

Rare Traffic Signs Recognition

Tingir Badmaev¹, Vlad Shakhuro¹ and Anton Konushin^{1,2,3}

¹*Lomonosov Moscow State University, 1-52, Leninskiye Gory, 119991, Moscow, Russia*

²*NRU Higher School of Economics, 11, Pokrovsky boulevard, 109028, Moscow, Russia*

³*Samsung AI Center, 5-C, Lesnaya street, 125047, Moscow, Russia*

Abstract

Recognition of road signs is an important part of the control systems of autonomous vehicles and driver assistance systems. Modern recognition methods based on neural networks require large well-labeled datasets. Marking up data is quite expensive, but it is even more difficult to mark up rare classes of objects. To solve this problem in this article, we use synthetic data. We improve the marking of the Russian traffic signs dataset (RTSD) in semi-automatic mode adding 9 thousand new road signs. We perform an experimental evaluation of the currently best classifiers and detectors in the task of recognizing road signs. To improve the performance of classification, we use stochastic weight averaging (SWA) and contrastive loss. The use of modern methods allows us to train a high-quality neural network on synthetic data, which was previously impossible, and significantly improves the metrics of recognition of both rare and frequent road signs.

Keywords

Computer vision, Image recognition, CNN, Image classification, Object detection

1. Introduction

Recently, self-driving cars control systems have been actively developing. In these systems, an important component is an algorithm for recognizing road signs. In addition to autonomous vehicles, sign recognition is used in driver assistance systems and for automating the processes of road maintenance services. Modern methods of object recognition in images are based on deep learning models. To get a high-quality detection model, a well-labeled dataset is usually required. Dataset markup process is expensive and time-consuming since it requires manual routine work, which must be rechecked taking into account the inevitable human errors. Synthetic data can reduce the complexity of data collection. They can be obtained quickly, for free, without errors in the annotation, and in almost unlimited sizes, which greatly reduces the cost of obtaining data.

The task of recognizing road signs by its nature has highly unbalanced classes. Many classes of signs are missing in the training samples, which complicates the training of models. It has been previously shown that using the realistic synthetic generation of datasets with rare classes of Road images augmentations[1] it is possible to improve models of traffic signs recognition.


GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27–30, 2021, Nizhny Novgorod, Russia

✉ tingir.badmaev@graphics.cs.msu.ru (T. Badmaev); vlad.shakhuro@graphics.cs.msu.ru (V. Shakhuro); anton.konushin@graphics.cs.msu.ru (A. Konushin)

🆔 0000-0002-3318-9618 (T. Badmaev); 0000-0002-1586-9257 (V. Shakhuro); 0000-0002-6152-0021 (A. Konushin)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In this work, we investigate the recognition of both frequent road signs and rare ones because their significance on the road is not lower than frequent ones. Firstly, we improve the current markup of the Russian traffic signs dataset[2] in semi-automatic mode. We add 9000 new signs to the markup: 4000 signs to the test set and 5000 signs to the train set. Secondly, we perform an experimental evaluation of modern methods of classification and detection of road signs. We neglect the combined classification method[3] and use a single neural network that is successfully trained on a mixture of real and synthetic data. To improve the performance of detection and classification of road signs, we use stochastic averaging of weights[4]. The resulting recognition method shows a substantial increase in metrics on both rare and frequent signs.

2. Related works

2.1. Classifiers

Neural networks are widely used in the problem of image classification after the appearance of AlexNet [5]. Facing the problem of overfitting deep neural networks, the Resnet [6] uses skip connections to solve the problem of gradient vanishing. This method is still popular today and is a universal backbone for many computer vision tasks. Some works explore various modifications of Resnet. For example, ResNeXt [7] uses group convolutions, and WideResnet [8] reduces the depth, but increases the number of channels in the convolutions.

2.1.1. Classification of rare classes

In the work [3], a Random Forest Classifier is trained on top of neural network features to separate road signs into rare and frequent signs. The idea of the method is that the number of objects of rare classes is significantly less than the number of frequent ones, so rare classes can be considered as an anomaly among the frequent ones. Further, a neural network is used to classify frequent signs and KNN - for rare ones.

Another approach is few-shot learning, the field of machine learning when we need to classify new objects having only a couple of examples for supervised learning. Some methods are related to the topic of K-shot learning, Zero-shot learning. The task of zero-shot is if we know the general attributes of a particular class, without having pictures of representatives of the class, then we have to build a classifier or generate synthetics based on them. Thus, work [9] proposes method for generating image embeddings based on variational autoencoder, and [10], [11] combine image features and the semantic representation of class attributes into one embedding. In K-shot learning, we have several, but still few, instances of a class for neural network training. Metric learning and classes centroids are often used in such cases, as in [12]. However, this is sometimes not enough when we have only 1 picture for a rare class. In this case, [13] introduces architecture to press the image with noise closer to the center of the class.

2.1.2. Self-supervised representation pretraining

Modern methods of pretraining allow the network to learn better features without markup. However, in SimCLR [14], BYOL [15], SimCLRv2 [16], it is necessary to have large computing

resources to obtain a good neural network. SimCLRv2 [16] shows if there is much less marked-up data than not marked-up data, then it is better to take a larger model for pretraining. This facilitates learning features with a larger representation capacity, although such models are more susceptible to overfitting.

2.1.3. Transformers

Recently, autoregressive models gain popularity. Vision Transformer [17], VIVIT [18], DINO [19] explore their usage in computer vision tasks however for their training, it is also necessary to large computing resources and large datasets for their pretraining. Therefore, with comparable samples and computing resources, transformers have lower performance than their CNN counterpart.

2.2. Object detection

Most detection methods can be divided into 2 categories: single-stage or two-stage. Examples of single-step methods are YOLO [20], SSD [21], RetinaNet [22]. Two-stage detectors, on the other hand, have better recognition metrics than single-stage ones but they lose in terms of efficiency. Faster R-CNN [23] uses the Region Proposal Network for hypotheses about objects in the picture and then refines or rejects the prediction. Work [24] investigates the idea of looking several times at the image and proposes to increase the number of parameters of the backbone. Although deeper neural networks are more accurate, their usage becomes expensive in terms of computing resources. Therefore, some works explore the topic of quantization of neural networks [25] to make models more lightweight and less resource-intensive for computing.

3. Investigated methods of recognizing road signs

3.1. Classification

Improved WideResnet in [3] is WideResnet trained using contrastive loss [26]. For each pair of images of the same class, it aims to minimize the distance between them in the embedding space, and for pairs of different classes - to maximize. And after the neural network training stage, Random Forest Classifier is trained by images embeddings to determine a frequent sign or a rare one. And if the sign is rare, then to predict the class the KNN is used, else - softmax layer. But in this work, we additionally apply SWA [4] during training. After a certain number of epochs, we increase the learning rate several times and train yet more epochs than in stage 1. Then the final weights of the neural network are obtained by averaging some checkpoints from stage 2 and updating batch normalization statistics as in the original [4].

3.2. Detection

FCOS [27] with modifications ATSS [28] and Generalized Focal Loss [29] is used as the base detector. FCOS is an anchor-free detector that predicts not the coordinates of the corner and the parameters of the rectangle but for each point, the distance to the boundaries of the box bounding the object, and two upper layers have been added to its Feature Pyramid Network

Table 1

The performance of the classifier with the use of synthetics and SWA. In all cases, we use RTSD as a base real part of training data, but additional synthetic dataset we are changing in experiments. Synthetic datasets are taken from [1].

Synthetic Dataset	SWA	Accuracy all	nn-output		Accuracy all	combined method	
			Recall rare	Recall frequent		Recall rare	Recall frequent
Cycled	-	86.71	73.43	87.61	93.98	75.46	95.23
Cycled	+	97.35	79.22	98.58	96.69	79.72	97.84
Styled	-	85.96	70.41	87.01	94.11	76.33	95.31
Styled	+	97.17	77.19	98.52	96.85	76.70	98.22

[30]. This modification facilitates raising the performance of detecting objects in the case of objects occlusions. ATSS [28] proposes a new algorithm of matching anchors with ground truth objects with adaptive matching with respect to the IOU between anchors and ground truth using their average and standard deviations. GFL [29] tries to solve the problem of different treatments of the predictive branch of the localization quality in train and test stages and also the problem of objects occlusion.

4. Additional RTSD markup

During the experiments, we have found that there are additional signs in the data that are not presented in the markup. Therefore, we decide to additionally markup the dataset in semi-automatic mode. False alarms of the detector we analyze manually. New detections that contain an object we add into the markup. Then we train the detector on an updated dataset without classes with new road signs and manually analyze false detections again. After 3 iterations of this process, we haven't seen images with missed road signs. Figure 4 shows the results of comparing the original and new markup, Tables 4 and 5 show the new markup. The majority of new detections is either small signs or blurry ones. After updating markup for one-class detection task we label new road signs in the same manner. We train a classifier to predict classes and manually correct errors.

5. Experiments

Table 1 shows the results of a WideResnet-based classifier trained using contrastive loss. We train networks for 10 epochs at the first stage with a decreasing lr starting at 0.001, then we reset lr to 0.015 and train networks another 25 epochs. To average weights, we used every 5 checkpoints from 10 to 35 inclusive. With the use of SWA, the performance of classification increases very much both for rare signs and for frequent ones. The results show that it is possible to neglect the combined model with SWA training. Using only a neural network without a random forest classifier and KNN is much faster in terms of execution time.

Table 2 shows the results of detectors. To compare with the previous PVANet [31] architecture



Figure 1: Examples of frames with road signs that were present in the markup (green rectangles) and signs that were added in semi-automatic mode (red rectangles).

Table 2

The performance of detectors trained using NN-Additional synthetics

Detector	Backbone	AUC - all	AUC - rare	AUC - frequent
ATSS	Resnet50	90.39	90.01	90.73
GFL	Resnet50	90.48	90.09	90.79
ATSS	Resnext50_32x4d	90.28	90.06	90.71
GFL	Resnext50_32x4d	90.47	90.19	90.79
PVANet	Resnet50	89.17	86.62	89.31

we train ATSS [28], Generalized Focal Loss [29] using mmdetection [32]. We employ a standard training scheduler with 12 epochs and decreasing lr. The PVANet results we take from Road images augmentations [1].

Table 3 shows the results of the performance of the detector’s work with the classifier. We cut the detector’s predictions from images and forward them to the best classifier - WideResnet trained with Cycled synthetic dataset from [1] and SWA. Due to the high-quality improvement of the classifier, the joint performance of recognition, i.e. detection with classes task, on rare

Table 3

The performance of the classifier on top of the detector, trained using NN-Additional synthetics

Detector	Backbone	AUC - all	AUC - rare	AUC - frequent
ATSS	Resnet50	88.87	70.45	89.01
GFL	Resnet50	88.85	70.04	89.02
ATSS	Resnext50_32x4d	88.66	70.18	88.85
GFL	Resnext50_32x4d	88.82	70.43	88.98
PVANet	Resnet50	86.16	64.96	86.70

Table 4

New markup

	old		new	
	Images	Signs	Images	Signs
Train set	47378	79896	47356	84910
Test set	11650	25613	11672	29591

Table 5

New signs distribution

	old		new	
	Frequent signs	Rare signs	Frequent signs	Rare signs
Train set	79896	0	84910	0
Test set	23991	1622	27790	1801

signs has increased significantly.

6. Conclusion

In this work, we improve the markup of the Russian traffic signs dataset [2] in semi-automatic mode and add 9 thousand new objects, performing an experimental evaluation of the best classifiers and detectors at present in the task of recognizing road signs. By improving the classification performance using stochastic averaging of weights (SWA) and contrastive loss, we show that it is possible to neglect the non-neural part of the classifier [3] architecture and use only WideResnet. Modern synthetics allow us to improve not only the recognition quality of rare signs but also on frequent ones.

References

- [1] A. Konushin, B. Faizov, V. Shakhuro, Road images augmentation with synthetic traffic signs using neural networks, arXiv preprint arXiv:2101.04927 (2021).
- [2] V. Shakhuro, A. Konushin, Russian traffic sign images dataset, Computer Optics 40 (2016) 294–300. doi:10.18287/2412-6179-2016-40-2-294-300.

- [3] B. Faizov, V. Shakhuro, V. Sanzharov, A. Konushin, Classification of rare traffic signs, volume 44, 2020, pp. 236–243. doi:10.18287/2412-6179-CO-601.
- [4] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, A. G. Wilson, Averaging weights leads to wider optima and better generalization, arXiv preprint arXiv:1803.05407 (2018).
- [5] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012) 1097–1105.
- [6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [7] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.
- [8] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146 (2016).
- [9] A. Mishra, S. Krishna Reddy, A. Mittal, H. A. Murthy, A generative model for zero shot learning using conditional variational autoencoders, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2188–2196.
- [10] S. Liu, M. Long, J. Wang, M. I. Jordan, Generalized zero-shot learning with deep calibration network, in: Advances in Neural Information Processing Systems, 2018, pp. 2005–2015.
- [11] L. Liu, T. Zhou, G. Long, J. Jiang, X. Dong, C. Zhang, Isometric propagation network for generalized zero-shot learning, arXiv preprint arXiv:2102.02038 (2021).
- [12] J. Snell, K. Swersky, R. S. Zemel, Prototypical networks for few-shot learning, arXiv preprint arXiv:1703.05175 (2017).
- [13] W. Xue, W. Wang, One-shot image classification by learning to restore prototypes, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 6558–6565.
- [14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.
- [15] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, arXiv preprint arXiv:2006.07733 (2020).
- [16] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners, arXiv preprint arXiv:2006.10029 (2020).
- [17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
- [18] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, arXiv preprint arXiv:2103.15691 (2021).
- [19] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, arXiv preprint arXiv:2104.14294 (2021).
- [20] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

- [21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.
- [22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
- [23] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv:1506.01497 (2015).
- [24] S. Qiao, L.-C. Chen, A. Yuille, Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, arXiv preprint arXiv:2006.02334 (2020).
- [25] B. Zhuang, L. Liu, M. Tan, C. Shen, I. Reid, Training quantized neural networks with a full-precision auxiliary module, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1488–1497.
- [26] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), volume 1, IEEE, 2005, pp. 539–546.
- [27] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.
- [28] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9759–9768.
- [29] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, arXiv preprint arXiv:2006.04388 (2020).
- [30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
- [31] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, M. Park, Pvanet: Deep but lightweight neural networks for real-time object detection, arXiv preprint arXiv:1608.08021 (2016).
- [32] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, D. Lin, MMDetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).