AI final

June 22, 2021

1 Introduction

"Driving a car by an Artificial Intelligent agent has been one of the greatest dream in AI for decades. In the past 10 years, significant advances have been achieved. Despite those great advances, there are two major challenges. The first challenge is how to drive safely with other "human drivers", a self-driving car should learn to anticipate others' behaviors in order to avoid these accidents. The other important challenge is how to scale-up the learning process. We propose to take advantage of the cheap and widely available dashboard cameras to observe corner cases at scale. Dashcam is very popular in places such as Russia, Taiwan and Korean. Many dashcam videos involving accidents have been recorded". Chan [3] proposed a method to learn from dashcam videos for anticipating various accidents and providing warnings a few seconds before the accidents occur: a Dynamic-Spatial-Attention (DSA) Recurrent Neural Network (RNN) to anticipate accidents before they occur.

Problem: The problem that we want to solve is to predict an accident before it occurs.

Importance: Many car accidents are caused by other human drivers. So it is important that a self-driving agent should learn to anticipate others' behaviors in order to avoid these accidents.

Learning to anticipate accidents is an extremely challenging task, since accidents are very diverse and they typically happen in a sudden. Based on [3], we modify its model by adding a branch which takes the pairwise object rela-

tion(feature) into consideration [5]. Also, we modify the original feature extraction method to Mask R-CNN [1].

2 Related work

From [3]:

A few works have been proposed to anticipate events — classify "future" event given "current" observation. Hoai and Torre [6] propose a maxmargin-based classifier to predict subtle facial expressions before they occur. Lan et al. [11] propose a new representation to predict the future actions of people in unconstrained in-the-wild footages.Our method is related to event anticipation, since accident can be consider as a special event.Kitani et al. [9] propose to forecast human trajectory by surrounding physical environment. Yuen and Torralba [16] propose to predict motion from still images.

There are also many works for predicting drivers' intention in the intelligent vehicle community. [2, 4, 10, 12] have used vehicle trajectories to predict the intent for lane change or turn maneuver.However, most methods assume that informative cues always appear at a fixed time before the maneuver. [7, 8] are two exceptions which use an input-output HMM and a RNN, respectively, to model the temporal order of cues. In order to address the challenges in accident anticipation, our method incorporates a RNN with spatial attention mechanism to focus on object-specific cues at each frame dynamically. But, these above methods focus on anticipating specific-maneuvers such as lane change or turn. In contrast, we aim at anticipating various accidents observed in naturally captured dashcam videos.

RNN with attention has been applied on computer vision tasks: video/image captioning and object recognition. RNN with soft-attention has been used to jointly model a visual observation and a sentence for video/image caption generation [15, 14]. Yao et al. [15] incorporate a "temporal" soft-attention mechanism to select critical frames and Xu et al. [14] demonstrate the power of spatial-attention mechanism for generating an image caption. Compared to [14], our proposed dynamic-spatial attention RNN has two main differences: (1) their spatial-attention is for a single frame, whereas our spatial-attention changes dy-

namically at each frame in a sequence; (2) rather than applying spatial-attention on a regular grid, we apply a object detector [13] to extract object candidates for assigning the spatial-attention.

---

[3] considers frame and object observations of each frame to predict if there will be an accident or not. We thought the relation between objects that they didn't consider is important too, since accidents are mostly involved with two agents(ex. crash between cars), so we add the branch that takes pairwise object feature into the origin model then do the spatial-attention work(See Fig.1).

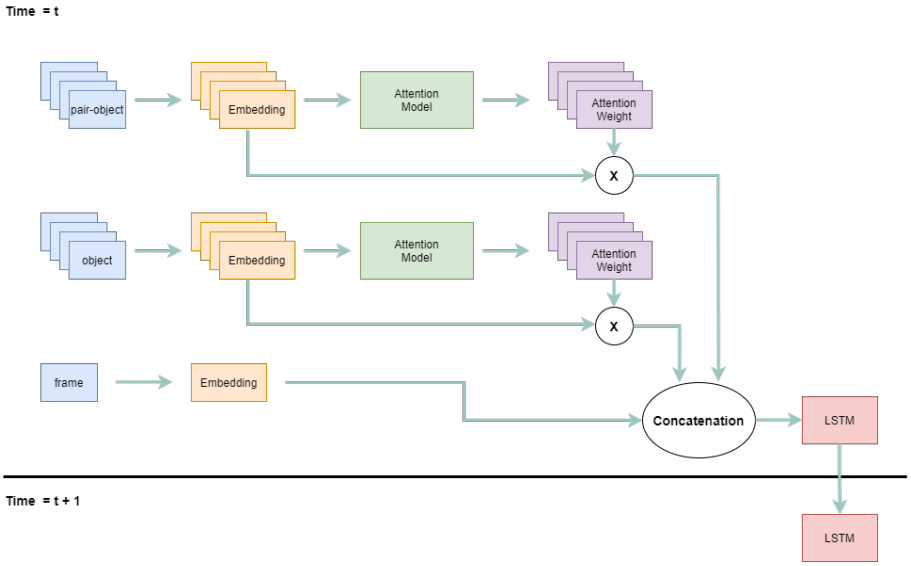

Figure 1: Flowchart

3 Methodology

Given a set of videos, for each frame, we detect objects to extract features and bounding boxes by Mask R-CNN. For t-th frame, we define our observations as $X_t = (x_1, x_2, ..., x_T)_t$ where x_i is the object observation of i-th object at this frame and we assume there are $T \leq 20$ objects in each frame. Besides single objects, we learn the spatial relation between each of them by extracting

union box features through an RoIAlign layer of a Mask R-CNN model denoted by $(z_{12}, z_{13}, ..., z_{23}, z_{24}, z_K)_t$, where z_{ij} denotes the union box features of i-th object and j-th object and we assume there are $K \leq 100$ union boxes. If the two objects are far, we ignore the union box of these two objects since there will be low chance to have an accident between these two objects. We take these objects' features and pairwise features as input of our standard RNN model based on LSTM and get the accident label y_t to specify at which frame the accident started. The feature extraction model Mask R-CNN is based on Feature Pyramid Network (FPN) and a ResNet101 backbone and is pre-trained on MS COCO dataset.

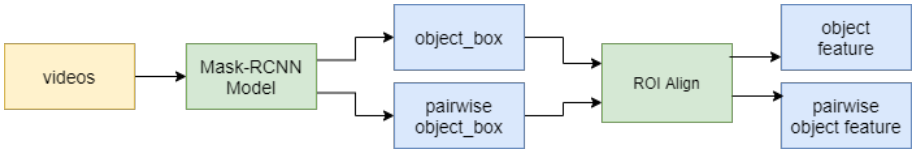

Figure 2: Feature extraction

Input data are dash-cam videos from [3] which are harvested from YouTube by many users. Our model contains two branches, one is single object and one is single object with pairwise object. For each frame, we first get the embedding features. After that we adapt soft-attention mechanism to take dynamic weighted-sum of spatially-specific object observations. Last, we concatenate two branches and take it as input of LSTM.

Loss: We use exponential loss for positive accident training videos and standard cross-entropy loss for negative accident training videos [3].

4 Experiments

Dataset: The number of training clips is about three times the number of testing clips: 140 training clips (46 positive and 94 negative clips) and 58 testing clips (16 positive and 42 negative clips). Feature extraction is time consuming which leads to small amount of dataset.

We will test our model by comparing the evaluation score of 2 branches (single object + pairwise object) and one branch (single object). Evaluation

score is to calculate precision= $\frac{TP}{TP+FP}$ (TP: True positive, FP: False positive).

	No pairwise obj.	With pairwise obj.
Training	0.9362	0.5507
testing	0.4253	0.4301

Figure 3: Precision

One branch model: We get precision= on the model that only considers single object.

Two branches model: We get precision= on the model that considers single object and pairwise object.

5 Conclusion

The pre-trained Mask R-CNN model is trained on MS COCO dataset which is not trained to detect street scenes and not related to any traffic accidents. If we finetune the last fully connected layers of Mask R-CNN on street scenes data, the performance of object detection will increase. Also, lack of computing resources makes the amount of data set is small.

Our experiment shows that precision is better on single object model, this is because our dataset is too small to train the model since the parameters of pairwise object is too huge for small dataset. If we increase the amount of dataset, pairwise object model will improve.

6 Github repo

https://github.com/nargoo0328/NYCU-AI-Final-Project

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[2] H. Berndt, J. Emmert, and K. Dietmayer. Continuous driver intention recognition with hidden markov models. in: Intelligent transportation systems, 2008.

[3] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[4] B. Frohlich, M. Enzweiler, and U. Franke. Will this car change the lane? - turn signal recognition in the frequency domain, 2014.

[5] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. Spatio-temporal action graph networks, 2019.

[6] M. Hoai, De la Torre, and F. Max-margin early event detectors, 2012.

[7] A. Jain, Raghavan B. Koppula, H.S., S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models, 2015.

[8] A. Jain, A. Singh, H.S. Koppula, S. Soh, and A. Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture, 2016.

[9] B.D. Kitani, K.M.and Ziebart, J.A.D. Bagnell, and M Hebert. Activity forecasting, 2012.

[10] P. Kumar, M. Perrollaz, S. Lefvre, and C. Laugier. Learning-based approach for online lane change intention prediction, 2013.

[11] T Lan, T.C. Chen, and S Savarese. A hierarchical representation for future action prediction, 2014.

[12] M. Liebner, M. Baumann, F. Klanner, and C. Stiller. Driver intent inference at urban intersections using the intelligent driver model, 2012.

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[15] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure, 2015.

[16] J. Yuen and A Torralba. A data-driven approach for event prediction, 2010.