Scene Graphs for Interpretable Video Anomaly Classification

Nicholas F. Y. Chen, Zhiyuan Du, Khin Hua Ng
DSO National Laboratories
Singapore
{cfangyew,dzhiyuan,nkhinhua}@dso.org.sg

Abstract

Video anomaly classification has various applications for law enforcement agencies, such as analyzing post-criminal activities. However, surveillance videos pose a great challenge to video classifiers due to the sparsity of anomalies within the video, the complexity of the events and the large intra-class variations. For example, current approaches using end-to-end deep learning struggle on the recently released UCF Crimes benchmark. Meanwhile, there has been recent interest in using scene graphs for tasks involving both computer vision and natural language processing, as scene graphs are interpretable and can enhance model performance. Using the intuition that anomalies might leave clues in the object-relation-object triplets of the scene graph, we train models to directly learn from these triplets and achieved an accuracy of 23.8%, which is competitive to state-of-the-art video classification models. We further show that the scene graphs are able to point at the anomalies within the video, and potentially offer semantic explanations, giving our models a form of interpretability.

1 Introduction

Anomaly classification from surveillance video is an important application for law enforcement agencies in predicting and detecting crimes in progress, and analyzing post-criminal activities. One of the biggest challenges in anomaly detection is the rarity of such activities, leading to limited data. Another challenge is the length of some anomalous activities, which can be short relative to the time period of the surveillance video. Recent anomaly detection methods [1, 2, 3] use motion and temporal aspects of the video and offer some form of visual interpretability, but do not explicitly explain the context of the anomaly.

Scene graphs [4] of images have been shown to improve performance of tasks which involve both computer vision and natural language processing, such as image captioning [5], visual question answering (VQA) [6] and generating VQA explanations [7]. In a scene graph, a node represents an object while an edge represents the relationship between objects, and they have seen increased interest recently [8, 9, 10, 11, 12] because they equip current deep learning models with interpretability, reasoning capabilities and the ability to generalize to other tasks.

In this paper, we explore using scene graphs for video anomaly classification, and potentially providing explanations. The intuition is that object relationships in the scene graph can indicate the nature of the anomaly, making the model interpretable. For example, the 'man has gun' triplet could predict a 'shooting' activity. This leverages on language not only to describe properties of activities, but also to provide explanations. Recently, [13] introduced a new data set for video anomaly classification which includes more videos than existing data sets, and selected from actual surveillance footage. This data set is a challenging benchmark because the videos are untrimmed, the activities are complex and the intra-class variations are large. Furthermore, state-of-the-art video

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

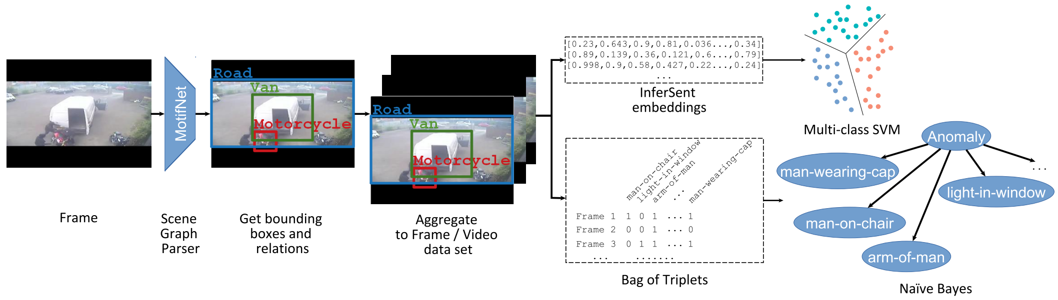

Figure 1: A schematic of our experiments.

classification models such as C3D [14] and TCNN [15] gave accuracies below 30%, making this data set a worthy challenge to tackle.

In section 2, we describe the representation of scene graphs and how we train models on them. We show that our models, which reason solely from scene graphs, are competitive to state-of-the-art deep learning models presented in [13]. Also, we show that the scene graphs can highlight the region in the video which contributes to the anomaly. In section 3, we present an analysis of the results and found that scene graphs do provide salient features in identifying anomalies even without taking temporal elements of the video into account.

2 Experiments

2.1 Implementation Details

We leave hyperparameters to the default settings in all our experiments (unless explicitly mentioned otherwise) to show that our results are due to the information embedded in scene graphs rather than excessive tuning. A schematic of our experiments can be found in Fig. 1. To ensure a fair comparison, all our experiments use the same 4-fold cross validation methodology as in [13].

Generating triplets Scene graphs were generated by running Stacked Motif Networks (MotifNet) [8], which were trained on 150 object classes and 50 relation classes from the Visual Genome data set, on the action recognition splits of the UCF Crimes data set [13] at a rate of 2 frames per second. The output, represented as lists of object-relation-object triplets, were filtered by discarding those with a total score (obtained by taking the product of the two object confidence scores with the predicate score) of less than 0.1.

Feature vectors The triplets in a specific aggregation period (explained later) were then converted to vector form in two ways. The first way is to use bag of triplets, where each element of the vector is a binary variable representing the presence of a triplet within the aggregation period, similar to bag of words. To exploit the semantic meaning of triplets, we also used InferSent sentence embeddings [16] (model version 2) to encode each triplet into a 4096-dimensional vector. Each aggregation period is represented by the average of the embeddings of the unique triplets within that specific period. For one experiment, we concatenated all triplets sequentially found in the aggregation period into one sentence, and converted it into an InferSent vector.

Two aggregation periods were chosen, namely the entire video and the entire frame (hence termed the *video*(V) and *frames*(F) data set respectively). For the *frames* data set, the classifier is trained to give a prediction for each frame, and predictions from frames of the same video will vote to give a prediction to the video.

Classifiers Since InferSent vectors have semantic meaning embedded in space, the support vector machines [17] (SVM) classifier is a reasonable choice. For bag of triplets features, we used the multinomial Naïve Bayes (NB) classifier to investigate whether the triplets are discriminative enough to recognize anomalous activities. In all NB models trained, the α smoothing parameter is set to 0.01 as the data set can be sparse.

Method	C3D [14]	TCNN [15]	NB(V)	NB(F)	SVM(V)	Cat+SVM(V)	SVM(F)
Accuracy (%)	23.0	**28.4**	23.8	18.3	18.5	15.2	14.1

Table 1: Comparing Sultani *et al.*'s [13] Activity Recognition results using C3D and TCNN to our method of using scene graph triplets with Naïve Bayes and SVM. Keys: Cat = concatenated triplets as sentences for InferSent; V, F = *video* and *frames* data set respectively.

	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	Normal	RoadAccidents	Robbery	Shooting	Shoplifting	Stealing	Vandalism	Avg
SVM(V)	0.0	0.0	27.1	4.2	6.2	0.0	6.2	0.0	29.2	22.9	0.0	**91.7**	**70.8**	0.0	18.5
Cat+SVM(V)	0.0	0.0	14.6	**10.4**	6.2	0.0	4.2	2.1	37.5	**56.2**	0.0	29.2	52.1	0.0	15.2
SVM(F)	2.1	**22.9**	22.9	0.0	**37.5**	6.2	16.7	0.0	0.0	0.0	0.0	68.8	20.8	0.0	14.1
NB(V)	**27.1**	6.2	**37.5**	8.3	18.7	**8.3**	14.6	**16.7**	**54.2**	41.2	**6.2**	54.2	35.4	**4.2**	**23.8**
NB(F)	20.8	8.3	8.3	0.0	22.9	6.2	**29.2**	10.4	18.7	18.7	0.0	56.2	56.2	0.0	18.3

Table 2: Class-specific recall (%) for models using outputs of scene graphs as features. Models in the top section of the table use InferSent embeddings while models in the bottom section use bag of triplets. Keys: Cat = concatenated triplets as sentences for InferSent; V, F = *video* and *frames* data set respectively.

2.2 Quantitative Results

Our best performing model, NB(V), performed better than the C3D baseline but worse than the TCNN baseline (see Table 1). For both NB and SVM classifiers, performance tends to be better on the *video* data set as compared to the *frames* data set, though all of our models perform better than random chance (7.1%). This indicates that the triplets have some discriminative power, and the voting mechanics induced noisy results. Another factor to consider is that the *video* data set is more dense than the *frames* data set and thus, a single instance can have more triplets in it. This suggests that each triplet is weakly discriminating towards an anomaly label, so models require an accumulation of triplets for better performance. The recall for each anomaly class is listed in Table 2. We see that each model is able to do well for some anomalies, and the best model (NB(V)) performs well for many anomalies. Shooting and vandalism are the hardest class to identify, and this will be explained later.

Additionally, we experimented with using the difference between adjacent vectors in the *frames* data set as features to see if changes between adjacent frames are useful, but we saw only minor gains relative to random chance. Moreover, we have also investigated how much of the triplets are discriminative enough to classify anomalies accurately. Using mutual information (MI), we train NB on the top 1% of features for *each* anomaly label and found that only some triplets are discriminative. More details can be found in the supplementary material.

2.3 Qualitative Results

From the MI feature selection above, we visualized some of the top 20 triplets to better understand our model. In some cases, the triplets do not make any semantic sense, but the detected region corresponding to the triplet corresponds to the anomaly (see Fig. 2). For example, in the 'explosion' anomaly, the triplet is 'person holding pizza' but the detection correctly highlights flames caused by an explosion. However some triplets do not make any sense at all, and these spurious correlations probably degraded the performance of our classifiers.

3 Discussion

Our quantitative results suggest that anomalous events leave clues in the object-relation-object triplets. The qualitative results suggest that the triplets, in its current form, cannot be taken literally. For example, flames are misclassified as 'light', as seen in the 'arson' anomaly. However, since the misclassification is consistent and unique, the model can differentiate arson from other anomaly classes. This also explains why the SVMs trained on InferSent features, originally intended to exploit semantic meaning, did not perform as well as the NB models. We have to keep in mind that the MotifNet only has a vocabulary of 150 object classes and 50 relation classes, so it is expected that

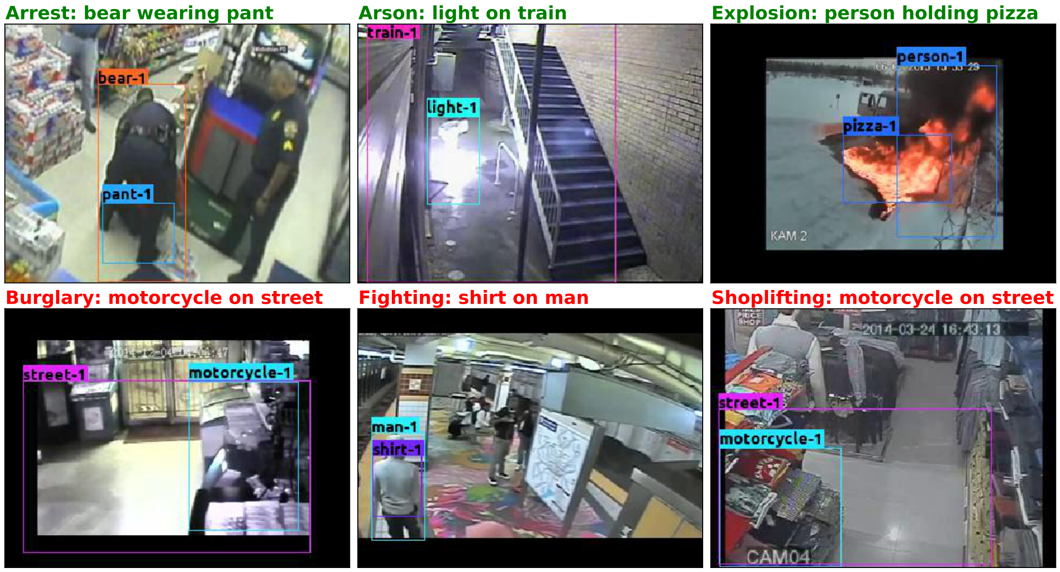

Figure 2: (Best viewed in color and zoomed in.) Top row: Positive examples where the anomaly was correctly identified by the triplets, though the triplets do not make semantic sense. Bottom row: Negative examples where the anomaly was falsely attributed to certain triplets. The titles correspond to the anomaly and its associated triplet.

many objects in the videos do not have a corresponding label in MotifNet. Furthermore, the videos are of 320×240 resolution which is quite poor, so objects can be misclassified quite easily. When the object detector component of scene graph parsers become accurate enough in giving correct object labels, then our approach can take advantage of semantic relationships of words for classification and offer semantic interpretability on top of the visual interpretability currently provided.

Related to this, NB(V) managed to predict some videos correctly in all classes (Table 2). This could be due to triplets being highly associated to some specific classes. However, some of the top few triplets identified the anomaly by exploiting the data set bias. Using MI for example, the top few triplets in 'abuse' are associated with bedrooms (e.g 'person on bed' and 'pillow in room'), where most of these videos take place. Another example is 'stealing', where many of the top triplets are associated with cars and motorcycles (e.g 'person on motorcycle' and 'car behind man') which appear in most of the 'stealing' videos. Despite this, the bag of triplets shows that it could be useful with anomalies that are associated with their environment like road accidents.

Lastly, the NB and SVM models did not take into account temporal information, *i.e.* changes in the graphs over time. Hence, anomalous activities which are identified by objects in a single frame (e.g. 'arson' and 'explosion' can be identified by fire) tend to be easier for our models to identify, whereas anomalous activities involving a sequence of actions (*e.g.* 'abuse' and 'vandalism') are probably better identified by a model which can track triplet changes in time.

4 Conclusions and Future Work

In conclusion, we trained video anomaly classifiers on scene graphs and found that a small proportion of the triplets have enough predictive power that our models perform better than random chance. Though we did not use temporal information, our models have competitive performance relative to state-of-the-art deep learning models. Our models offer visual interpretability as they are able to point at anomalous regions in the video, and potentially offer semantic interpretability when the object detector component is improved.

Future work includes the integration of object tracking so that analysis of dynamic scene graphs is possible, leading to machine understanding of how an anomalous event unfolds in time. When this becomes possible, a natural extension would be to perform anomaly detection where the time stamps of the anomalous period within the video can be produced.

References

[1] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[2] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016.

[3] Yingying Zhu, Nandita M Nayak, and Amit K Roy-Chowdhury. Context-aware activity recognition and anomaly detection in video. *J. Sel. Topics Signal Processing*, 7(1):91–101, 2013.

[4] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[5] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. *arXiv preprint arXiv:1809.07041*, 2018.

[6] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint*, 2017.

[7] Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind. Generating natural language explanations for visual question answering using scene graphs and visual attention. *XAI 2018*, page 45, 2018.

[8] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[9] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[10] Yikang Li, Wanli Ouyang, Zhou Bolei, Shi Jianping, Zhang Chao, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *ECCV*, 2018.

[11] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. *arXiv preprint arXiv:1808.00191*, 2018.

[12] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: learning object-agnostic visual relationship features. *arXiv preprint arXiv:1808.00171*, 2018.

[13] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[15] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *IEEE international conference on computer vision*, 2017.

[16] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[17] Zeyi Wen, Jiashuai Shi, Qinbin Li, Bingsheng He, and Jian Chen. ThunderSVM: A fast SVM library on GPUs and CPUs. *Journal of Machine Learning Research*, 19:1–5, 2018.