ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Random Signal Cut for Improving Multimodal CNN Robustness of 2D Road Object Detection

Robin Condat, Alexandrina Rogozan, Abdelaziz Bensrhair

Normandie Univ - INSA Rouen, LITIS
Rouen - France

Abstract. Given the large number of deep neural network proposals using only RGB images for 2D object detection for Advanced Driver-Assistance Systems, we propose MMRetina, a CNN taking multimodal data (RGB, Depth from Stereo, Optical Flow, LIDAR) as input for detecting road objects and their 2D localization. We introduce a new data augmentation method, we called Random Signal Cut, to make our multimodal CNN more robust to sensor malfunctions or breakdowns. The experiments show on KITTI dataset that using multimodal data with Random Signal Cut improves significantly CNN robustness without lowering its overall performances when all sensors are well functioning.

1 Introduction

Detection of road traffic actors is nowadays a challenging task and plays an important role in the field of Advanced Driver-Assistance Systems (ADAS). In order to avoid human errors in driving, which are the majority of road accidents causes, a lot of work has been done on 2D and 3D object detection for autonomous driving car. Although the KITTI object detection benchmark [1] relates that the state-of-the-art algorithms are able to achieve about 90 % average precision (AP) in road object detection, there are still improvements needed to get a very high accuracy in real time and real world environments. In the last decade, the emergence of Convolutional Neural Networks (CNN) has made it possible to obtain very high performances in real time.

Moreover, most of the work on 2D road object detection focus only on ideal conditions, where all sensors used are well calibrated and work perfectly. Unfortunately, in real world environments, one or even more sensors may be uncalibrated, partially work, or even be out of order. For this reason, we can not rely on a single sensor, but on several ones working together. However, using multimodal embedded ADAS does not prevent the risk of sensor malfunction or failure. Indeed, severe weather, hilly road or sudden breaking may uncalibrate a sensor, render it out of use or even damage it. Thus, it is necessary to prove that the multimodal detection of road traffic actors is still robust. Therefore, one must also estimate the impact of sensor malfunction or breakdown within an ADAS, not only the contribution of the multimodal approach in ideal laboratory conditions.

In this paper, we first propose MMRetina (MultiModal RetinaNet), a new CNN detector which takes multimodal data as input, from a sensor system consisting of 2 RGB stereo vision cameras and a LIDAR, in order to detect both the class and the 2D localization of road objects (car, pedestrian, cyclist, among others). We then introduce an original data augmentation technique, we called Random Signal Cut, to strengthen our network during a sensor breakdown or malfunction. Finally, we evaluate the robustness of our approach for the task of 2D detection on the challenging KITTI object detection benchmark.

2 Related Work

In the context of autonomous driving, most of the existing work use only RGB images for object recognition, object detection or semantic segmentation. It can be explained by the fact that there are not many databases including several modalities. When using multimodality, authors usually combines RGB images with depth from stereo vision [2, 3], optical flow [3, 4], LIDAR [4, 5] or thermal images [6]. In [4] it is presented a boosting-based sliding window solution for object detection that exploits information from RGB images, optical flow and LIDAR front view. Multi-View 3D networks are introduced in [5], where a sensory-fusion framework takes both RGB images and LIDAR point cloud, in front view and bird's eye view, as input and predicts oriented 3D bounding boxes. To the best of our knowledge, we believe that there are never been any work combining RGB images, depth from stereo, optical flow and LIDAR point cloud for ADAS and moreover with deep learning approach. This is potentially due to the fact that depth from stereo and LIDAR represent correlated information, therefore there are usually considered redundant.

A few work exist concerning CNN robustness in case of noisy or missing input data. In [7], a RGB-D architecture for object recognition and a data augmentation scheme are presented, for robust learning with depth images by corrupting them with realistic noise patterns.

3 Improving Multimodal CNN Robustness

We propose a multimodal CNN which takes, as input, data from a system including 2 RGB stereo vision cameras and a LIDAR. We use Semi Global Block Matching algorithm [8] for RGB images from these cameras to extract the depth image, and Farneback algorithm [9] for two temporal adjacent images from left camera to extract the optical flow image. We project LIDAR point cloud on left camera plane and then, we apply a linear interpolation, to provide the LIDAR front view image. Finally, we have, in addition to the left camera RGB image, a depth image from stereo vision (DP), an optical flow image (OF) and a LIDAR front view image (LD), allowing for 4 signals. For this work, we consider each of these 4 signals as a 3-channel image.

3.1 Our baseline Multimodal CNN

We propose MMRetina, a new multimodal mid-fusion 2D object detector (see Fig 1), based on RetinaNet [10], taking as input RGB images, depth from stereo,

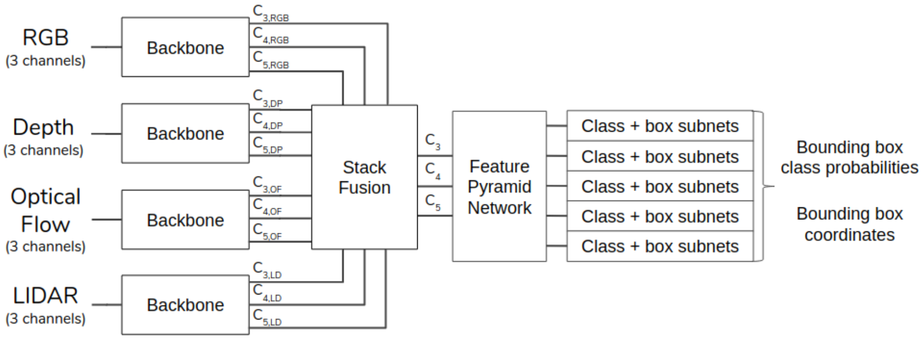

Fig. 1: Our Baseline Multimodal Mid-Fusion 2D Object Detector.

optical flow and LIDAR front view image. Its architecture offers a good compromise between computation speed and high object detection accuracy. Our network is first composed of 4 backbones with shared weights in order to have fewer parameters to optimize. For each signal $k \in \{RGB, DP, OF, LD\}$, its corresponding backbone provides 3 feature maps : $C_{3,k}$, $C_{4,k}$ and $C_{5,k}$. Mid-fusion is performed by stack fusion of these feature maps to obtain 3 fused feature maps : C_3, C_4 and C_5. These backbones are combined with a Feature Pyramid Network that extracts relevant multi-scale feature maps. Then, a first subnetwork classifies backbone outputs by predicting the probability of object presence at each spatial position, for each defined anchor and for each object class. Finally, in parallel of the classification subnetwork, a second one performs convolution bounding box regression for predicting the object location with respect to anchor box if an object is detected.

3.2 Random Signal Cut

The main goal is to make our proposed CNN more robust to sensor malfunctions or breakdowns. For that, we have to evaluate our CNN not only with complete multimodal input data, but also with multimodal input data with missing signals, simulating data produced by a system with defecting sensors. In our work, we consider that a sensor is either fully functional, or out of order and produce a null signal, that is to say a black 3-channel image. For instance, our system with one RGB camera breakdown would provide a RGB signal and an optical flow signal from the remaining functional camera, and a LIDAR signal. The CNN would receive 4 signals as input, including a black 3-channel image for depth signal. Table 2 shows the signal combinations according to the possible cases of breakdown that may occur. Our first idea was to train our network not only with complete, but also partial multimodal input data, depending on these possible signal combinations. However, the risk is to unbalance CNN learning by rejecting one or several signals, less present than the others.

To avoid this phenomenon, we propose a new data augmentation technique called Random Signal Cut (RSC). This approach aims to generate partial multi-

Benchmark	E	M	H
Car	89.66	79.53	69.52
Pedestrian	59.63	41.63	36.97
Cyclist	43.71	28.00	24.62

Table 1: Detection AP (%) of MM-Retina on KITTI object detection benchmark [1] according to three level of difficulties : Easy (E), Moderate (M) and Hard (H).

Case	System configuration	Available signals			
		RGB	DP	OF	LD
1	2 cameras + LIDAR	✓	✓	✓	✓
2	1 camera + LIDAR	✓	✗	✓	✓
3	2 cameras	✓	✓	✓	✗
4	1 camera	✓	✗	✓	✗
5	LIDAR	✗	✗	✗	✓

Table 2: Available signals according to system configuration.

modal input data for the training dataset. For each modality signal, we assign a Random Signal Cutoff Rate (RSCR), between 0 and 100 % corresponding to the percentage of input data signal absence, replaced by a black 3-channel image, during CNN training. It is to be noticed that each signal cut is made independently of the other signals. However, we lock the possibility to have less than one signal available in order to avoid null data in our network training. RSCR of each signal could be different, but in these experiments, all signals have the same RSCR. Finally, one can consider the RSCR values of one multimodal CNN as hyperparameters having to be optimized.

4 Experiments

Experiments are divided in two parts. First, we evaluate our baseline CNN on KITTI object detection benchmark [1]. Second, we analyze Random Signal Cut impact on multimodal CNN robustness in case of single or multiple breakdowns. For all the CNNs developed for our experiments, we use samples of KITTI training dataset, which consists of 7481 images with 39597 relevant objects labeled in 7 different classes. We divide them into 6 disjoint folders for a 6-fold cross-validation with 5 folders for training, and one for validation, in order to have enough data for training (more than 6000 images) and for validation (approximately 1250 images). All CNNs developed are trained on the training set with ADAM optimizer, a learning rate of 10^{-4} and a batch size of 8. We train a network during 400 epochs and keep its version with the lowest validation loss for evaluation. We rescale the input data such that their shorter size is fixed to 400 pixels, therefore the average size for each modality signal of input data is 1330×400 pixels per channel. For each CNN backbone, we chose ResNet18 [11], because of computational resource constraint.

4.1 Evaluation of the baseline multimodal CNN

For the first part, we train 6 identical models from our baseline CNN without RSC, with 6-fold cross validation, and select the one that got the best performances on its validation set. We evaluate its object detection performance on KITTI benchmark test set, using the PASCAL criteria with detection Average

	No sensor breakdown	Cases with several sensor breakdowns			
RSCR	Case 1	Case 2	Case 3	Case 4	Case 5
0 %	82.87 ± 0.94	74.45 ± 1.10	23.28 ± 1.51	15.47 ± 2.81	0.97 ± 0.82
2.5 %	83.14 ± 0.70	81.57 ± 0.56	59.40 ± 2.34	52.40 ± 1.98	48.38 ± 2.64
5 %	82.33 ± 1.14	81.25 ± 1.05	64.50 ± 1.97	57.59 ± 2.36	55.06 ± 1.59
10 %	82.75 ± 1.03	81.98 ± 1.05	69.03 ± 0.86	62.74 ± 1.05	62.01 ± 3.08
15 %	**83.84 ± 0.89**	**82.44 ± 1.49**	70.47 ± 1.93	65.80 ± 2.60	66.87 ± 2.71
20 %	82.64 ± 0.95	81.92 ± 1.19	71.23 ± 1.23	66.57 ± 0.97	67.93 ± 1.68
25 %	82.79 ± 1.33	82.76 ± 1.38	72.88 ± 1.84	68.33 ± 2.29	70.43 ± 2.06
50 %	82.29 ± 1.47	81.44 ± 1.20	**74.18 ± 2.86**	**71.22 ± 2.66**	**72.88 ± 2.28**

Table 3: Mean Average Precision (%) ± Standard Derivation of our multimodal CNN depending on RSCR applied to its input signals with 6-fold cross validation.

Precision (AP %) proposed in [1] with an overlap of 70 % required for cars and an overlap of 50 % for pedestrians and cyclists. Table 1 shows overall performances of our proposed CNN for car, pedestrian and cyclist detection. We obtain satisfying performances in car detection, although less successful than state-of-the-art best solutions[1]. However, we obtain lower accuracy in pedestrian and cyclist detection, because these objects are smaller and less present than cars in KITTI dataset. Thus we believe that these are more difficult to detect. The best solution to avoid this problem is to enlarge the input data size but we could not do it because of computational resources lack.

4.2 Analysis of RSC impact on multimodal CNN robustness

For the second experiment, we compare several versions of our proposed multimodal network with different RSCR. We consider CNNs trained without RSC technique (classical learning) as CNNs with an RSCR of 0 %. We study their overall performances (Mean Average Precision on 7 object classes) and their robustness in case of failure. Table 2 shows all cases of breakdown and the available signals according to sensor system functioning. CNNs performances are measured for all these cases listed with their unavailable signals replaced by black 3-channel images. We evaluate our models on their validation set, since we do not have the possibility to test them all on KITTI benchmark test set. We show in table 3 that RSC use does not influence overall performances for our multimodal CNN in ideal laboratory conditions (Case 1) with a Mean Average Precision between 82 and 84 %. On the other hand, RSC use gives significantly better performances in case of failure (Cases 2 to 5) than classical learning without RSC. We believe that our method made our CNN able to detect road objects by extracting available information from each signal only when it is available, independently of the other signals. Finally, we notice that, like other data augmentation methods, RSC slows down CNN overfitting.

[1] State-of-the-art best algorithms performances can be viewed on `http://www.cvlibs.net/datasets/kitti/eval_object.php`

5 Conclusion

In this paper, we propose MMRetina, a new multimodal mid-fusion CNN road object detector and Random Signal Cut, an original data augmentation method for increasing CNN robustness to several malfunctions or breakdowns. Random Signal Cut makes significantly CNNs more robust without lowering its overall performances when all sensors are well functioning. Moreover, this technique is generic and could be applied for other applications than road object detection. Its impact could be improved with different adapted rates for each signal depending on its characteristics. In future work, we will test Random Signal Cut with different modality-adaptative rates, by optimization techniques. We will also improve our overall performances by extending RSC to cross-dataset training, where one could use several datasets for training, even if they do not have all needed multimodal signals available.

6 Acknowledgment

The present work was performed using computing resources of CRIANN (Normandy, France).

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, June 2012.

[2] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *CoRR*, abs/1608.07711, 2016.

[3] D. O. Pop, A. Rogozan, F. Nashashibi, and A. Bensrhair. Improving Pedestrian Recognition using Incremental Cross Modality Deep Learning. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, April 2019.

[4] A. Daniel Costea, R. Varga, and S. Nedevschi. Fast boosting based detection using scale invariant multimodal multiresolution filtered features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. *CoRR*, abs/1611.07759, 2016.

[6] Y. Zheng, I. H. Izzat, and S. Ziaee. GFD-SSD: gated fusion double SSD for multispectral pedestrian detection. *CoRR*, abs/1903.06999, 2019.

[7] A. Eitel, J. T. Springenberg, L. Spinello, M. A. Riedmiller, and W. Burgard. Multimodal deep learning for robust RGB-D object recognition. *CoRR*, abs/1507.06821, 2015.

[8] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008.

[9] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[10] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.