Group: MLG55
Real Time Sentiment Analysis

Mentor: Gopichand Kotana

Mithlesh Kumar
mithlesh@iitk.ac.in

Faizanurrahim Ansari
afaizan@iitk.ac.in

Prince Gaurav
princ@iitk.ac.in

Puneet Kumar Verma
puneetkv@iitk.ac.in

Abstract

Many people have used Convolutional Neural Networks to classify sentiment of an image into two classes: positive and negative class. In this project, we explore CNNs to classify an image into five sentiment classes: fear, happpiness, love, sadness and violence. We also use these classifiers to predict real time sentiment of videos. Later we do a thorough analysis of results obtained.

1 Introduction

In the modern Internet era, people are posting images on social media in bulk quantity. Given the huge amount of information available on social media related to any event in the form of visual content, a significant amount of information can be extracted to analyze general public sentiments. Attempts have been made to predict box office earnings, election results etc. by analyzing online users' sentiments.

In this project, our aim is to perform real time sentiment analysis of videos. First we fine-tune 3 types of CNNs to predict the sentiment of an image into 5 categories: fear, happiness, love, sadness and anger. Next we extract frames of a video and feed it into fine-tuned model for the task of real time sentiment prediction. Our focus is on the classification of the general scene of the image and not on the facial expressions of the characters present in it.

2 Related Work

There has been significant work done in the field of sentiment classification of images containing facial expressions. Following are some of the papers which influenced us most.

Visual Sentiment Prediction with Deep Convolutional Neural Networks by Xu et. al.[1] uses a pretrained Convolutional Neural Networks (CNNs) on a large-scale image dataset for object classification, and then the learned parameters of the network are transferred to the task of sentiment prediction for generating image-level representations. Robust image sentiment analysis using progressively trained and domain transferred deep networks by You et al.[2] uses VGG- ImageNet and its architectural variations to study sentiment analysis on Twitter and Flickr datasets. Campos et al.[3] uses fine-tuned CNN to classify images into positive and negative sentiments. The closest work to our project has been done by Poria et al.[4]. This work involves emotion and sentiment recognition from speech, textual data and facial expressions on YouTube dataset.

3 Our Work

We used similar ideas from the previous works on sentiment classification to fine tune a CNN pre-trained on a large scale database like ImageNet for object classification. This involves a technique known as transfer learning in which the parameters learned while solving one problem is used to solve another similar problem. To do this we took a CNN pre-trained on ImageNet for classification on objects into 1000 categories, and removed its final layer to include a new fully connected layer with 5 nodes corresponding to the 5 emotions. Now we train this model on our dataset.

3.1 Data Collection

We have used Flickr's API[5] service to query for images using each of the emotional categories - Fear, Happiness, Love, Sad, and Violence - as search query parameters to collect the image metadata. The images were obtained using the 'relevance' sorting criteria to get the best possible results. The images were then downloaded using this metadata using the API.

We obtained a total of 3899 images for training with around 780 images in each category. For validation we obtained another 500 images with 100 in each category. We have used held out validation technique since cross fold validation takes a lot of time in training. Finally for testing, we obtained 1000 images with 200 in each category.

4 Experiments

We experimented with 3 popular CNNs namely - VGG16, Resnet50 and Alexnet. In all our experiments, we used standard image preprocessing techniques like random zooming, rotating and rescaling using Keras Image Preprocessing. Stochastic gradient descent was used for optimization with a starting learning rate of 10^{-4}, Nesterov momentum of 0.9 and a decay of 10^{-6}.

4.1 Fine tuning VGG16

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 34%.

4.2 Fine tuning ResNet50

Since VGG16 was trained on task of object recognition and our task is closer to scene classification than object recognition, we tried using ResNet50 which was trained on both scene-centric data (MS COCO) as well as object-centric data (ImageNet). We expected it to give better results. However that did not happen and we got an accuracy of around 27%. We suspect that this was possibly because ResNet50 is much deeper than VGG16 and due to our smaller dataset the fine tuning did not happen as expected.

4.3 Fine tuning with AlexNet

We finally tried using a model with fewer layers keeping in mind our smaller dataset. Here, we trained last two layers on our dataset, keeping rest of the layers non-trainable. We trained it with 80 epochs, keeping batch size 128. We did got a better results this time with a testing accuracy of around 37%.

5 Results

In this section, we examine the results of AlexNet model which was fine-tuned on our dataset to get better insights of our model. Following plot (1) shows the plots of validation accuracy and cross-entropy loss versus time.

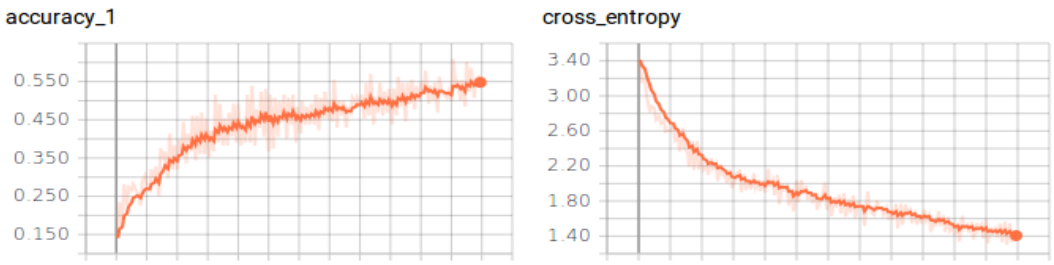

Figure 1: Plots of validation accuracy and loss function Vs time

5.1 Analysis of Results

In order to get better understanding of what our network was learning, we make a table showing the percentage of predictions made by our model for each sentiment class. Based on the table (1), we can make following conclusions about what the model is learning for each category.

- *Fear:* This class has low accuracy and high variance. The model is getting confused with sadness and violence class and classifies most of the images as sadness and violence.
- *Happiness:* The model seems to be learning faces and bright colours in this class. We observe that most of the images classified as happiness are images containing human face in it. It is classsifying more number of images into violence class. One reason could be that it is classifying images with group of people as violence.
- *Love:* Our training dataset has mostly pink coloured images and heart shaped images in this class. So, the model seems to be classifying images with pink and red colour in it as love. This class has the least accuracy among all classes. One possible explanation for this could be that labelling for this class is quite poor.
- *Sadness:* This class has reasonable accuracy. This model seems to be learning dark colours and people with sad faces in it.
- *Violence:* This class has the best accuracy and least variance among all classes. The model seems to be learning group of people with placards in this class. One reason for its accuracy could be that labelling for this class is unambigious.

Table 1: Confusion matrix

Actual class label	Predicted class label %				
	fear	happiness	love	sadness	violence
fear	16	11	11.5	37.5	24
happiness	17	22	14	14.5	32.5
love	15.5	15.5	11.5	29.5	28
sadness	23.5	10.5	13.5	36.5	16
violence	24.5	5.5	8.5	6	55.5

6 Possible reasons for low accuracy

- To understand our results better, we perform sentiment analysis on our dataset by clubbing the happiness andn love class into one class called positive and the remaining three classes into another class named negative. We obtain a testing accuracy of 71.1% 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.

- Psychological factors also comes into play while taking about emotions. Some images might invoke different emotions in different poeple. This leads to a poor correlation in the dataset and reduces the accuracy.
- The source from which the dataset was obtained i.e Flickr was not reliable. A better approach would be to use a crowd sourcing platform like Amazon Mechanical Turk and labelling the images based on the majority vote.

7 Future Work

- The emotion that a frame in a video depicts depends on what happened in the previous frames. This information is not made use of in our model. So in future work Seq2Seq model can be used to pass the parameters of the previous frames to predict the emotion of the current frame.
- It is not just the images in a video that influences the emotions of the viewers but a combination of audio and video. So we can use an ensemble of models that predict the emotions using the audio and the video to predict the final emotion.
- Our model classifies the general scene rather than the facial expressions of the poeple present in it. So if a person is crying in front of a beautiful background, it could be classified as happy. Therefore, we can use an ensemble of a model that predicts the emotion using the facial expressions of people (if present) and a model that classifies the general scene.

References

[1] Can Xu and Suleyman Cetintas and Kuang-Chih Lee and Li-Jia Li. "Visual Sentiment Prediction with Deep Convolutional Neural Networks". *arXiv preprint arXiv:1411.5731*, 2014.

[2] Quanzeng You and Jiebo Luo and Hailin Jin and Jianchao Yang. "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks". *arXiv preprint arXiv:1509.06041*, 2015.

[3] Victor Campos and Brendan Jou and Xavier Giró-i-Nieto. "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction". *Elsevier*, 2016.

[4] Soujanya Poria and Erik Cambria and Newton Howard and Guang-Bin Huang and Amir Hussain. "Fusing audio, visual and textual clues for sentiment analysis from multimodal content". *Elsevier*, (50-59), 2016.

[5] dataset. `https://github.com/philadams/flickr-images-grab`.

(a) Love

(b) Violence

(c) Fear

(d) Happiness

(e) Sadness

Figure 2: Some examples of true positives

(a) Fear (sadness)

(b) Happy (sadness)

(c) Love (sadness)

(d) Sadness (Fear)

(e) Violence (Happiness)

Figure 3: Some examples where our model misclassifies. Actual class is shown in brackets.