2021 IEEE 10th Global Conference on Consumer Electronics (GCCE) | 978-1-6654-3676-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GCCE53005.2021.9622059

A Lightweight Deep Learning Model for MOT

Xiu-Zhi Chen
Dept. Computer Science and Information Engineering
National Taipei University of Technology
Taipei, Taiwan
t107599006@ntut.edu.tw

Jhen-Hao Li
Dept. Computer Science and Information Engineering
National Taipei University of Technology
Taipei, Taiwan
t109598018@ntut.edu.tw

Yen-Lin Chen
Dept. Computer Science and Information Engineering
National Taipei University of Technology
Taipei, Taiwan
ylchen@mail.ntut.edu.tw

***Abstract*—Multiple object tracking (MOT) is a high complexity computer vision task, it has to detect multiple target objects in frames and extract their features for matching. Through deep learning techniques, MOT can be solved much easier while getting more accurate results, however it is still hard to be adopted for real-time applications because of the high computational cost. In this research, we applied pruning method on JDE, which is a well-known MOT model structure, and proposed a lightweight JDE model that obtained acceptable accuracy with lower computational cost. Evaluate our proposed model through real-world surveillance video, the result has shown that comparing to original JDE, detection accuracy had decreased about 15.65% in average, although computational efficiency had increased about 71.54% in average, which proved that the proposed model is possible to be used for real-time applications.**

Keywords—lightweight deep learning model, multiple object tracking, pruning method

I. Introduction

There are lots of computer vision tasks were implemented through deep learning techniques nowadays, for example, RCNN serious[1][2][3] and YOLO[4][5][6][7] are well-known object detection and recognition models. The performance of these models were not high enough at first, but after improving, both accuracy and efficiency were good enough to be adopted for real-world applications. For multiple object tracking (MOT) task, proper method that satisfied real-world application requirement haven't exist yet. Joint detection and embedding (JDE) model[8] is one of the most latest MOT deep learning model, it's accuracy and efficiency are quite high under low resolution, but as the resolution gets higher, it's efficiency will be decreased, and is unable to reach real-time requirement. In this research, we are going to apply pruning method on JDE model, and proposed a lightweight JDE model that obtains acceptable accuracy, with lower computational cost. We took MOT17[9] as our training dataset, and evaluated the performance through a real-world surveillance video sequence, download from Pixabay[10], and shown that comparing to original JDE, detection accuracy had decreased about 15.65% in average while the computational efficiency had increased about 71.54% in average, which proved that the proposed model is possible to be used for real-time applications.

II. Methods

Existing MOT method mostly were tracking-by-detection, it detect targets first, than extract their features. These kinds of methods suffer from efficiency problems. To deal with this problem, JDE combines two parts into single model, which extracts the features as it gets the detection results. Fig. 1 shows the schematic diagram of JDE.

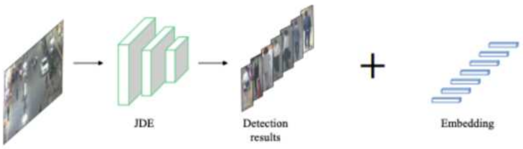

Fig. 1. Schematic diagram of JDE.[8]

The architecture of JDE is similar to YOLOv3[6], it added an embedding part into model structure. We trained the whole model through the prepared dataset, and let the embedding parts learned that how to extract useful features to complete the accurate matching between the current objects and those in the previous frame. We took JDE as our basic model, and analyzed that there are 116 layers in it. To lower the computational cost of it, we applied pruning method on the top 55 layers.

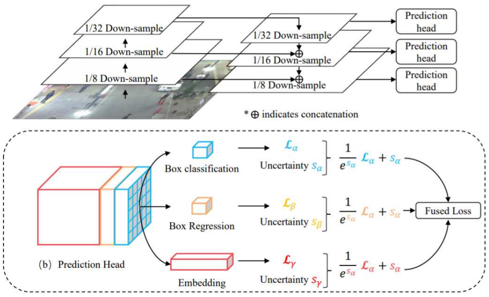

Fig. 2. Architecture of JDE.[8]

The main concept of pruning models is to analyze which group of filters has the highest contribution, by reserving those filters and remove the others, the number of weights become fewer. The previous method had claim that the weights value of filters is a valuable reference for the importance of filters[11], as a result, we trained our model followed by the concept of it. Fig. 3 shows the flowchart of the adopted pruning method, we set a target pruning rate for each layer before starting the model training. As the training began, the L^2-Norm of weights from filters in each layer will be calculated. After sorting the values, the last ***N*** filters of each layer was set as 0, indicated that the filters was pruned.

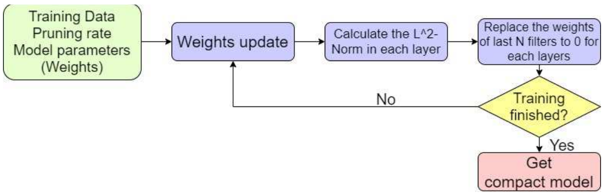

Fig. 3. Flowchart of pruning method.

Although the pruning method mentioned above lowered the number of weights, applying the same pruning rate on each layer is not a good approach. As we know, the receptive field of the units between the deep layers and the shallow layers are different. Referred to previous researches, we knew that having different pruning rate for different layers could get better pruning results[12], therefore, we divided the top 55 layers in JDE in to three parts and defined different pruning rate for them, shown as TABLE I. We took the layer division rules shown in TABLE I into the pruning method shown in Fig. 3 to come up with our proposed lightweight deep learning MOT model.

TABLE I. LAYER DIVISION RULES

Level	Condition	Pruning Rate
Shallow	Layers < 12	12.50%
Medium	13 <= Layers < 37	50.00%
Deep	38 <= Layers < 55	87.50%

III. EXPERIMENTS

A. Experimental Setup and Details

MOTChallenge[9] is a well-known benchmark for single-camera MOT, we took MOT17 as our training dataset for our model training. There are 7 different scenes in MOT17, each scene has three different sequences. Separated the sequences into frames, we divided 15948 frames for training set, and 17757 frames for testing set. The input image size of JDE was set as 1088x608, and the model was trained over 100 epochs using Nvidia GeForce GTX 1080 Ti GPU with PyTorch 1.5.0 deep learning framework. The video sequence for evaluation is a real-world surveillance video sequence, download from Pixabay[10], the captured frame of it is shown as Fig. 4.

Fig. 4. Captured frame of the evaluation sequence.

B. Results

After the model training, we observed the number of trainable weights and compared it with the original one. TABLE II shows the comparison result, it includes 66,950,747 trainable variables in the original model, and 58,071,087 trainable variables left after pruning by our proposed method, which compacts about 13.26%.

TABLE II. COMPARISON OF THE TRAINABLE WEIGHTS AMOUNT

Model	Trainable weights
Original JDE model	66,950,747
Proposed model	58,071,087 (-8,879,660)

To proof that the performance of the compact model is still functional, we applied the models on the evaluation sequence, and calculated the detection accuracy and efficiency every 20 frames. Although comparing to JDE, the detection accuracy is lower about 15.65% in average, the efficiency increased about 71.54% in average, which is a big improvement for developing real-time applications, the details are shown in TABLE 3.

TABLE 3. COMPARISON OF THE DETECTION ACCURACY AND EFFICIENCY

Frame Index	Number of detected pedestrian		Efficiency (fps)	
	JDE	*Proposed model*	*JDE*	*Proposed model*
20	38	28	7.52	6.77
40	37	31	7.75	9.88
60	37	33	7.86	11.58
80	35	31	8.06	12.75
100	39	35	8.15	13.58
120	33	26	8.29	14.24
140	30	22	8.48	14.76
160	31	21	8.64	15.26
180	33	31	8.71	15.60
200	33	32	8.76	15.86
220	32	27	7.81	16.11
240	28	23	8.89	16.38
260	28	23	8.94	16.58
280	29	25	8.99	16.72
300	31	26	8.99	16.84
320	28	24	9.04	16.95
340	33	30	9.08	17.07
Average	32.64	27.53(-5.11)	8.47	14.53(+6.06)

IV. CONCLUSION

In this research, we took JDE as our basic model, and applied pruning method with our self-defined pruning rate rules. Experimental results show that the proposed model is truly compacted and obtain acceptable performance on MOT task. In further, we would do more studies and try to increase the stability of object detection and it's accuracy.

ACKNOWLEDGMENT

This study is conducted under the "New generation transport deep learning intelligent system" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China, and also financially sponsored by the Ministry of Science and Technology of Taiwan under the granted No. MOST 109-2628-E-027-004-MY3.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," Region-Based Convolutional Networks for Accurate Object Detection and Segmentation, vol. 38, no. 1, pp. 142-158, May, 2015.

[2] R. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015.

[3] S. Ren, K. He, R. Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June, 2016.

[4] J. Redmon, S. Divval, R. Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[6] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv: 1804.02767.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv: 2004.10934.

[8] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards Real-Time Multi-Object Tracking," arXiv: 1909.12605.

[9] P. Dendorfer, A. Ošep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, and L. Leal-Taixé, "MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking," arXiv: 2020.07548.

[10] Pixabay. [Available] https://pixabay.com/

[11] Y. He, G. Kang, X. Dong, Y. Fu, and Y. Yang , "Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks," arXiv: 1808.06866.

[12] L. Li, J. Zhu, and M.-T. Sun,"Deep Learning Based Method For Pruning Deep Neural Networks," in 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 2019.