Violent Image Classification

A B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Ashutosh Mishra
(1601CS06)

under the guidance of

Abyayananda Maiti

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY PATNA
PATNA - 800013, BIHAR

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Violent Image Classification**" is a bonafide work of **Ashutosh Mishra** (**Roll No. 1601CS06**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Patna under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Abyayananda Maiti**

Assistant/Associate Professor,

May, 2020

Department of Computer Science & Engineering,

Patna.

Indian Institute of Technology Patna, Bihar.

Acknowledgements

It is injustice to write just a page of acknowledgements for all the people who have made it possible for me to write this thesis, but these are the only means available to show my eternal gratitude. First and foremost I would like to thank my supervisor Dr. Abyayananda Maiti for his thorough guidance, support, encouragement and counsel throughout the project. Without his invaluable advice and assistance it would not have been possible for me to complete this thesis. I would also like to give my most sincere thanks to Miss. Jyoti (PhD scholar) , IIT Patna for her invaluable guidance and advice on the topic. I am also grateful to Mr. Sanket Patil, a B.Tech scholar for his contribution in this project as a project partner. I am also indebted to all my family members, my friends and my well wishers for their inspiration and constant encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The need to identify violent and potentially disturbing images is at an all time high in the present scenario. With web applications like Facebook, Instagram, Whatsapp, etc. it has become very easy for anyone to upload and share any image they want. The content of these images can be anything. From pictures of pets and food to pictures of violence and gore, there is no limit to what a user can share. This project aims to develop a model to identify images which have violence in them, as well as the capability to incite violence in other people.

1.1 Background

A violent image can be defined as any image that has any of the following characteristics:

- Physical violence - fighting
- Blood, injuries, etc
- Hints of violence - pictures of riots, angry mobs, etc
- Images enabling violence - terrorist signs, nazi symbols, radical propoganda, etc

The purpose of developing such an automated detection system is to protect the general public from seeing such images without warning and to stop the intentional spread of such images by extremist groups to gain popularity. Another adverse effect these images could have is a civil unrest between different political, social and economic groups within society.

1.2 Categories

The images classified are branched into two main categories: Violent and Non Violent. The violent category is further divided into:

1. Radical violence
    - Riots because of religion or difference thereof
    - Signs, boards, other such objects containing inciting content
    - Physically violent images linked to radicalism
2. Non radical violence
    - The "normal" form of violence
    - Violence without any political motivation, clear of any radical ideologies
    - Physically violent images containing blood, gore, etc
3. Violent extremism
    - Extremist groups using violence to further their propaganda
    - Images of terrorist groups, beheadings, etc. Examples: ISIS, Al-Qaeda
4. Non violent extremism
    - Similar to violent extremism, without the physical violence
    - Signs and symbols supporting the extremists. Examples: Nazi symbols, logos of terrorist groups, propaganda posters, etc

1.3 Dataset

The dataset has been prepared by the **CSE** department of **IIT Patna** and consists of 8089 training images, 809 validation images and 1214 test images. The images are RGB, and each image has been resized to 224x224, per pixel mean subtracted.

Class	Number of images
Training	8089
Validation	809
Test	1214

Table 1.1 Image split

1.4 Overview

This report contains analysis and results of the work done on the following **problem statement:**

Given an image, classify it into one of the five categories:

- Radical violence
- Non radical violence
- Violent extremism
- Non violent extremism
- None

The following sections will describe related works, initial approaches, better models, thought process behind the methods, analysis and results.

Chapter 2

Review of Prior Works

2.1 Related publications

- **ImageNet Classification with Deep Convolutional Neural Networks - Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. [AK12]**

  The classifier in this paper was built for image classification on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2010) dataset. This paper served as a starting point for the classification task. As described in section 3.1, ReLU non linearity has certain advantages and hence was used in this project too. The paper describes a basic model consisting of 5 convolutional layers, two dense layers and one output layer. We modified this model by changing the filter sizes and numbers to better suit our needs.

  Figure **2.1** shows the first baseline model.

- **Visualising and Understanding Convolutional Networks - Matthew D. Zeiler and Rob Fergus [DF14]**

  Our second architecture was derived from a model presented in this paper. We again reduced the filter size from 11x11 to 7x7 and stride from 4 to 2 thus performing better. However, the authors trained the model on two GPUs, and we trained on

only one. Each RGB image was preprocessed by resizing the smallest dimension to 256, cropping the center 256x256 region, subtracting the per-pixel mean (across all images) and then using 10 different sub-crops of size 224x224 (corners + center with(out) horizontal flips)

Figure **2.2** shows the second baseline model.

- **Deep Residual Learning for Image Recognition - Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun [HZRS16]**

  This paper explains the advantages of using a Residual Network, or a ResNet to solve the problem of overfitting and vanishing gradients. We used a modification of ResNet50 as one of the models.

  Figure **5.2** shows the ResNet50 model.

- **ResnetCrowd: A Residual Deep Learning Architecture for Crowd Counting, Violent Behaviour Detection and Crowd Density Level Classification - Mark Marsden, Kevin McGuinness, Suzanne Little, Noel E. O'Connor [MMLO17]**

  This paper provided useful insight and paradigms for violence detection. The idea of making a high level decision tree using individual detectors was inspired from this paper.

2.2 Conclusion

We studied several papers related to image classification, residual networks, and convolutional networks. We implemented and tested a few models which were based on these papers.

Fig. 2.1 Model based on the first paper

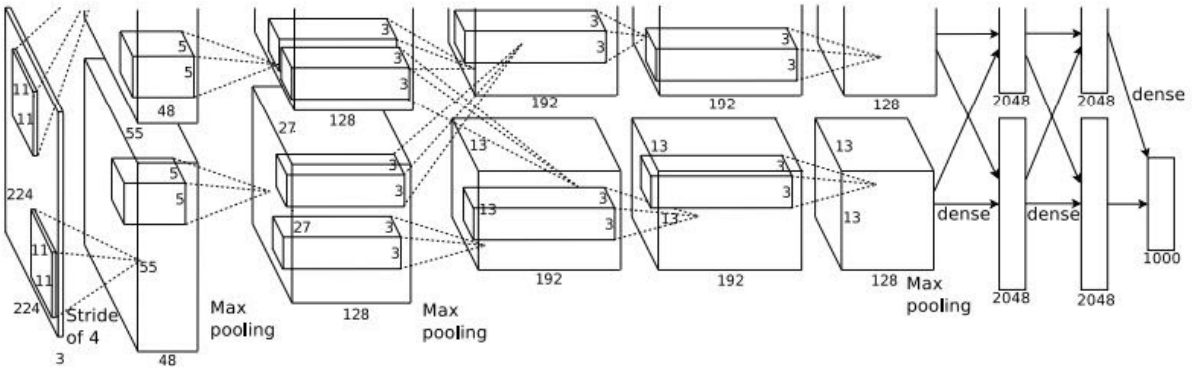


Fig. 2.2 Model based on the second paper

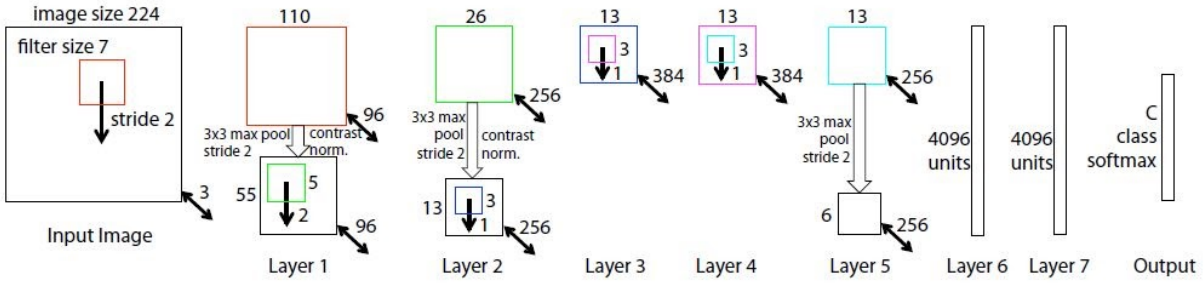
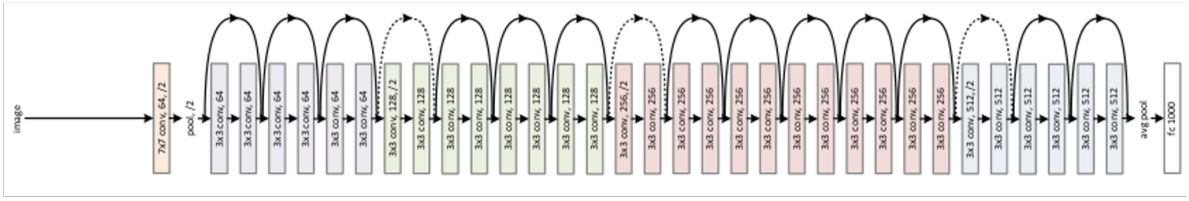


Fig. 2.3 Model based on the ResNet50



Chapter 3

The first model

The first model is based on Figure **2.1** [AK12]. There are 5 convolutional blocks, each followed by a maxpool layer. There are two fully connected layers at the end followed by a softmax layer which outputs the final classification. The specific layer by layer breakdown can be found in table **3.1**. All the convolution layers have "same" padding.

3.1 Results

The results obtained are tabulated in **table 4.2**. Performance of the model has been measured in terms of accuracy.

	Layer content
Input layer	-
conv1a	64 filters, 3x3
conv1b	64 filters, 3x3
maxpool	filter size: 2x2
conv2a	128 filters, 3x3
conv2b	128 filters, 3x3
maxpool	filter size: 2x2
conv3a	256 filters, 3x3
conv3b	256 filters, 3x3
maxpool	filter size: 2x2
conv4a	512 filters, 3x3
conv4b	512 filters, 3x3
maxpool	filter size: 2x2
conv5a	512 filters, 3x3
conv5b	512 filters, 3x3
maxpool	filter size: 2x2
FC	size: 1024
FC	size: 1024
Softmax	5-way

Table 3.1 Layers of the first model

	Loss	**Accuracy**
Training	0.6154	0.7689
Validation	0.7333	0.7188

Table 3.2 Model 1 performance

Chapter 4

The second model

The second model is based on Figure **2.2** [DF14]. There are 5 convolutional blocks. The first two layers are followed by a maxpool layer. The next three layers are directly connected. The fifth layer is followed by maxpooling. There are two fully connected layers at the end followed by a softmax layer which outputs the final classification. The specific layer by layer breakdown can be found in table **6.1**. All the convolution layers have "same" padding.

	Layer content
Input layer	-
conv1a	96 filters, 7x7, stride=2
maxpool	filter size: 3x3, stride=2
conv2a	256 filters, 5x5, stride=2
maxpool	filter size: 3x3, stride=2
conv3a	384 filters, 3x3, stride=1
conv4a	384 filters, 3x3, stride=1
conv5a	256 filters, 3x3, stride=1
maxpool	filter size: 3x3, stride=2
FC	size: 4096
FC	size: 4096
Softmax	5-way

Table 4.1 Layers of the second model

4.1 Results

The results obtained are tabulated in **table 4.2**. Performance of the model has been measured in terms of accuracy.

	Loss	**Accuracy**
Training	0.5066	0.8125
Validation	0.7337	0.7370

Table 4.2 Model 2 performance

Chapter 5

The third model

The third model was not based on purely classifying images using neural nets. Instead, we tried to approach the problem in a different way. After thoroughly going through our dataset we found out that almost all violent images have one or more of the following:

- Fire
- Blood
- Weapons
- Armed personnel(police, military, terrorist)
- Signs, posters, flags

We implemented detectors for each of the above attributes. The main idea was as shown in **figure 5.2**

5.1 Weapon detection

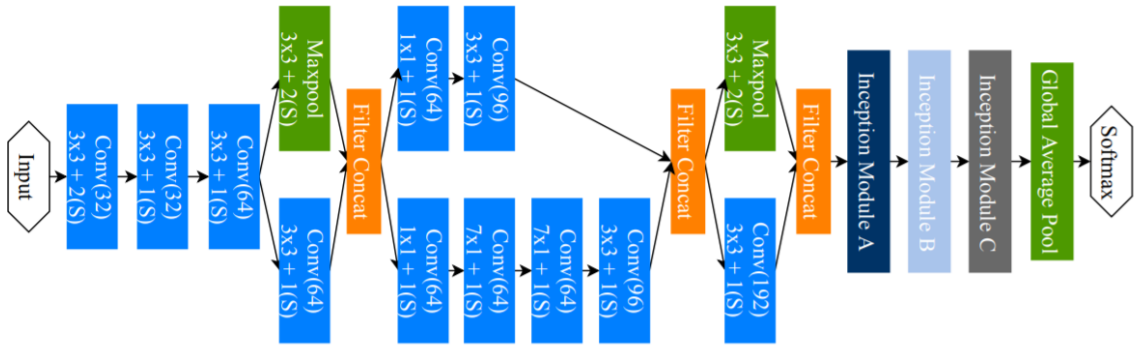
Weapons can be one of the most important visible feature of a violent scenes. Presence of weapon can directly indicate towards an violent scene. The aim is to identify and classify weapons in a robust way so that we can predict the presence of violence in a given frame.

We will be using faster R-CNN to achieve this goal of classifying weapon. Tensorflow and Google object detection API will be used as a platform to run this model for object detection.

5.2 Fire detection

We used Google's InceptionV4-OnFire architecture to build a model for detecting fires. The architecture is shown in **figure 5.1**

Fig. 5.1 Architecture of InceptionV4-OnFire network



5.3 The tree

First condition: Image contains visual violence

If the image contains any one of the following, the condition is satisfied: blood, fire, weapons.

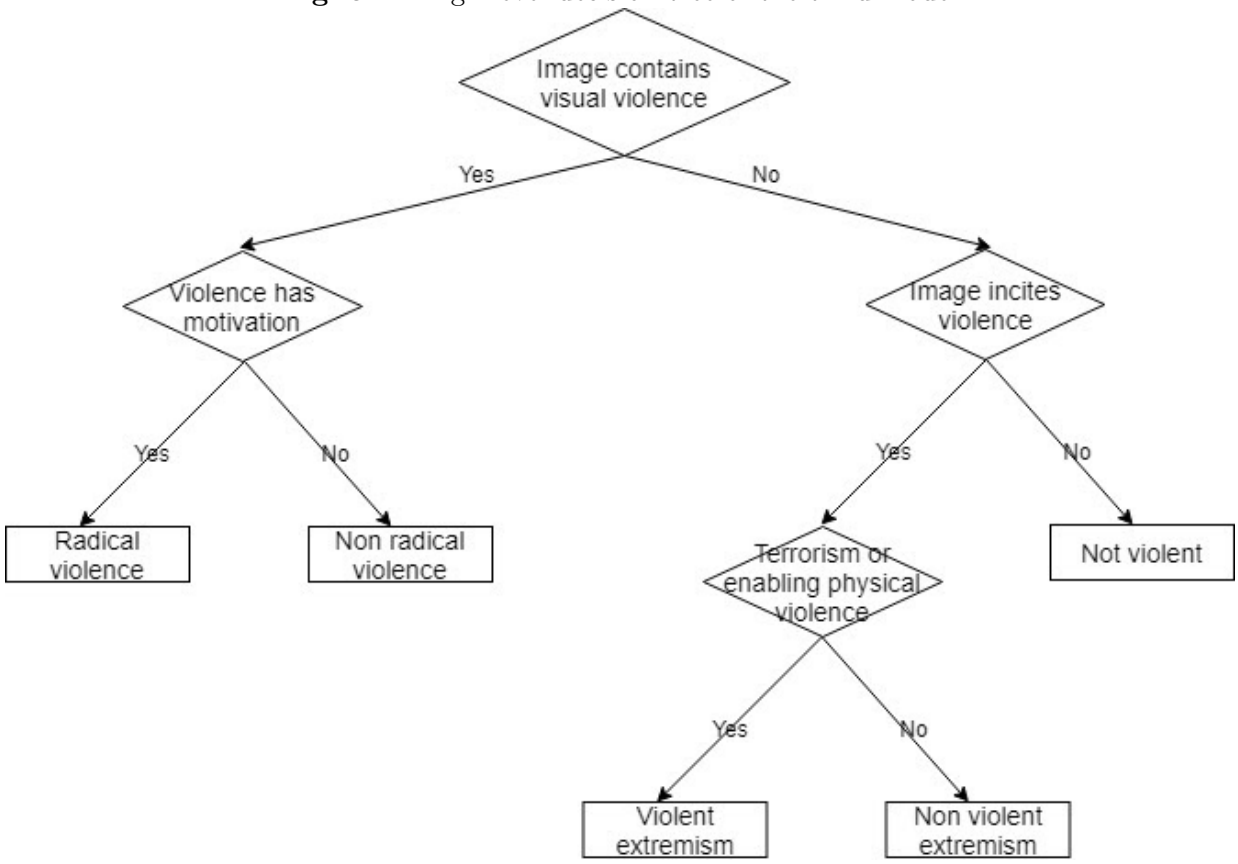
Second condition: Violence has motivation

If the image contains any one of the following, the condition is satisfied: signs, posters, flags.

Third condition: Image incites violence

If the image contains any one of the following, the condition is satisfied: propaganda, terrorists, flags or logos or extremist groups.

Fig. 5.2 High level decision tree of the third model



Fourth condition: Content

If the image contains any one of the following, the condition is satisfied: terrorism, enabling physical violence.

The fire and weapons detectors were easy to train, they had online datasets available as well as readymade models.

However, the trouble was with the blood and police/terrorist detectors. The blood detector would incorrectly classify images with high amounts of red and dark red as blood. The police/terrorist detector marked everyone with a weapon as either terrorist or police with nearly 50-50 accuracy.

Since the detectors themselves did not function as expected, this model was quickly discarded.

Chapter 6

The final model

The final model which we tried was the deepest one yet. We took the ResNet50 architecture [HZRS16], removed the top layer, added two densely connected layers (size: 4096) followed by a 5 - way softmax function. This model was by far the slowest to train.

6.1 Results

The results obtained are tabulated in **table 6.1**. Performance of the model has been measured in terms of accuracy.

	Loss	**Accuracy**
Training	0.4122	0.8856
Validation	0.8005	0.6850

Table 6.1 Model 4 performance

We find that although training accuracy has improved significantly, validation accuracy has decreased. Although ResNets usually solve the problem of overfitting, but it seems like in this case the model is too complex and the dataset is too shallow. Hence the increase in training performance but a loss in validation performance.

This indicated that we need to either expand our dataset, or reduce the complexity of our model by decreasing a few layers.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This project has aimed to implement models which can aid in classifying images into violent and non violent categories.

The results of the models implemented are as follows:

Model	Validation Accuracy
VGG16 based	0.7188
Second model	0.7370
ResNet based	0.6850

Table 7.1 Model accuracies

The third model was not fully tested on the entire dataset, as the individual detectors themselves were not classifying properly.

We can see that a medium complexity model performs the best. The first model also has a good accuracy measure, albeit being a simple model. The second model is a bit more complex than the first one, and it shows in the accuracy. The third ResNet based model surprisingly performs the worst. Our hypothesis is that the model is too deep and the dataset is too shallow; causing overfitting *even though we are using a resnet.*

7.2 Future work

7.2.1 ResnetCrowd

A good approach for estimating crowd sentiments would be to incorporate the ResNetCrowd model - it has multi task learning which optimizes the model even further. The transfer learning capability of ResnetCrowd can be used to make more robust models.

7.2.2 Other models

A few other promising models include Densenet and Darknet-19. However, before using any more deep image classification models, I would strongly recommend expanding the dataset, which brings me to the last point.

7.2.3 Dataset

Expanding the dataset. I believe this is a crucial point to be addresses before moving further. To capture some abstract concept of *violence*, we need more than just image and object detection mechanisms. We need deep networks which capture higher dimensional features related to violence. To train these networks effectively, we need a much bigger dataset.

References

[AK12] Geoffrey E. Hinton Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 10, 2012.

[DF14] Matthew D.Zeiler and Rob Fergus. Visualising and understanding convolutional networks. *ECCV 2014: Computer Vision – ECCV 2014*, 12, 2014.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[MMLO17] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E. O'Connor. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. *CoRR*, abs/1705.10698, 2017.