EXTREMISM VIDEO DETECTION IN SOCIAL MEDIA

An Undergraduate Research Scholars Thesis

by

AKASH RAO

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. James Caverlee

May 2021

Majors: Computer Science
Mathematics

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Akash Rao, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Research Faculty Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

ABSTRACT

Extremism Video Detector in Social Media

Akash Rao
Department of Computer Science and Engineering
Department of Mathematics
Texas A&M University

Research Faculty Advisor: Dr. James Caverlee
Department of Computer Science and Engineering
Texas A&M University

Social media has grown to become a fundamental part of our lives over the past two decades and with its growth, the misuse of the platform for extremist purposes has become common. The wide reach of social media has allowed extremist groups to take advantage of the platform to spread terrorist propaganda and fear. Therefore, the need for a robust extremist detector in social media is evident.

As an attempt to combat this problem, we present techniques to detect various forms of extremism in videos crawled from *Twitter*, a social media to share short posts. We build upon existing deep neural networks used for action classification and create a model capable of recognizing certain common extremism types. Additionally, we also expand on logo/object detection models for the same purpose. We then use these models against a sample space of roughly 2 million unlabelled videos to test the accuracy of these models.

DEDICATION

To my family, instructors, and friends who supported me throughout the research process.

ACKNOWLEDGMENTS

I would like to thank my faculty advisor, Dr. James Caverlee, and Majid Alfifi for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

The data set, namely videos, used for testing/training in Extremism Video Detector on Social Media were provided by Majid Alfifi and Infolab.

All other work conducted for the thesis was completed by the student independently.

1. INTRODUCTION

As social media grows as a platform, it has become the first destination for many to access the news, share thoughts about any matter, keep up with one's country's political situation etc. For example, 50%-60% of adults in the US say they use popular social media platforms like YouTube, Twitter or Facebook at least once a day with about 18% of US adults relying on social media as their primary source of news [1, 2]. This widespread use allows influencers from around the world to spread their messages to anyone with access to the internet. However, with its spread and utility also comes shortcomings such as abuse of the platform for terrorist propaganda, spread of extremism and even insinuation of violence. For example, YouTube, a popular video sharing platform, removed roughly 170,000 videos were removed for hate speech or violent acts promoted by foreign terrorist organizations in Q4 2020 out of approximately 65 million minutes of content uploaded during the same time [3, 4]. Therefore, an automatic system of detecting such content before it receives too much attention is of great necessity.

In order to flag such type of content, social media companies employ three main techniques. First is the usage of automatic checking systems that check for uploaded content against other content that has been flagged and basic sanity check against the text and other meta data uploaded as part of a social media post. Secondly, companies employ experts from around the world to manually check certain content to ensure policy guidelines in their respective platforms. And finally, they use a system that allows users to report content that they think violates certain guidelines, through which content can be subsequently taken down.

There is little public knowledge of the specific methods that guide these three main approaches, however, in literature automated methods have received the most attention with an emphasis on inference from textual content. This may be in the form of tweets from Twitter, blogs from Tumblr, or video meta data such as title, captions and comments from YouTube. However, we believe that there is valuable information to be gained from the visual elements of social media

posts such as images and videos. Take for example the images presented in **Figure 1.1**, we can see that this type of imagery containing an armed man, waving the ISIS flag and walking down the street is clearly extremist and demonstrating radical actions. If this post were accompanied by text describing the situation, flagging it might be relatively easy, however, even without any accompanying text we can notice patterns that would immediately strike this content as suspicious.

Figure 1.1: Example of clearly extremist video

Based on this reasoning, we propose an extremism video detector that takes into account specifically visual input in the form of short video clips to identify unseen videos as possibly extremist. We first expand on the work in the field of action recognition and hope that extremism can be incorporated as a comparable class to the set of action classes commonly considered. Next, we approach the problem with in a more ad hoc manner and hope to exploit logo/symbol detection to extract extremist videos from our large unlabelled dataset of short video clips. With these two approaches we hope to create a robust classification model that is capable to detecting extremist/radical content avoiding some of the pitfalls mentioned below.

In regards to our approach of creating a video extremism/radicalization detector we anticipate and face the following challenges,

1. *Lack of a standard dataset*. Unfortunately, there does not exist a commonly used dataset of extremist/radical videos upon which to train/evaluate our model and more importantly compare it to other models designed to do the same. However, we will address this problem in the dataset section.

2. *Class imbalance*. In a production setting, there could be only a fraction of a percentage of videos that could potentially contain traces of extremism or radical content. And this might even vary from geographic location, language, time of day etc. This means that our model should not be too eager to classify lots of videos as suspicious, but at the same time not avoid classifying any videos as extremist in order to achieve a higher accuracy.

3. *Computational efficiency*. Although image classification has become quite efficient in modern times, video classification adds another dimension to this problem, and hence can require additional resources and time to build more sophisticated models. We need to keep this in mind when building our classifier.

Our contributions discussed in the remainder of this thesis will mainly include the discussion of a semi-novel technique to the heavily studied field of predicting and identifying extremism from social media content. Additionally, we also propose, implement and compare two different approaches of using short video clips that we crawled from Twitter in order to train a model and test its classification performance on unlabeled data. We will also briefly mention a few more approaches that have not been implemented but that we see potential in in performing the task that we are tackling. Finally, we introduce a new dataset that could potentially serve as a standard video dataset wherein further research on video extremism detection could be based on.

2. RELATED WORK

The area of utilising social media content in order to either classify future content as extremist or potentially radical, or perform extremist activity forecasting has been widely studied. Moreover, the area of computer vision, specifically action recognition and video classification is a very hotly studied research area as well. But, we feel that our research lies within the intersection of these two fields but is heavily inspired by some of the works from this domain. In this section, we will perform a literature review of our best knowledge assessing different methodologies and techniques employed by these papers.

2.1 Online Radicalization Prediction and Identification using Social Media

As social media has expanded to become a crucial interpersonal communication medium, terrorist organisations have taken advantage of the system for purposes such as recruitment of new members, spreading of their ideologies and messages, displaying violent and extremist content to emphasise their mission and even provide early warning to terrorist attacks on countries. Therefore, the use of algorithms and artificial intelligence to use these patterns to prevent such activities on social media has been widely studied. For example, [5] uses a classic K-nearest neighbours (KNN) and support vector machine (SVM) in order to classify tweets as belonging to a class of hate promoting tweets or unknown class. More concretely, tweets are described by a feature vector $\{f_1(I), \dots f_m(I)\}$ where $f_i(I)$ is the i'th instance value for a tweet, I, with a total of m discriminatory features and the specific distance function used is an l_2 norm or Euclidean norm. Even in studies that consider visual content such as [6]'s attempt to identify extremist content in video sharing sites, attempts are made to utilise video meta data such as video title, description and comments followed by classification through SVMs. In a literature review of different studies researching the use of social media posts for online radicalization identification we can see this heavy reliance of textual data from **Figure 2.1** [7]. Specifically, we see a greater number of papers relying on social media content like tweets, blogs and meta data and applying corresponding classification tech-

niques that have been used traditionally for such textual data. Techniques such as these that rely heavily on textual information can pose certain problems. For instance, variation in the types of textual data differing in writing style, grammar errors, misspelt words or even language variation among data can introduce noise into the model and increase the complexity of the problem from a practical standpoint.

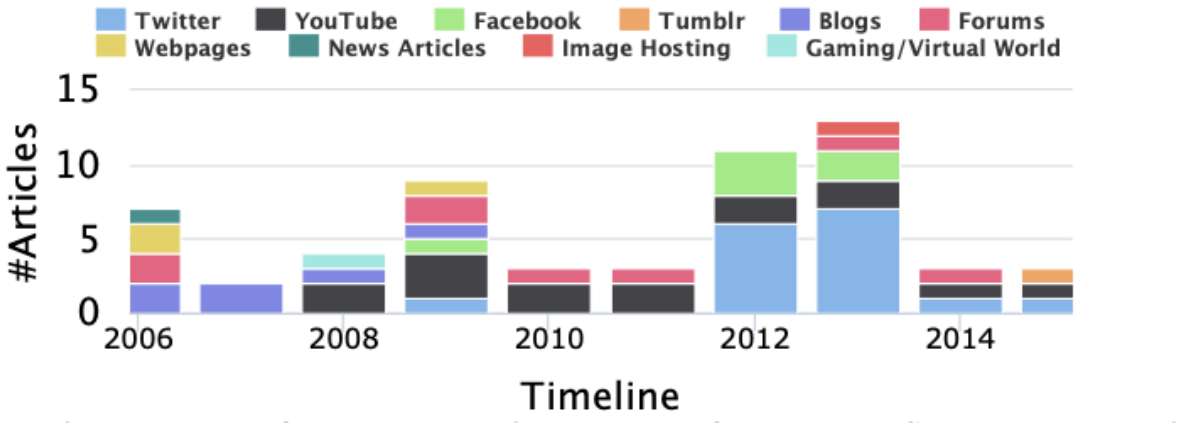

Figure 2.1: Number of publications over time and the datasource for online radicalization detection

Although these problems need to be overcome in the analysis of text-only social media posts, we believe that for posts that are accompanied with visual aids such as videos or pictures, such content can be more universal and allow us to neglect problems associated with natural language.

2.2 Video classification and Action Recognition

The first real large scale implementation of video classification mirroring the successes of image recognition arguable came from Karpathy et. al in 2014. In their paper, they show that convolutional neural networks (CNNs) and their variations are capable of reproducing accuracy levels seen in image/object recognition tasks but in action recognition [8]. Specifically they use a dynamic dataset of 1 million YouTube videos, *Sports-1M*, of sports clips with 487 classes and achieve approximately 80% accuracy in top 5 video category. More importantly and relevant to

our research is that they were able to achieve greater on smaller datasets by employing transfer learning and retraining/ fine-tuning only top-k layers of an already pretrained model. The central idea here is the pretrained models are quite capable of capturing generic features that are domain-independent and reusing this learning has been proved to be beneficial. More recently, Hara et. al. have shown great advances in action recognition using very deep CNN architectures with 3D CNN kernels as opposed to just 2D kernels as seen in image recognition tasks. Moreover, they use residual neural networks (ResNets for short) in order to create bypass connections in the network and facilitate training of very deep models that are capable of capturing very high level features [9]. For example, their models acheive a 84.4% Top-5 accuracy on the extremely large Kinetics dataset with over 700 classes and 650,000+ videos clips. In fact, we hope to expand on their work, which is open source [1] , and apply transfer learning on their 156 layer deep model.

[1] https://github.com/kenshohara/3D-ResNets-PyTorch

3. METHODS

3.1 Phase I: Classification into Extremist Sub-classes

In Phase I of experimentation, our problem very much resembles that of traditional action classification in videos, except we are now considering labels such as "Dead Person", "Bombing" etc. instead of "Applying Lipstick" or "Push Ups". We call the set of these labels $Labels_{extremism}$. If we consider a specific video, v, and determine a confidence score array from our model, $\{s_1, s_2, \ldots s_n\}$ where s_i corresponds to how confident our model is that v belongs to the i'th subclass in $Labels_{extremism}$. The main idea behind separating videos into different sub-classes of extremism is to further help the model understand the intricacies of each type of extremist video. If instead, we were to simply consider the set of labels extremism and unknown, there would a wide range of videos falling into each category making it harder to narrow down labelling for a specific unseen video.

In **Figure 3.1**, we can see a visual depiction of the Hara et. al. 156 layer deep ResNet model that we have employed to perform classification into extremist sub classes. This model was trained on the Kinetics dataset and hence predicts confidence scores for 700 actions that the input video could contain. In **Figure 3.2**, we see how this model is now fine tuned by adjusting the weights of the last few layers (represented by the orange in the figure), therefore, transferring knowledge from the Kinetics domain, but then redesigning the model to predict classes that belong to $Labels_{extremism}$.

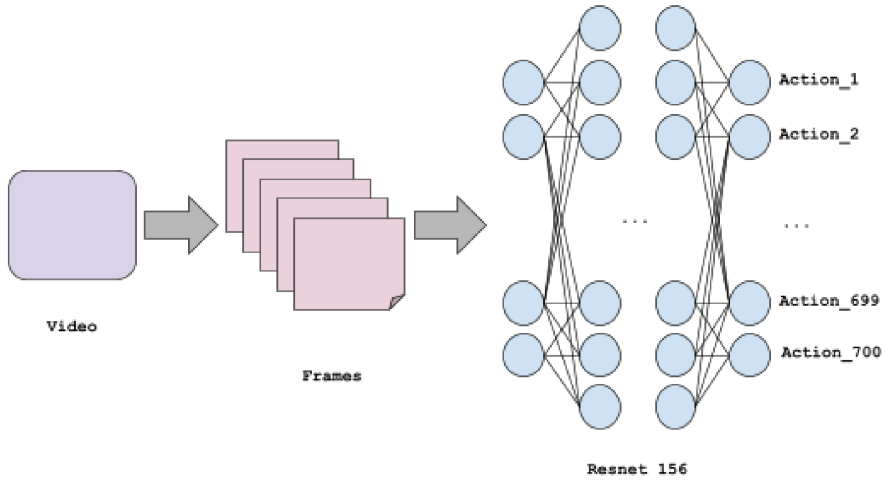

Figure 3.1: Visual Depiction of Hara et. al. Model

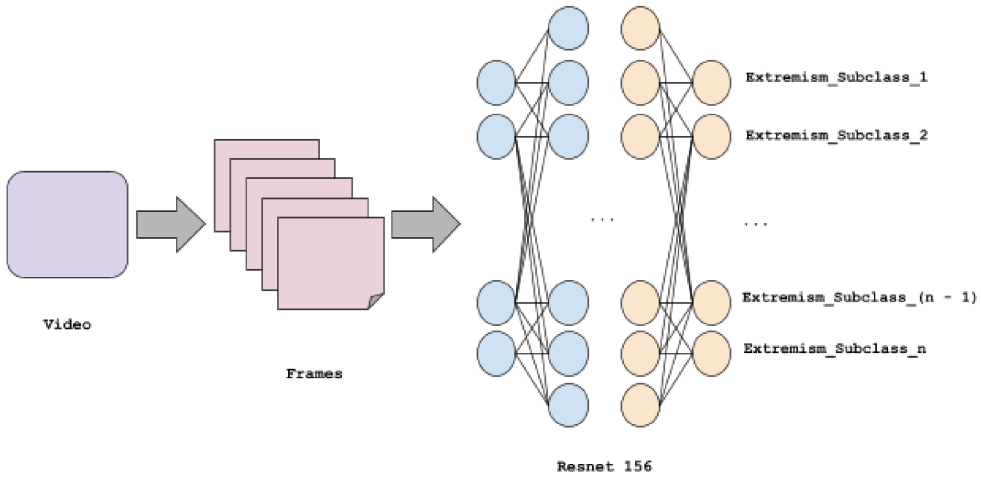

Figure 3.2: Visual Depiction of Fine Tuning Last Layers

3.2 Phase II: Classification based on Iconography

In this second phase, we attempt to use logo/object detection in order to classify videos as extremist. This idea was spawn about because a large portion of our labelled dataset contains ISIS videos, which contain a lot of iconography and symbols. We hope to use these patterns in order to identify more ISIS videos from our unlabelled dataset. For example, in **Figure 3.3**, we see the

ISIS flag waving in the top right hand corner of the video clip. This can be seen on several of their videos, albeit in different locations and irregularly throughout the videos.

Figure 3.3: Typical location of iconography in videos

Figure 3.4: Examples of some more ISIS-related iconography (ISIS Flag, Man Making and Al Hayat)

In **Figure 3.4**, we can see examples of a few more instances of the ISIS flag and also additional symbols that also occurs with regular frequency.

We pose this as an object detection problem, where the objects are namely the symbols in the

videos. Suppose we define our list of icons as $Icons = \{i_1, i_2, \ldots i_n\}$, then our model takes a single frame, f, from a video and outputs a confidence score $\{s_1, s_2, \ldots, s_n\}$ where s_j corresponds to how confident our model is that icon i_j is detected in the image. For this purpose we will be using a Single Shot Multi Box Detector (SSD) network architecture that has been quite popular in this domain [10]. We mainly choose this model over other types of architectures such as You Only Look Once (YOLO) and variations of R-CNNs because of their great performance in terms of accuracy as well as speed at which they can perform object detection [10, 11]. More specifically, on the PASCAL Visual Object Classes Challenge 2007 with classes like person, bird, bottle, chair etc. SSD architecture models attained 79.3% accuracy whereas YOLO and Faster R-CNN architectures attained 63.4% and 73.2% respectively [10, 12].

We will be extending off of the work[1] from Machine Learning Engineer, Hugo Zanini who created an end-to-end object detection solution using Tensorflow. This application was based on a model built to detect kangaroos, but we modify it to instead detect ISIS flags, since this particular symbol occurs with greatest frequency and is clearly distinct from the rest of the symbols which are all variations of Arabic writing could potential result in inaccurate classifications when we are using, say nonextremist content that also happens to include Arabic text in the video.

[1] https://blog.tensorflow.org/2021/01/custom-object-detection-in-browser.html

4. EXPERIMENTAL SETUP

In this section, we first expand on the dataset used for experiments along with their limitations. Next, we describe the different approaches implemented to achieve our objective of detecting whether videos belong to a specific subclass of extremism. We will also portray the reasoning behind the choices made along the way as well as the results of each experiment.

4.1 Dataset

Since there is no commonly used dataset for extremism video classification we use a custom dataset crawled from Twitter restricted to the Middle Eastern region. In total, we have 2,840,588 videos of varying length, subject material, aspect ratio and resolution. Of the roughly 2.8 million videos, we have manually identified 56 videos containing different types of extremism particularly graphic scenes and terrorist propaganda demonstrated by the organization ISIS. For some context, in **Figure 4.1** we see examples of non extremist videos acquired from our dataset and in **Figure 4.2** we see examples of some non-graphic extremist videos.

In order to facilitate classification we further narrowed down these videos into sub-classes of extremism that are summarized in **Table 4.1**.

Table 4.1: Sub-classes of Extremism labelled in Dataset

Extremism sub-class	**Video count**
Air force Bombing	5
Bombing	14
Dead Person	6
ISIS Flag	6
Militia Crossfire	10
Militia Marching	6
Total	47

We observe that there are only a limited number of videos in each class relative to other standard

Figure 4.1: Screen captures of non-extremist videos from dataset

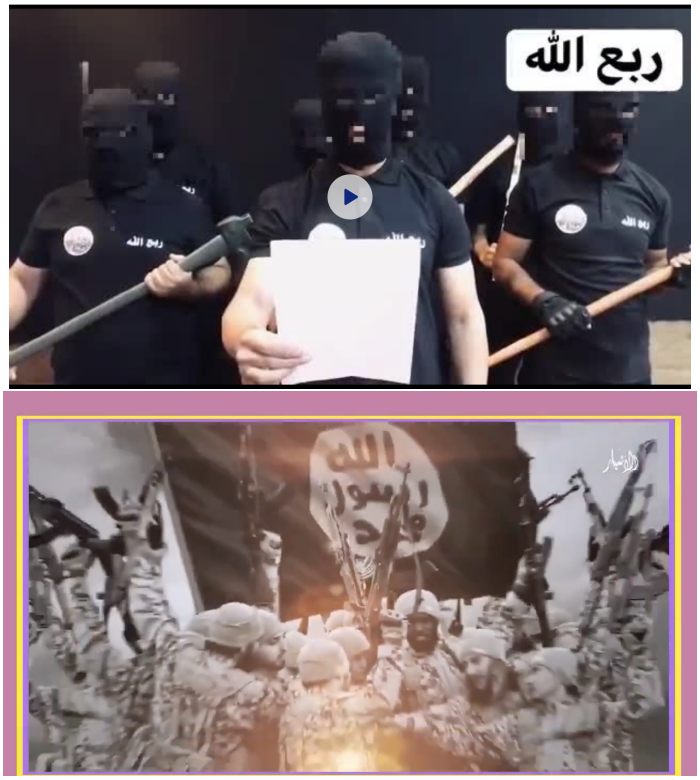

Figure 4.2: Screen captures of extremist (non graphic) videos from dataset

video classification datasets such as UCF-101 with > 100 videos in each category and Kinetics-700 with $600 - 1150$ videos in each of the 700 classes. However, we hope to exploit [9]'s *ResNet* model's fine tuning capacity which has been proven to transfer training from one domain to another, especially with such limited training data. Transfer learning has been quite beneficial, since pretrained models have shown promise due to their ability to learn generic features (like edges and shapes) that are common across multiple datasets and even generic actions like body-motion or human-object interaction etc. [8].

In order to faciliate objection detection in images, we split each of our videos into 10 evenly spaces frames and then manually provide bounding boxes for the different iconography present in our extremist videos. We also use the help of AWS Mechanical Turk to both hasten and validate this process by manual labelling from several workers. In **Table 4.2**, we have summarized the different logos and their corresponding number of occurances in the dataset.

Table 4.2: Iconography from manually labelled extremist video dataset

Iconography/Symbol name	**Total number of images containing symbol**
ISIS Flag	473
Al Hayat	95
Man Making	297
Total	865

4.2 Experiments

As previously mentioned our first phase of experimentation expands on the work by Hara et. al. whose work has been open sourced and available on GitHub. Our next phase of experiments involving ISIS iconography will be based on the work of another popular paper whose experimental setup has also been made public on GitHub. All the experiments were performed on `Intel 8-core i7 4820k 3.7 GHz` Linux machine with `65GB` physical memory and `Nvidia GeForce RTX 2080TI` with `11GB` of graphics memory.

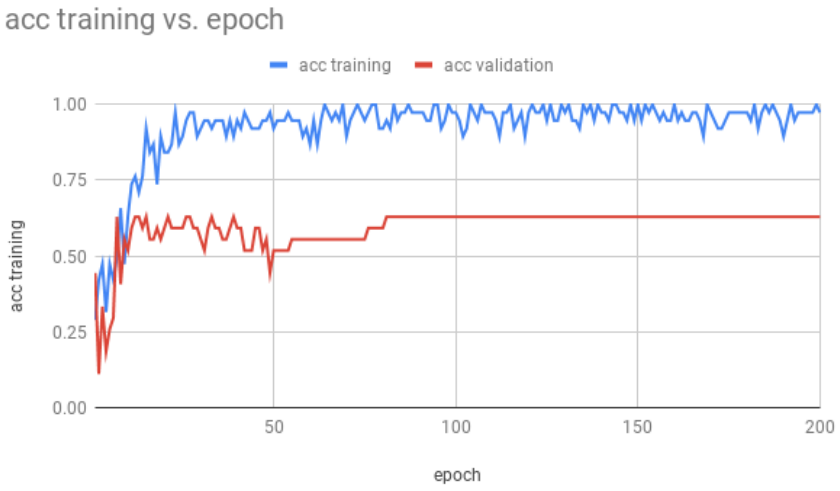

Figure 4.3: Testing and Validation Accuracy for Video Classification Model Based on Actions

4.2.1 Phase I: Classification into select sub-classes

4.2.1.1 Training

We perform training in two stages due to the lack of a large well labelled extremism to begin with, however, we hope that by using a pretrained model and then fine-tuning on a smaller subset will still allow for accurate classification with scores $\geq 0.85 - 0.90$.

In the first stage, we use the approximately 60 videos with 6 extremism classes to train our model. The resulting testing/validation accuracy is portrayed in **Figure 4.3**. As can be seen from the graph, despite a small number of videos in our training set, we are still able to achieve a validation accuracy of roughly $80 - 82\%$.

We then use the aforementioned model to run inference on roughly 500 thousand videos and manually judge the model's performance and filter out correctly classified videos from each extremism subclass for improving the model further. We repeat the same training procedure with a 80/20 split with the previously labelled dataset and the newly obtained dataset combined.

4.2.1.2 Testing

The testing stage involves running the model against the entire subset of unlabelled videos and accumulating the results. We perform testing in two main stages also - Transformation of Videos and Inference. Firstly, we convert all the videos into `jpeg` image frames collected at every one second interval. Next, we load the model and feed every video's frames to obtain the corresponding inference. Overall time taken to complete converting the videos to frames is about 16 hours with 10 jobs running, and inference by the model is also another 16 hours.

4.2.1.3 Evaluation

We perform evaluation of the model's performance by manual checking for correctness. We filter out videos with accuracy scores of > 0.95 (note that the sum of the scores among all classes is 1) and we check whether the video indeed falls into the predicted class.

However, even during the initial results we notice strong red flags indicating that our model is largely overestimating the number of extremist videos in our larger sample size. For example, of the 48,000 unlabelled videos fed into the model for inference roughly 10,000 videos were labelled as following into one of our extremist sub-classes with a confidence score of greater than 0.95. However, from some of the videos it is quite evident the reason for misclassification. For example, in several of the videos labelled as bombing we notice a large bright yellow spot like reflection on water or a zoomed in video of a yellow flower.

$$precision@k = \frac{|classified\_extremist \cap actually\_extremist|}{|actually\_extremist|} \qquad \text{(Eq. 1)}$$

4.2.2 Phase II: Classification based on iconography

4.2.2.1 Training

For training we start with a standard $80-20$ training and validation split of our $(frame, bounding\_box)$ pairs for the 473 manually labelled samples of ISIS Flags. We then train our SSD architecture Ten-

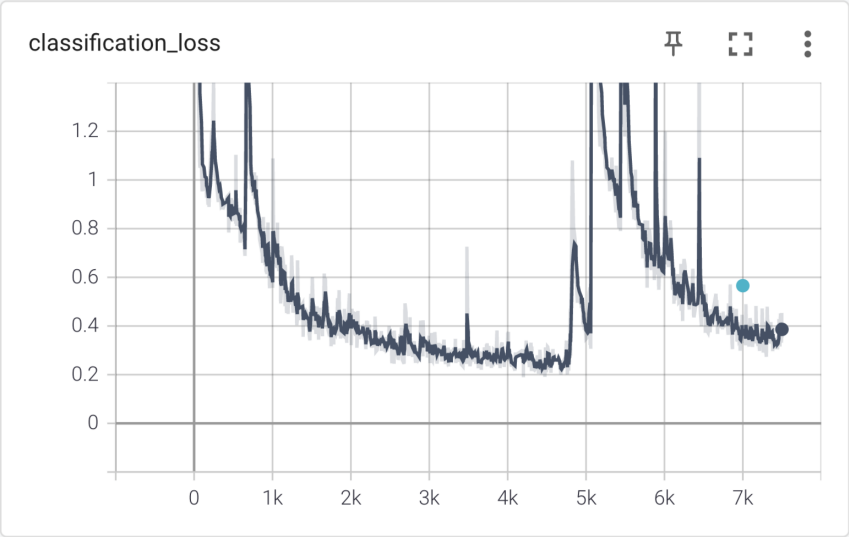

Figure 4.4: Training Loss for Icon Detection Model

sorflow model for 7500 steps with a running time of 3 hours with about 1.5 seconds per step. The training loss is provided in **Figure 4.4** and we can see that our model performs best after about 5000 steps.

4.2.2.2 Evaluation

Finally, we use as $precision@IoU$ and $recall@IoU$ to determine the relevance of the frames picked by our model as having extremist content and therefore judge the overall performance of the model. These metric allows us to especially determine how well the classification model does in this extremely imbalanced set of classes. A more formal definition of these techniques is presented below.

Suppose we have a predicted bounding box for an image $bbox_{pred}$ and a ground truth bounding box that we have manually labelled $bbox_{truth}$, then let $bbox_{pred} \cap bbox_{truth}$ be the area of overlap between these two boxes and $bbox_{pred} \cup bbox_{truth}$ be the total area of the union of these two boxes. Then we can define Intersection over Union as,

$$IoU = \frac{bbox_{pred} \cap bbox_{truth}}{bbox_{pred} \cup bbox_{truth}} \qquad \text{(Eq. 2)}$$

Figure 4.5: Predicted bounding boxes (left) and ground truth bounding box (right)

Therefore, $IoU = 1$ would be that our predicted bounding box exactly overlays the grouth truth and $IoU = 0$ means there is no overlap at all. Next, we define true positives as those validation images with $IoU > \delta$, false positives as those images with $IoU < \delta$ and false negatives as images with no predicted bounding box. With these definitions we have,

$$precision@[IoU = \delta] = \frac{|true\_positives@\delta|}{|true\_positives@\delta \cap false\_positives@\delta|} \quad \text{(Eq. 3)}$$

and,

$$recall@[IoU = \delta] = \frac{|true\_positives@\delta|}{|true\_positives@\delta \cap false\_negatives@\delta|} \quad \text{(Eq. 4)}$$

Below we have listed the evaluation metrics from the trained model,

```
Average Precision @[ IoU=0.50 ] = 0.480
Average Precision @[ IoU=0.75 ] = 0.088
Average Recall @[ IoU=0.50:0.95 ] = 0.355
```

As we can see the Average Precision, which is simply the mean of the precision over all the tested images, for $\delta = 0.50$ is 0.480. This roughly translates to half the images having good predicted bounding boxes. However, since in our detector the exact precision of the predicted bounding box compared to the ground truth bounding box doesn't matter and we are more interested in whether

the model even produces a bounding box for an image with any reasonable level of certainty (confidence level > 80). This is more clearly illustrated in the **Figure 4.5**. In this image we can see predicted boxes on the left do a reasonably good job in detecting ISIS flags and therefore videos with extremist content, but as per the evaluation metrics for object detection this would be considered poor.

Finally, we use this new model to run inference on our 40,000 unlabelled videos. The general setup involves splitting each video into 10 equally spaces frames which are then fed into the model to detect ISIS Flag's. Here we use a confidence threshold of 80% and for any images that are predicted to have an ISIS Flag over this threshold, the corresponding video is tagged as extremist. From this method we find that only 132 videos were tagged as extremist as compared to the 10,000 videos from the action recognition technique. Therefore, that is a reduction by almost hundredfold reducing the number of videos that need to be manually checked. However, based on our inspection all the 132 videos classified are false positives. Furthermore, we do not know the exact composition of extremist versus non-extremist videos from our unlabelled dataset, only the fact that there is a heavy class imbalance.

5. CONCLUSION

In this body of work, we have demonstrated the use of visual aids such as video clips in order to identify more videos in social media as potentially extremist or containing radical content. This technique allows us to bypass some of the challenges with only textual based extremism social media classifiers such as noise from variations in natural language and the general volatility of human language and written text.

Additionally we have shown that ResNet's employed in action recognition tasks from video clips can be used to make meaningful predictions on videos containing traces of extremism, especially deep ResNet's that are then fine-tuned to classify videos that are in domains quite different from those that were used to train the models. Next, we also engineer a more ad hoc approach to utilise symbols and flags, that are quite common across ISIS related extremism videos, to more accurately predict radical videos from a large subset of unlabelled videos. We have also experimentally compared the aforementioned two methods and see that the ad hoc method does indeed perform better when it comes to classification of extremist content, however, it comes with the downside of being organization specific and can easily overlook videos that do not contain any symbolic representations of the corresponding organizations.

Finally, we see potential in utilising auditory decision variables into our classifier similar to how we exploited symbols. These are clear indicators of videos from certain terrorist organizations and could prove to be very accurate.

REFERENCES

[1] B. Auxier and M. Anderson, "Social media use in 2021."

[2] M. Jurkowitz and A. Mitchell, "Americans who get news mostly through social media are least likely to follow coronavirus coverage."

[3] "YouTube community guidelines enforcement – violent extremism." https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en.

[4] "YouTube for press." https://blog.youtube/press/.

[5] S. Agarwal and A. Sureka, "Using KNN and SVM based one-class classifier for detecting online radicalization on twitter," pp. 431–442.

[6] T. Fu, C. Huang, and H. Chen, "Identification of extremist videos in online video sharing sites," in *2009 IEEE International Conference on Intelligence and Security Informatics*, pp. 179–181, 2009.

[7] S. Agarwal and A. Sureka, "Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats,"

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, IEEE.

[9] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet?," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, IEEE.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection,"

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector." https://github.com/weiliu89/caffe.