project is that we use transfer learning to train these models on grayscale pre-processed images and test them on RGB images from the "wild".

For the architectures mentioned, we try different filter sizes, network depths, and update rules to see how they impact performance. We also use saliency maps on the existing architectures to understand the differences in what they look for in "wild" images.

5. Datasets

Currently, we are using a dataset from the Kaggle "Challenges in Representation Learning: Facial Expression Recognition Challenge" competition [7]. The provided dataset for this challenge consists of 35887 48 × 48 pixel grayscale images. These are preprocessed images that are centered and adjusted with faces occupying almost the same amount of space in each image. We divided them into 28709 images for training, 3589 for validation, and 3589 for testing. This dataset contains the pixel values and emotion label for each image. The labels are numerical and represent one of seven categories of facial expressions: anger (0), disgust (1), fear (2), happiness (3), sadness (4), surprise (5), and neutrality (6). However, among them, it is important to note that there are only 547 images in the disgust category. We discuss the effect of

this on our results later in the paper.

For our test data, we have tried to test our models both on the test images from the Kaggle dataset (grayscale images) as well as sampled wild images from the Labeled Faces in the Wild database [14]. The Labeled Faces in the Wild database consists of 13,000 unconstrained images of public figures collected from the internet. These images are not centered or processed in any way, except that we convert them to grayscale for consistency before we run our models on them.

6. Experiments and Results

Before we ran our three-layer CNN, we first sanity checked the loss and gradients, which matched our expectations. Since we have seven classes of emotions, we got a loss of 1.94, which is what we expected. We then trained our network for 5 epochs with a batch size of 50. We experimented with different weight scales and numbers of hidden dimensions and learning rates and found the following as our current optimum.

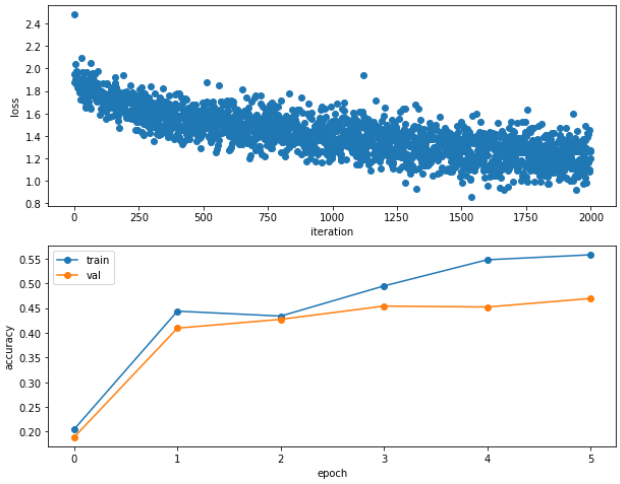

Fig 1. The training loss is shown over 2000 iterations and the training and validation accuracies are shown over 5 epochs. We set the number of hidden dimensions as 500 and used a weight scale of .001. regularization of .001, learning rate of 1e-4, and the RMSProp update rule.

In Fig. 1 above, we see that training loss quickly decreases over 2000 iterations and both training and validation accuracy increase over 5 epochs. Our final accuracy with the shallow three-layer CNN was 49.74%.

In addition to playing around with different hyperparameters for the three-layer CNN, we also ran an experiment on what update rule works best for facial emotion recognition. In Fig. 2, we found that RMSProp update rule works best, even over the Adams optimizer. This was surprising to us since the Adams update rule has bias correction and momentum, which we thought would lead to better performance. However, it seems that for facial emotion recognition, bias correction and momentum take away from the benefits of having an adaptive learning rate. It was also interesting to note how unstable SGD was compared to the other update rules. Momentum, bias correction, and decaying learning rates all create more stable updates as expected.

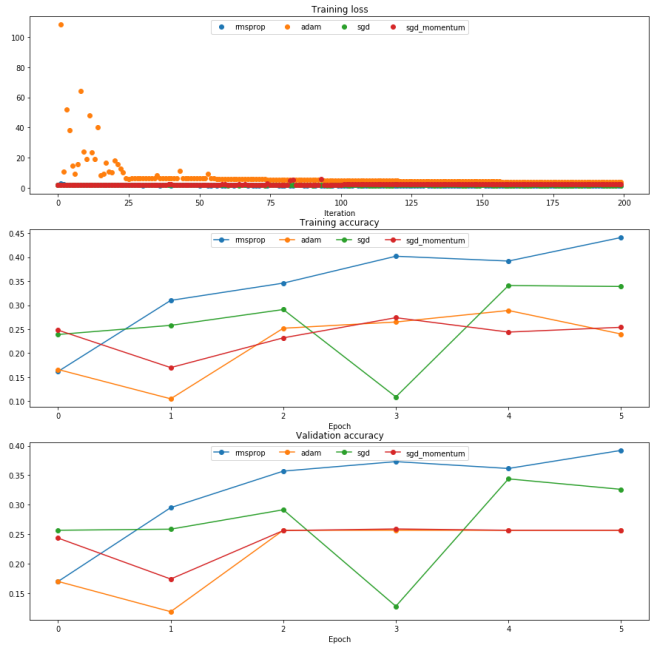

Fig 2. The training loss is shown over 2000 iterations and the training and validation accuracies are shown over 5 epochs for different update rules (RMSProp, Adam, SGD, SGD with momentum)

		Predicted						
		Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Actual	Anger	35.6%	0.0%	11.8%	9.6%	30.5%	2.0%	10.4%
	Disgust	30.9%	23.6%	3.6%	5.5%	29.1%	1.8%	5.5%
	Fear	13.3%	0.0%	27.7%	8.5%	33.5%	9.3%	7.8%
	Happy	4.7%	0.0%	4.8%	68.4%	17.3%	1.4%	3.5%
	Sad	9.4%	0.2%	9.6%	9.3%	57.6%	3.4%	10.6%
	Surprise	4.3%	0.0%	14.2%	6.3%	10.6%	60.1%	4.6%
	Neutral	8.0%	0.3%	5.1%	12.6%	29.6%	3.2%	41.2%

Fig 3. Confusion matrix for the three-layer CNN

Looking at the confusion matrix shown in Fig. 3, there are a few interesting patterns to notice. Except for disgust and fear, the three-layer CNN classifies an image into the correct emotion class the majority of the time. Fear is most often classified as sadness and disgust is most often classified as anger. Fear and sadness share the same characteristics of pulled apart lips and tense foreheads and are often both present in an expression, so this may be why they are confused by the CNN. Similarly, disgust and

anger share the same characteristics of burrowed eyebrows, narrow/pursed lips, and glaring eyes. They are also emotions that tend to occur together in an expression.

This again goes to show how ambiguous and subjective the task of emotion recognition is. Although there may be a "misclassification" of fear as sadness and disgust as anger, if we allow for multiple emotions to be considered correct, then the three-layer CNN correctly classifies emotions the majority of the time.

Another pattern to notice is that there are much fewer misclassifications for happiness and many more misclassifications for disgust overall. For disgust, the higher number of misclassifications may be due to the lack of many images in the training data from this class. For happiness, we reason that the higher number of correct classifications for this emotion is due to not only more "happy" images in the training data, but also due to the unique features that represent being happy. These features, such as a smile and relaxed facial muscles, rarely occur in tandem in images with the other emotions. Thus, it seems to be easier for the three-layer CNN to correctly classify happy expressions.

Shallow CNN vs. Deep CNN

To understand how performance differs based on the depth of the neural networks, we studied the performances of deeper neural networks from existing architectures such as AlexNet, VGG-16, Inception and Inception-Resnet. We used the pretrained VGG-16 on Imagenet from the TensorFlow-Slim image library [16] and fine-tuned the last three fully-connected layers on FER2013. This gave us a test accuracy of 34% which was the baseline for our later training. The training loss and training accuracy for the fully trained AlexNet, VGG-16, Inception and Inception-Resnet are shown in Fig. 4 below.

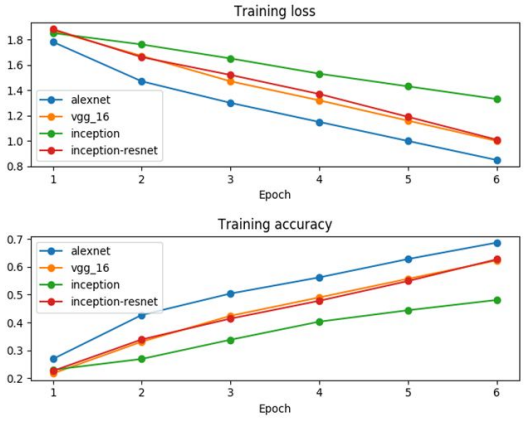

Fig. 4 Training loss and training accuracy for AlexNet, VGG-16, Inception and Inception Resnet after different epochs.

We used a learning rate of 1e-4 for all four models and got a test accuracy of 51%, 49.4%, 46.3% and 45.7% for AlexNet, VGG-16, Inception and Inception-Resnet respectively. As for all these existing model

The performance can be further improved by fine tuning the hyper-parameters. However, as mentioned earlier, wild image facial expression recognition is a challenging task not only for the machine learning algorithms but also for humans too. Experiments in [17] show human can only achieve 53% accuracy in terms of Fleiss kappa to classify AFEW video clips without extra information such as audio track. Besides, we would like to study how the trained models performs on completely new, wild dataset without pre-center or any pre-processing. Thus instead of spend lots of time to improve the model accuracy by several percent, we choose to study how the neural network gets to learn and transfer its training to images from new dataset. Fortunately, saliency map provide a tool for us to probe into the neural network.

Saliency Maps for Deep Neural Networks

To get a better understanding of what the neural network learns during training, we studied how the saliency maps change over different training epochs, as shown in Fig. 5. For example, when learning the expressions "sad", "happy", and "fear", the AlexNet model is initially confused with gradients diverging everywhere. However, after two or three epochs, the model learns to focus on the eyebrow for the sad image, mouth for the happy image and eyes for the image with fear.

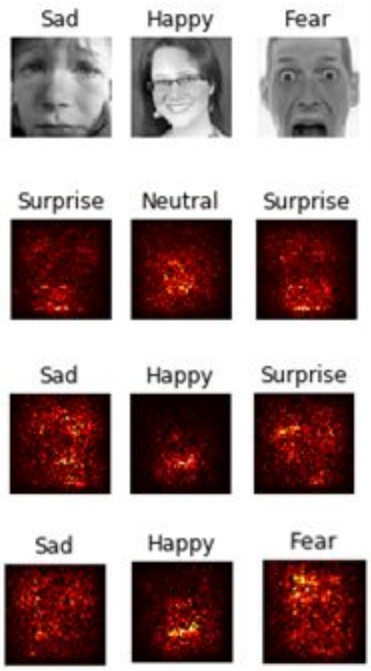

Fig. 5 The samples images from Fer2013 dataset (first row), saliency

map and the result expression labels after first epoch (second row), after two epoch (third row) and after three epoch (fourth row)

We also compared the saliency maps of the several test images from the models trained from scratch with both AlexNet and VGG-16 architectures as shown below:

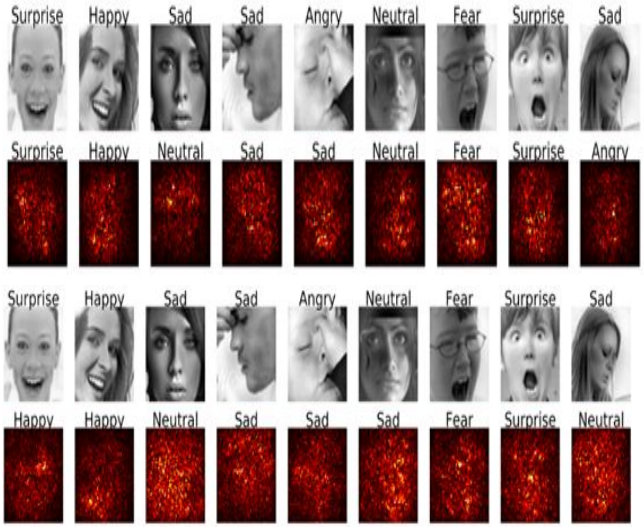

Fig 6. Saliency maps for AlexNet (the top row) and VGG-16 (the bottom row)

After training, the test accuracy for AlexNet was 57% and 55% for VGG-16. We found that the saliency maps are quite different with these two models even though they have very similar test accuracies. In general, the saliency map from Alexnet is darker than that from VGG-16 and converges closer to the face area. Especially for VGG-16, facial expression recognition tends to be wrong when the model fails to find the face itself. In addition, for both models, the mouth and eyes are the key points for expression detection.

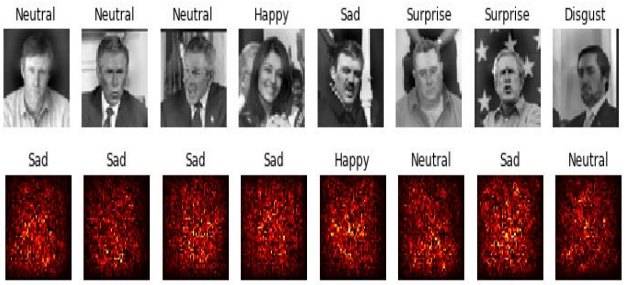

Fig 7. The original images and the saliency map from Labeled Faces in the Wild database from the AlexNet

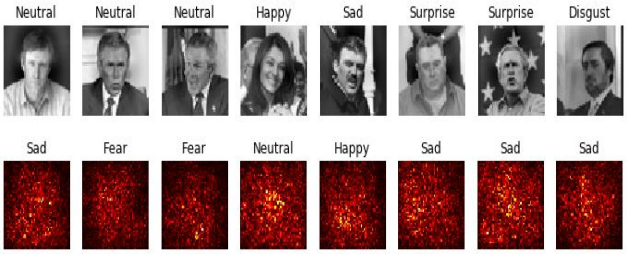

Fig 8. The original images and the saliency map from Labeled Faces in the Wild database from the VGG-16

After training the AlexNet and VGG-16 models on images in the FER2013 database, we tested them on 8 random images from another dataset of Labeled Faces in the Wild database. Since the images from the Labeled Faces in the Wild database are not centered, both AlexNet and VGG-16 failed to detect the face area. Thus, they could not find the key points such as the mouth or eyes as they did before. Overall, considering the whole training process, both models are only exposed to the centered face area and thus had no opportunity to learn to detect faces. On the other hand, to improve the accuracy on wild images, based on our experience as human for expression detection, we need to further incorporate a dynamic searching algorithm to move the active detection bound box around or do a face segmentation first.

7. Conclusion

The three-layer CNN we implemented and the existing architectures we trained gives us accuracies between 49 and 57%. Experiments on hyperparameters and update rules showed us that regularization is critical to not overfit to this kind of dataset and that the RMS Prop update rule is best for facial emotion recognition data. When tested on wild images, the CNNs performed decently well and we gained an understanding of what features they look for and why certain misclassifications occur.

References

[1] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C¸ . G¨ulc¸ehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. *In Proceedings of the 15th ACM on International conference on multimodal interaction, pages 543–550. ACM,* 2013.

[2] Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE,* 2016.

[3] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. *In Systems, Man, and Cybernetics, Computational Cybernetics and*

Simulation., 1997 IEEE International Conference on, vol. 4, pages 3732–3737. IEEE, 1997

[4] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 2, 7

[5] M. Kenji. Recognition of facial expression from optical flow. IEICE TRANSACTIONS on Information and Systems, 74(10):3474–3483, 1991. 2

[6] Barros, P., Weber, C., and Wermter, S. Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction. *In Humanoid Robots (Humanoids), IEEE-RAS 15th International Conference on (pp. 582-587). IEEE*, 2015

[7] Kaggle competition: Challenges in representation learning: Facial expression recognition challenge, 2013.

[8] The fifth Emotion Recognition in the Wild (EmotiW) 2017 challenge will be held at ACM International Conference on Multimodal Interaction (ICMI) 2017, Glasgow. EmotiW 2017

[9] Fan, Y., Lu, X., Li, D., & Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 445-450). ACM, 2016

[10] Bargal, S. A., Barsoum, E., Ferrer, C. C., and Zhang, C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 433-436). ACM, 2016

[11] Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., & Chen, Y. HoloNet: towards robust emotion recognition in the wild. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 472-478). ACM, 2016

[12] Ding, W., Xu, M., Huang, D., Lin, W., Dong, M., Yu, X., & Li, H. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 506-513). ACM, 2016

[13] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. 2012.

[14] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[15] Code from Assignment 2 of CS231N was used

[16] TensorFlow-Slim library, https://github.com/tensorflow/tensorflow/tree/master/tensorflow/contrib/slim

[17] T. Gehrig and H. K. Ekenel. Why is facial expression analysis in the wild challenging? In Proceedings of the 2013 on Emotion recognition in the wild challenge and workshop, pages 9–16. ACM, 2013. 3