Video Anomaly Detection With Self-Attention

Pedro Contipelli  
pedrocontipelli@knights.ucf.edu  
Praveen Tirupattur  
praveentirupattur@gmail.com  
Dr. Mubarak Shah  
https://www.crcv.ucf.edu/person/mubarak-shah/

Abstract

In this paper, we explore a weakly supervised method for anomaly detection using self-attention. Since making temporal annotations for videos can be very time consuming, we only look at weak video-level labels during training. This means that, given a video, we know that it is either normal or contains an anomaly, but not where that anomaly may be. Features are extracted from video clips and then passed through our encoder-decoder architecture with multi-head attention. Our loss function incorporates both the reconstruction loss of the predicted features and the classification loss of the predicted labels. Our goal is to show that the application of self-attention to this problem can offer a better contextual understanding of what defines an anomaly and improve upon existing state of the art methods. Additionally, we have been able to visualize the attention map activations per video in the UCF Crimes dataset and understand various relationships formed between parts of each video in our network. We show that when self-attention has been applied to our model, it can offer a 1.02% improvement in AUC score.

1 Introduction

The ability to detect an anomaly in a given video sequence is an extremely useful thing to be able to do. An anomaly can be considered to be any event that is a deviation from what is normal or expected. In our research, an anomaly could range to be anything from assault, robbery, and shootings to explosions and road accidents. The motivating idea is that many lives could be saved if we are able to automate the task of video anomaly detection accurately and reliably. A sufficient algorithm could be implemented to work on a live video feed and notify the correct first responders faster than any human would be able to do so. With that being said, this is still a very newly developing field and problem with much room for exploration.

Work on the problem of Anomaly Detection has come a long way. In 2013, one dataset which was very popular is the UCSD dataset. All of the videos in this dataset were acquired with a stationary camera in a fixed position, making it a vastly simpler problem than it is today. Now with the advent of social media and the increasing necessity of content-detection
from hand-held camera recordings with motion and many other factors, the standards we require for the problem are much more difficult to meet.

The three existing approaches to video anomaly detection are unsupervised, weakly supervised, and fully supervised. In an unsupervised training, no information is given about the training videos. In a weakly supervised approach, each video is labeled as either containing or not containing an anomaly somewhere in that video. And in a fully supervised approach, we are given temporally localized (clip-level) annotations highlighting exactly where in each video an anomaly is located.

We believe that self-attention could be very useful in a weakly-supervised setting. Self-attention has been a growing alternative to LSTMs that often performs better and trains faster. It has been shown to be very useful in machine reading, abstractive summarization, and image description generation. Using the newer UCF Crimes dataset, we split each video into several clips and apply self-attention on extracted features of those clips to hopefully get a better contextual analysis of the underlying features that correspond to an anomaly and apply this in our detection model. In this paper, we trained our model using C3D features extracted from those clips and then tried to predict which clips were anomalous in each video.

2 Related Works

The most recent (state of the art) work on video anomaly detection that we have been comparing our results to comes from two main papers. The first of these is [1], which presents a multiple instance learning model in which videos of each category (anomalous and not anomalous) are taken in pairs and the loss function formulation is such that it causes the network to learn to push the anomaly scores of positive instances and negative instances far apart. This multi-instance learning is a feature we have run experiments with and without, among many other different approaches and loss formulations. In [2], the use of label noise cleaning is applied by looking at feature similarity and temporal consistency across anomaly clips. We did not use label noise cleaning in our approach; however, we have considered adding in fully connected layers with some dropout before the self-attention layer after seeing their model use those before the feature similarity and temporal consistency graphs. Lastly, we also reimplemented and took inspiration from the many different possibilities for multi-instance learning loss functions presented in [3] for our experiments.

3 Method

3.1 Self-Attention

The novelty in our approach comes in the application of self-attention to this problem. Self-attention is a method of relating the different positions of a single sequence in order to compute a representation of the same sequence. Using self-attention, our model can be trained to learn and understand the context that each clip provides about every other clip of a video. This is very important in the act of classification, as it is otherwise very difficult to label individual clips as anomalous or not without the surrounding context, even for humans. In the process of our classification, we also come to understand the complex relationships between segments of the video and how they are dependent on one another by analyzing attention map activations for the videos in our dataset.
3.2 Architecture

Our architecture is depicted in Figure 1. Given an input video, we divide it into non-overlapping clips, each of 16 frames. We pass these clips through a C3D model pre-trained on Sports1M dataset. The extracted features are adjusted into 32 bags using interpolation. We use a batch size of 16 to train our model. So our input to the network is 16x32x4096 (16 batch size, 32 bags per video, 4096-dim feature per bag). At a high level, our model takes as input the interpolated features from each clip, encoded, fed through N layer(s) of self-attention, decoded, and also simultaneously sent through a classifier. Our classifier consists of X fully connected layer(s) halving in dimension and fed into ReLU function for non-linearity until the final M dimension to 1 layer which is passed through a sigmoid function for our final binary classification output between 0 and 1 (0 being normal; 1 being definitely anomalous).

3.3 Loss Formulation

We attempted experiments with many multi-instance learning approaches. Our best model used the following loss-formulation. For non-anomalous videos, we add the predicted score of the maximum anomaly predicted index and predicted feature score of the maximum anomaly predicted index to the loss. This is in order to reduce reconstruction and predicted anomaly scores of the maximum clip. For anomalous videos, we adding one minus the predicted score to the loss in order to train the model to raise the maximum clip’s predicted score.

3.4 Outputs and Evaluation Metric

We graphed the ROC curve and measured our results against existing state of the art methods using the Area Under the Curve (AUC) metric. This allows us to measure the accuracy of our results independent of threshold for anomaly detection. Additionally, we also visualized attention map activations on a per video basis for our best models.
4 Experiments

4.1 Dataset

UCF Crime contains 14 different classes of videos: Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, Vandalism, and Normal videos. It consists of 950 normal and 950 anomalous videos for a total of 1900 videos. The videos average 7247 frames in length, and the dataset contains overall approximately 128 hours of content. Video-level annotations are provided with the training set naturally leading us to use a weakly-supervised approach and frame-level annotations are provided with the testing set for accurate testing and evaluation. Since this dataset contains a larger number of longer surveillance videos with more realistic anomalies than many Anomaly Detection datasets, we believe that once we can achieve state of the art on it, we can apply our model to others.

4.2 Parameters

We achieved our best result of 77.49% AUC score using self-attention with the following parameters. We used 32 bags per video in our MIL with a batch size of 32 as well. We extracted features using C3D. Our model’s input dimension was 4096. We used a single layer of multi-head attention with 8 different heads. We had 3 fully-connected layers in our classifier going from 4096 to 2048 to 1024 to 1 in dimensionality. We trained for 100 epochs with a learning rate of 1e-4 and dropout 0.4 using ADAM optimizer.

4.3 Results

We have found in general, both with and without self-attention, that Xavier Normal weight initializations can offer best performance, with more stable training. We also have achieved a better AUC Score (77.49% vs 76.47%) with a single layer of attention than we have without attention.
While both our training loss and validation loss have been very stable, displaying the curves we expected in their decreasing, our AUC Score curves have been less stable. This leads us to believe that we should explore alternate loss formulations, as the network is training perfectly but has a hard time generalizing to the complicated problem of anomaly detection.

We have also looked at attention map activations for videos in the data sets and found that higher activations tend to correspond with normal event detection, and lack of activations usually correlates with anomaly in those time steps. We believe this is due to the reconstruction loss in our formulation.
5 Future Work

Analyzing videos from the dataset and our attention maps, we have seen and identified two main issues still needed to address to solve this problem in the future. There is an issue of anomalies that only take up a small portion of the frame, and so do not register because our approach is fully temporal, and we have not attempted spatial localization of anomalies. There is another issue of change in camera behavior, position, brightness, focus, etc which can cause false alarms in our anomaly detection with self-attention.

References