

Emotion Detection and Sentiment Analysis of Images

Vasavi Gajarla

Georgia Institute of Technology
vgajarla3@gatech.edu

Aditi Gupta

Georgia Institute of Technology
agupta448@gatech.edu

Abstract

If we search for a tag "love" on Flickr, we get a wide variety of images: roses, a mother holding her baby, images with hearts, etc. These images are very different from one another and yet depict the same emotion of "love" in them. In this project, we explore the possibility of using deep learning to predict the emotion depicted by an image. Our results look promising and indicate that neural nets are indeed capable of learning the emotion essayed by an image. These kinds of predictions can be used in applications like automatic tag predictions for images uploaded on social media websites and understanding sentiment of people and their mood during/after an election.

1 Introduction

Nowadays, people share a lot of content on social media in the form of images - be it personal, or everyday scenes, or their opinions depicted in the form of cartoons or memes. Analyzing content like this from social media websites and/or photo-sharing websites like Flickr, Twitter, Tumblr, etc., can give insights into the general sentiment of people about say Presidential elections. Also, it would be useful to understand the emotion an image depicts to automatically predict emotional tags on them - like happiness, fear, etc.

As a part of this project, we aim to predict the emotional category an image falls into from 5 categories - Love, Happiness, Violence, Fear, and Sadness. We do this by fine-tuning 3 different convolutional neural networks for the tasks of emotion prediction and sentiment analysis.

2 Related Work

There exists an affective gap in Emotion Semantic Image Retrieval (ESIR) between low-level features and the emotional content of an image reflecting a particular sentiment, similar to the well-known semantic gap. A lot of previous work tries to address this issue, using both handcrafted features and neural networks. We briefly

describe some of the most influential work that addresses this issue.

Visual Sentiment Prediction with Deep Convolutional Neural Networks proposed by Xu et al. [2] use Convolutional Neural Networks pretrained on Object recognition data to perform sentiment analysis on images collected from Twitter and Tumblr. Robust image sentiment analysis using progressively trained and domain transferred deep networks by You et al. [3] uses VGG-ImageNet and its architectural variations to study sentiment analysis on Twitter and Flickr datasets. Recognizing image style by Karayev et al. [4] experiment with handcrafted features like L^*a^*b color space features, GIST and saliency features on Flickr style data, Wikipaintings and AVA Style data. Emotion based classification of natural images by Dellagiacoma et al. [5] uses MPEG7 color and edge descriptors to perform emotion classification and compares the results with Bag of Emotions method.

3 Our work

Firstly, we finalized the emotional categories to perform the classification on. We then collected data from Flickr for these categories. We experimented with various classification methods on our data - SVM on high level features of VGG-ImageNet, fine-tuning on pretrained models like RESNET, Places205-VGG16 and VGG-ImageNet - more details follow in Section 4.

3.1 Deciding Emotion Categories

We are inspired by the 6 emotion categories defined by the famous psychologist Ekman in [1] - Happiness, Sadness, Fear, Disgust, Anger and Surprise. These emotions are used by several other papers in this domain ([5], [7]). However instead of 'Anger' we chose 'Violence' as we feel it is more suited to the applications we are looking at - more and more people are turning to social media to show their support for causes or protest against something (For example, a lot of social messages poured in after terrorist attacks in Paris in November, 2015). For the same reason, we add 'Love' as an emotion category. To limit ourselves

to 5 emotions we dropped ‘Disgust’ and ‘Surprise’. So our final emotion categories are: **Love, Happiness, Violence, Fear and Sadness.**

3.2 Data Collection

We collected data for the selected 5 emotional categories - Love, Happiness, Violence, Fear, and Sadness from Flickr. The following are the steps involved in collecting this data.

Step 1: Querying the image metadata.

We used the Flickr’s API service to query for images using each of the emotional categories - Fear, Happiness, Love, Sad, and Violence - as search query parameters (along with the license flag set to ‘creative commons’) to collect the image metadata (like server and farm ID numbers on which the image is stored) after sorting the result set by interestingness to make sure that images fitting to the emotional categories are retrieved on priority.

Step 2: Downloading the images from Flickr server.

Once the image metadata is retrieved, it was possible to use that metadata to directly access the images from Flickr servers instead of going through the Flickr API [8].

For this project, we collected 9854 images in total with ~1900 images in every category. We split the data in each category such that 75% of the data (~8850) is used for training, and 25% of the data (~1000) is used for testing.

4 Experiments

In this section we demonstrate all the experiments we conducted on our data.

4.1 One vs All SVM

The idea is to use a pretrained neural network to get the feature representation of our dataset and then use this representation as an input to SVM and classify the data using one vs all SVM classifiers. We used VGG-ImageNet [9] as the pretrained model.

The following are the steps involved in this experiment:

Step 1: Pass the data (both train and test) as input to VGG-ImageNet model.

Step 2: Store the activations from second-to-last fully connected layer of the network as feature vectors.

Step 3: Train a one vs all SVM classifier for each emotion category.

Step 4: For each of the test image, find the maximum of the scores from each SVM to get its predicted category label.

We obtained an accuracy of 32.9% on the test data. The obtained confusion matrix is shown in Figure 1.

4.2 Fine-tuning VGG-ImageNet

The results obtained for the One vs All SVM classifier are better than chance (20%) but still far from good. Therefore, we experimented with fine-tuning the VGG-ImageNet model. We added a dropout layer followed by a fully connected layer and trained it on our dataset. We obtained a maximum testing accuracy of 38.4%.

4.3 Sentiment Analysis

The confusion matrix (Figure 1) indicates that there exists more confusion within each of the positive (Love, Happiness) and negative (Violence, Fear, Sadness) emotions. To understand our results better, we perform sentiment analysis on our dataset. We obtain a testing accuracy of 67.8% which indicates that the network is learning to distinguish between the overall positive and negative categories, but is having trouble classifying within each of the positive and negative sentiments. We suspect this is because the network is learning colors - positive sentiments usually have bright colored images and negative sentiments correspond to dark images.

4.4 Fine-tuning VGG-Places205

VGG-ImageNet is pretrained on ImageNet which consists mostly of iconic images. However, our dataset has both iconic, as well as non-iconic images (for example, in Violence category, many images contain protests which have a lot of people). Therefore, we experimented with fine-tuning a model pretrained on a scene-based database called Places205. We hypothesize that this will work better, since our task is closer to scene classification than object recognition or detection. We fine-tune the VGG model trained on the Places205 database. Using this method, we obtain a testing accuracy of 40.9% on the emotion classification task and an accuracy of 68.7% on the sentiment analysis task. The network’s confusion matrix can be seen in Figure 1.

4.5 Fine-tuning ResNet-50

We also experiment with fine-tuning a pretrained ResNet-50. Because this network is very deep (has 50 layers), and is trained on both scene-centric data (MS COCO) as well as object-centric data (ImageNet), we expected it to give better results. We did get better results with this approach - 40.9% of testing accuracy on emotion classification task

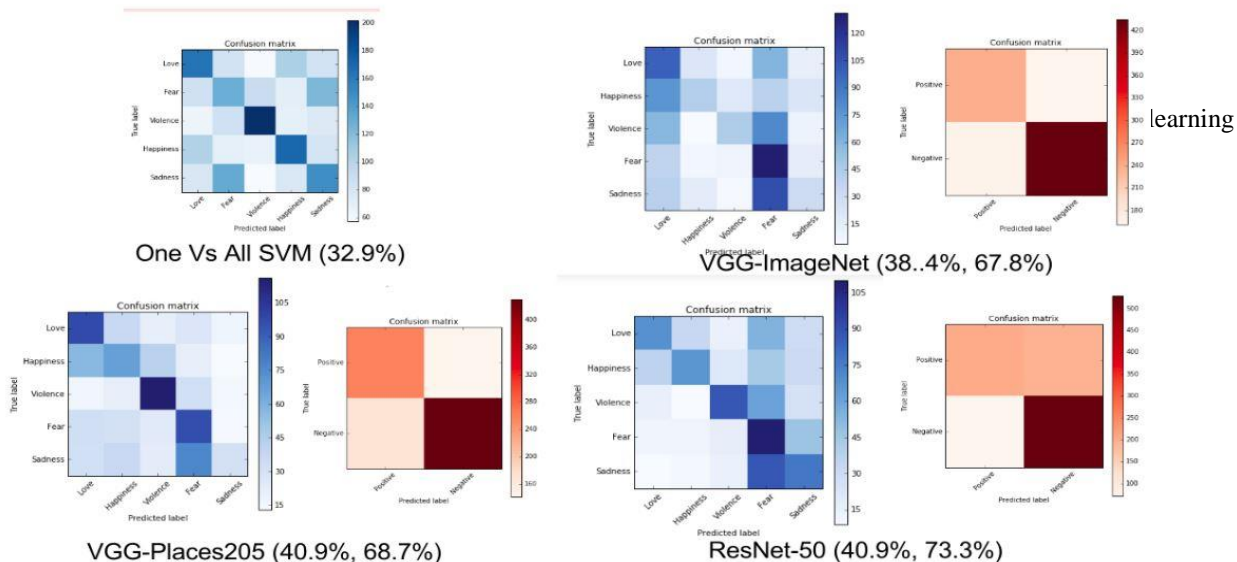


Figure 1: (row-wise starting from top left) Confusion matrix of emotion classification for One vs All SVM, Confusion matrix of emotion classification and sentiment analysis for CGG-ImageNet, 205PlacesVGG-16 and ResNet-50

and an accuracy of 73% on sentiment analysis task. The confusion matrix of this network can be seen in Figure 1.

In all of the fine-tuning experiments, it has been observed that standard data augmentation techniques like mirroring and random cropping of images increase the accuracy. Having a higher learning rate for the later layers also yields a better accuracy. We used Stochastic Gradient Descent optimization for all the networks, and the base learning rate at a start value of 0.0001 gave better accuracies.

5 Analysis of Results

In this section we examine the results of the ResNet-50 model fine-tuned on our dataset, to get a better understanding of what our network is learning. Based on the results, we can draw the following conclusions about what the model is learning for each category. See Table 1 for more results.

Love: The model seems to be learning sunsets, hearts, red color, flowers and serene landscapes for this category. A lot of sunsets which are tagged as “Happiness” get classified as love.

Happiness: The model seems to learning faces, bright colors and images depicting Bhutanese people! Hence an image depicting a sad Bhutanese woman get classified as Happiness. We also observe that 80% of the images that are tagged as happiness are face images. Hence this category is biased towards faces.

Violence: This category has the best accuracy. A close look at the true positives of this category tells us why. Images depicting Violence are very different from images

posters, demonstration scenes, officers and text, ‘Je Suis Charlie’ in particular. It probably because of the labeling for this class is unambiguous.

Fear: The model seems to be learning dark, haunting images and spiders.

Sadness: This class has the maximum variation. The model seems to be learning a lot of face images.

Table 2 shows some misclassified instances. These examples seemed to have been rightly classified (emotion-wise) but fall under misclassifications with respect to the actual tagged label. For example, the first image in the row of fear is dark and dull, which we suspect has led the network to classify it as fear but it was actually tagged as happiness. These examples reiterate the importance of having a clean dataset.

6 Comparison to Baseline

We compared our results to some other approaches like the ones mentioned in ‘Related Work’ section ([2], [3], [4] and [5]). These methods have been trained and tested on different datasets and the comparison is not fair, but we believe it does give an idea about how the methods which use hand-crafted features for emotion classification and a few methods which use deep learning techniques for sentiment analysis perform. Table 3 and Table 4 show the datasets used by various methods and the accuracies achieved.

Method	Datasets Used	Accuracy
MPEG7	Manually Curated	0.591
Fusion x Content	AVA Style	0.581
Fusion x Content	Flickr	0.368
Fusion x Content	Wikipaintings	0.473
Places205-VGG16	Flickr	0.409
VGG-ImageNet	Flickr	0.384
ResNet-50	Flickr	0.409

*Fusion x Content: L*a*b, Hist, GIST, Saliency

Table 3: Comparison of our experiments on emotion classification with respect to a few methods using hand-crafted features for the same task

Method	Datasets Used	Accuracy
FC7	Twitter & Tumblr	0.649
3CONV-4FC	Flickr & Twitter	0.644
3CONV-2FC	Flickr & Twitter	0.657
2CONV-3FC	Flickr & Twitter	0.654
2CONV-4FC	Flickr & Twitter	0.665
Places205-VGG16	Flickr	0.687
VGG-ImageNet	Flickr	0.678
ResNet-50	Flickr	0.733

Table 4: Comparison of our experiments on sentiment analysis with respect to a few methods using deep learning for the same task

7 Challenges

The problem of labeling images with the emotion they depict is very subjective and can differ from person to person. Also, due to cultural or geographical differences some images might invoke a different emotion in different people - like in India people light candles to celebrate a festival called "Diwali", however in western countries candles are lit, most of the times, to mark an occasion of mourning.

In addition to the above, there is another fundamental challenge involved in tackling the problem of extracting emotional category from images. A person could have tagged an image as, say "Fear", it could be because the image makes them feel that emotion when they look at it (Figure 2a), or the subjects/objects (people, animals, abstract art, etc.) in the image show that emotion (Figure 2b). This kind of image tagging can confuse the network.



Figure 2a: Image which causes fear in viewer



Figure 2b: Image with a woman experiencing fear

The above problems can be solved by collecting clean data using a crowd-sourced framework like Amazon Mechanical Turk and labeling the images based on the consensus reached by the workers, provided they are instructed to tag the images purely based on the emotion the image depicts rather than any personal opinions of the workers.

We also observed that the results improved when we augmented the input data to pretrained neural networks by performing mirroring and random cropping on them. This shows that large amount of data definitely helps in fine-tuning the network better.

8 Future Work

Our experiments demonstrated that Deep Learning does give promising results in both the classification of emotions as well as in performing sentiment analysis, even on the raw data collected directly from Flickr. The next step could be to run these experiments on bigger and

cleaner datasets and see if they would improve the results. We believe they would. It would also be interesting to see what the salient regions in the images are, perhaps by using [6], and feed these preprocessed images to pretrained neural networks. The idea behind this is that humans would decide the emotion detected by an image by looking at some salient regions in the image. By notifying the network of these beforehand might help in the training process.

9 Conclusion

Our results show that deep learning does provide promising results with a performance comparable to some methods using handcrafted features on emotion classification task, and also a few methods using deep learning for sentiment analysis. Emotion classification in images has applications in automatic tagging of images with emotional categories, automatically categorizing video sequences into genres like thriller, comedy, romance, etc.

References

- [1] P. Ekman, W. Friesen, and P. Ellsworth. Emotion in the human face. Pergamon Press New York, 1972.
- [2] Xu, Can, et al. "Visual sentiment prediction with deep convolutional neural networks." *arXiv preprint arXiv:1411.5731* (2014).
- [3] You, Quanzeng, et al. "Robust image sentiment analysis using progressively trained and domain transferred deep networks." *arXiv preprint arXiv:1509.06041* (2015).
- [4] Karayev, Sergey, et al. "Recognizing image style." *arXiv preprint arXiv:1311.3715* (2013).
- [5] Dellagiacoma, Michela, et al. "Emotion based classification of natural images." *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*. ACM, 2011.
- [6] Judd, Tilke, et al. "Learning to predict where humans look." *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 2009.
- [7] Ulinski, Morgan, Victor Soto, and Julia Hirschberg. "Finding emotion in image descriptions." *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2012.
- [8] For downloading Flickr images:
http://graphics.cs.cmu.edu/projects/im2gps/flickr_code.html
- [9] http://www.robots.ox.ac.uk/~vgg/research/very_deep/


























<p>Love</p>					
<p>Happiness</p>					
<p>Violence</p>					
<p>Fear</p>					
<p>Sadness</p>					

Table 1: Examples of true positives for each of the categories.





Love	 <p>Fear (Sunset)</p>	 <p>Happiness (Dog)</p>	 <p>Sadness (Flower)</p>	 <p>Violence (Bright)</p>
Happiness	 <p>Fear (faces)</p>	 <p>Love (face)</p>	 <p>Sadness (Bhutanese)</p>	 <p>Violence (Bright colorful)</p>
Violence	 <p>Fear (Text)</p>	 <p>Happiness (Officers)</p>	 <p>Love (Je suis charlie)</p>	 <p>Sadness (Officers)</p>
Fear	 <p>Happiness (Dark, gray scale image)</p>	 <p>Love (dark, haunting)</p>	 <p>Sadness (Spider-like)</p>	 <p>Violence (face)</p>
Sadness	 <p>Fear (Face)</p>	 <p>Happiness (Face)</p>	 <p>Love (Grave scene?)</p>	 <p>Violence (Crying face)</p>

Table 2: The row shows predicted class, and columns are actual class.