NLQ-Security: Surveillance Video Analysis with Natural Language

Brandon Contreras

May 2021

1 Abstract

Analysis of surveillance video is a time intensive and tedious task that involves watching hours of footage in order to find information on targets of interest. There are multiple tools in use that aim to solve this problem through video analytics. Natural Language Query-Security (NLQ-Security) is a natural language object retrieval tool that can search surveillance video for timestamp ranges that satisfy a given query describing an object of interest.

2 Introduction

The use of Closed Circuit TeleVision (CCTV) has rapidly grown ever since the first commercial system was introduced in the U.S in 1949 [4]. CCTV has been heavily deployed in public spaces from retail centres to medical facilities for a variety of different purposes [7]. One of the wider known use cases for CCTV is for security purposes such as identifying shoplifters or trespassers. Along with the growth of CCTV surveillance use has been the large amounts of video data that CCTV creates [5]. This footage must be combed through by a human operator in order to find relevant information. If an investigator is interested in the video footage of the past few days this can lead to hundreds of hours worth of footage that needs to be reviewed [5]. This task can prove daunting to a security team. It has been shown that a person can not stay focused for more than 20 minutes when conducting surveillance and monitoring tasks on security videos [7].

3 Background

The field of Video analytics (VA) has increased the utilization of CCTV video data greatly since it's inception. VA is the ability to automatically process video data to detect temporal and spatial events [2]. VA has transformed traditional video surveillance in the past two decades and is expected to increase the already billion dollar industry of surveillance solutions market [2]. Traditional video surveillance requires manpower which is both expensive and time consuming [5]. VA utilizes video surveillance data to automatically detect for a wide range of events depending on the industry and task at hand. Some VA tasks include license plate recognition, loitering detection, facial recognition, and counting people [5].

4 Data

Microsoft's Common Objects in Context (MS COCO) is a large-scale image data-set that contains features for object detection, segmentation, and image captioning. MS COCO aims to advance computer vision by providing an image data-set that contains annotated images of common objects in complex scenes [3]. The data-set contains features for identifying objects in their natural context as opposed to other image data-sets commonly used in computer vision which contain only labeled iconic images such as ImageNet [3]. RefCOCO is a data-set that contains a subset of the images found within MS COCO and adds referring expressions describing objects contained within the images. These referring expressions were captured from an interactive game interface and are 3.5 words on average. Approximately 41% of the referring expressions found in RefCOCO use the location of the object of interest to describe it [8]. By generating referring expressions from humans in a relaxed setting these captions are a good representation of how humans describe objects found in images.

Image data-sets such as Imagenet or Pattern Analysis Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC) contain labeled instances of iconic images whereas MS COCO contains instances of non-iconic images [3]. An image that is said to be a good representation of an object, scene, or place is described as a canonical representation. There have been various methods studied to define and measure iconicity. Iconicity is argued to be a relative property that is more subjective in nature than an absolute property. This is the basis for the definition of an "iconic" image. For example, an iconic image of a horse would be a picture in which a horse is centered in the image, is unobstructed from view, and takes up the majority of the image's area. On the other hand, an non-iconic image would be an image that is "noisy" i.e many subjects are present in the image and the subject of the non-iconic image might be partially obscured. An example of a non-iconic image for the concept of "bike rider" would be an image taken in Time Square where a bike rider is present in the scene but is obstructed from view and many other objects are present [9]

Extensive efforts have been made to ensure that the majority of the images found in MS COCO are non-iconic in nature. Images from the amateur photography site Flickr were collected which contains fewer iconic images than from other sources such as a Google Image search. Researchers utilized pairwise combinations of object categories in order to find images that were non-iconic when collecting images to represent the 80 object classes of the MS COCO data-set . These pairwise combinations would yield searches such as "dog + car" when searching for images to represent the "dog" class.[3]

The RefCOCO data-set was chosen over other image data-sets for the fact that it contained a majority of non-iconic images and more importantly contained caption descriptions for objects within the image. VA for security purposes primarily deals with non-iconic images as it is deployed in the real world where scenes are complex and the subject of interest will rarely be found in a clean iconic form. A significant challenge for VA use in security is developing models that are capable of crowded urban scene analysis, which are inherently complex scenes [2]. The non-iconic images in the MS COCO data-set are a better representation of the real world in which security VA is used. Non-iconic images data-sets have been found to produce image recognition models that are better at generalizing [3]. The referring expressions within the RefCOCO data-set enables the use of natural language processing models to find meaning in the captions which can then be used to create a link between natural language and image data. These two aspects of the RefCOCO data-set are crucial in order to create a model that is capable of searching video using natural language query.

5 Model and Approach

The purpose of NLQ-Security is to create a tool that is capable of using natural language query to search surveillance footage for timestamp ranges that satisfy the given query. Given an mp4 video file, NLQ-Security will allow users to input text queries to search the video for time ranges in which their object of interest is present in the video. For each timestamp range found, the image frame with the highest probability of satisfying the given query is returned with a bounding box displayed around the subject of interest.

When NLQ-Security first intakes a video it samples one frame from each second of the video to produce a collection of still images. After these images are collected, they are processed with Facebook's DEtection TRansformer (DETR) framework to produce a set of object detection features for each image. These features include probabilities and predicted bounding box coordinates for 91 different classes of objects. The 91 classes of objects are taken from the MS COCO data-set on which DETR was trained on [1]. For each image, only the bounding box coordinates for object detections with over 0.7 confidence are kept for later feature extraction.

After the initial processing of the video, additional features are generated with OpenAI's Contrastive Language-Image Pre-Training (CLIP) neural network. CLIP has been trained on image-phrase pairs and embeds text and images into the same vector space based on similarity. By embedding text and images into the same vector space, a cosine similarity can be calculated between the two embeddings which shows how closely related they are [6]. Embeddings for the text query, each video frame, and each of the bounding box image detections for the respective video frame from the previous step are generated. In addition to these features, two cosine similarity values are calculated for each frame-image detection pair. The cosine similarity between the video frame embedding and text query embedding is created, as well as the cosine similarity between the bounding box image detection and text query embedding. In addition to the features generated from the DETR and CLIP models, a set of positional features are created for each object detection that was produced from the DETR model. These positional features include: relative x and y coordinates of the bounding box, the relative area, normalized distance from the center, and the ratio of the image.

Once the video has been processed by the mentioned steps, a collection of 1,545 features are gathered for each video frame and image detection pair created. The features for each pair are fed into a neural network that has been trained on the RefCOCO dataset to obtain the probability that the sample satisfies the given text query. Frame and object detection pairs that are above 0.75 probability are then linked back to the timestamp that they occur at in the video. Any frames that are within five seconds of each other are grouped together to produce time ranges that start at the first occurrence and end at the last occurrence of a frame that satisfies the given text query. The frame-object detection pair that scored the highest probability within the time range is then listed with its respective frame and displayed to the user along with a bounding box around the subject of interest of the text query.

The data the neural network was trained on was created from the RefCOCO data-set in a somewhat similar manner to the video processing described. The RefCOCO dataset contains images that are each annotated with 2-4 referring expressions. Each referring expression contains information related to an object within the image that can be of one of 80 classes. The data utilized from the referring expression are the caption sentences used to describe the object(2-3 caption sentences per referring expression) and the bounding box coordinates in which the object of interest lies. For each image, a positive training sample is produced for each caption sentence. The positive training sample includes CLIP embeddings of the caption sentence, the main image, and the referring expression object sub-image. Positional features for the referring expression object sub-image are calculated as well as the cosine similarity between the image embedding and the caption sentence embedding and the cosine similarity between the sub-image and the caption sentence. Negative samples are created for each caption sentence and the features that represent all other referring expressions for the image besides the object which the caption sentence describes. In total 1,545 features are generated for each sample.

The neural network's architecture consists of three hidden layers that contain 64 nodes each. The input and hidden layers each have a ReLU activation function. The loss is calculated with PyTorch's binary cross entropy with logits function. The model was trained for 50 epochs on a training set of 456k samples generated from the RefCOCO data-set.

6 Evaluation

Although this project is intended to be used on video data, the model was trained on image data as it was more feasible to gather. The training set consisted of 120k positive samples and 336k negative samples for a total of 456k samples. As the model is trying to find whether an object of interest is present in the given image or not, the problem is modeled as binary classification. An imbalance in the ratio of positive to negative samples was created purposefully as the occurrences of a desired object in given security footage are expected to appear less often in the real world.

After 50 epochs of training, the neural network scored 85.1% accuracy on the test set. The baseline accuracy for the training and test sets is 73.6%

as the negative samples make up the majority of the samples for both sets. This accuracy is only marginally better than the baseline.

Precision	Recall	F1
.85	.85	.85

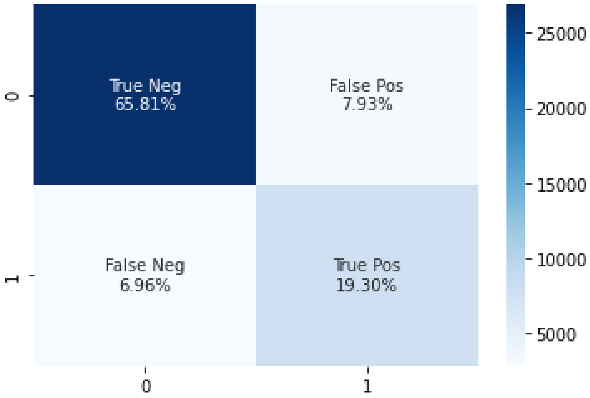

Figure 1: Confusion Matrix for performance on test set

7 Implications

NLQ-Security is intended to be used as a VA tool that can assist in processing many hours of surveillance footage in order to find objects of interest for an investigator. Ideally, NLQ-Security would reduce the number of man hours needed to process surveillance footage and help investigators locate the precise timestamps of footage that they need in order to find out more information about their target of interest.

However In its current state NLQ-Security should not replace human analysis of surveillance footage. The chance that NLQ-Security misses an occurrence of the target of interest is too high. This is due both to its poor accuracy and the fact that it only samples one frame for each second of video footage, potentially missing many times when the target of interest would appear in the video. It can be used though to get a better understanding of the content of surveillance footage and can yield useful information.

8 Conclusion

The use of advanced VA is transforming the way that security and surveillance is conducted around the world, for better and for worse. VA will most likely see increased use in the future as the ability to produce high quality video and analyze it is growing cheaper and more accessible [5]. NLQ-Security aims to solve just one type of problem that investigators may face when implementing video security for either commercial or private use.

References

[1] N. Carion et al. *End-to-End Object Detection with Transformers.* en. [online] arXiv.org. Available: 2021. URL: https://arxiv.org/abs/2005.12872.

[2] Ieeexploreieeeorg. *Video Analytics for Surveillance: Theory and Practice [From the Guest Editors.* en. [online]. Available: 2021. URL: https://ieeexplore.ieee.org/abstract/document/5562674.

[3] T. Lin et al. *Microsoft COCO: Common Objects in Context.* en. [online] arXiv.org. Available: 2021. URL: https://arxiv.org/abs/1405.0312.

[4] E. Longdin and A. *The history of CCTV – from 1942 to present - PCR.* en. [online] PCR. 2021. URL: Available:https://www.pcr-online.biz/2014/09/02/the-history-of-cctv-from-1942-to-present/.

[5] I. Olatunji and C. Cheng. *Video Analytics for Visual Surveillance and Applications: An Overview and Survey.* en. [online] Semanticscholar.org. Available: 2021. URL: https://www.semanticscholar.org/paper/Video-Analytics-for-Visual-Surveillance-and-An-and-Olatunji-Cheng/b918b9a7573956d3d718e88e988bfa7d16856779.

[6] A. Radford et al. *Learning Transferable Visual Models From Natural Language Supervision.* en. [online] arXiv.org. Available: 2021. URL: https://arxiv.org/abs/2103.00020v1.

[7] C. Shan et al. *Video Analytics for Business Intelligence.* da. Berlin, Heidelberg: Springer, 2012.

[8] L. Yu et al. *Modeling Context in Referring Expressions.* en. [online] arXiv.org. Available: 2021. URL: https://arxiv.org/abs/1608.00272.

[9] Y. Zhang, D. Larlus, and F. Perronnin. *What makes an Image Iconic? A Fine-Grained Case Study*. en. [online] arXiv.org. Available: 2021. URL: `https://arxiv.org/abs/1408.4325`.