ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Real-time capable Ensemble Estimation for 2D Object Detection

Lukas Enderich and Simon Heming

Robert Bosch GmbH - Computer Vision Research
Robert-Bosch-Straße 200, 31139 Hildesheim, Germany

Abstract. Deep neural networks tend to make overconfident predictions. Although ensemble methods improve the predictive performance by producing better calibrated confidences, they are computationally expensive. Thus, we propose a real-time capable ensemble method for object detection that significantly improves the performance with only a minor increase in runtime. Our method diversifies the prediction of the class probabilities on the anchor space using multiple classification heads. A regularization further increases the diversity of the heads, making ensemble distillation unnecessary. On the KITTI benchmark dataset, our approach increases the mean average precision of an SSD based network from **0.58** to **0.71**.

1 Introduction

Deep neural networks (DNNs) are state-of-the-art in many machine learning challenges, outperforming classical methods in computer vision, object detection, and speech recognition [1, 2]. However, there are still a number of problems when DNNs are used in real-time and safety-critical systems. This includes, for instance, their ability to generalize, as DNNs poorly quantify uncertainty and tend to make overconfident predictions [3, 4].

Ensemble methods improve the predictive performance of DNNs, providing better calibrated confidences by averaging over the ensemble output [4]. Ensemble methods can for example be based on a certain number of independently trained networks [5] or a monte-carlo dropout model [6]. However, these methods significantly increase the computational complexity during inference since multiple forward passes need to be calculated.

Therefore, recent work has focused on distilling both the diversity and the knowledge of an ensemble into a single network [3]. The model used in [3] consists of one network body and multiple networks heads. Thus, it is trained to approximate the predictions of each ensemble member with its corresponding network head. Since the body is shared among all heads, the computational complexity is significantly lower compared to the distilled ensemble.

However, the setup used in [3] only works for image classification. Furthermore, training and distilling an ensemble is time-consuming, especially for more complex tasks such as object detection. Therefore, we transfer the multi-head approach to object detection by making two major contributions:

- We show how to efficiently diversify an anchor-based object detector by using multiple classification heads. In this way, different class probabilities are predicted for each anchor that can be averaged during inference.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

- We use suitable regularization to increase the diversity of the classification heads during training. This replaces the distillation of a pretrained ensemble, which significantly reduces the required training time.

2 Related Work

Ensemble methods use a number of different models to estimate uncertainty based on the variance among predictions: The more the outputs differ at inference time, the more the input is expected to be outside the generalization area [4]. Furthermore, averaging the ensemble predictions increases the predictive performance compared to a single network of the same size [5, 6]. In fact, the higher the variance among the predictions, the more balanced the averaged class scores. Since ensemble methods have high computational and memory requirements, recent work tried to make them more efficient. This can be done either by distilling the diversity of an ensemble into a network with multiple heads [3], or by learning the conditional predictive distribution of an ensemble [7, 8].

Unlike image classification, **object detection** predicts both class and location [9]. Two-stage detectors first make suggestions about object locations using predefined anchors [10, 11]. These locations are then projected into the feature space to predict the corresponding object class. In contrast, single-stage detectors predict class scores for each object category in each anchor, including the background class [12]. Such networks consist of one feature generating backbone, one classification head to predict the class probabilities for each predefined anchor, and one localization head to predict the bounding box regression for each anchor. Both heads consist of multiple layers to process feature maps of different resolution [12]. Afterwards, the anchors are processed into objects.

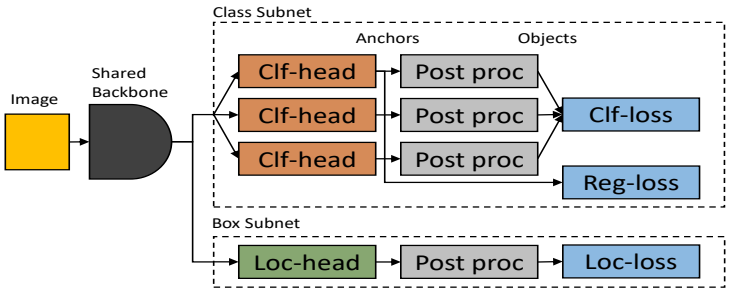
Single-stage detectors are computationally less complex than two-stage detectors. Since our approach emphasizes real-time capability, we use the widely available Single Shot MultiBox Detector (SSD) for our experimental setup [12]. However, our approach can be applied to **any** anchor-based detector.

3 Model

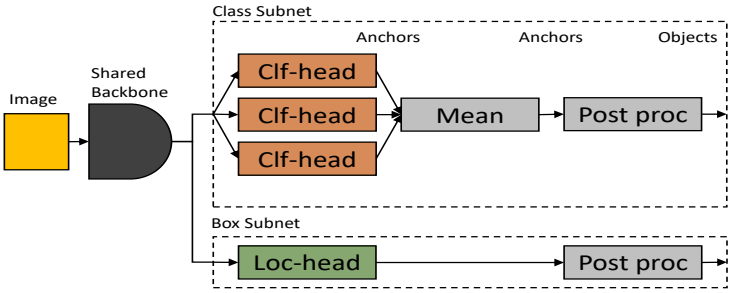
Since object detectors like the SSD predict both the class and the location of objects, uncertainties can be modeled for both types of prediction. Given that misclassifications and undetected objects (such as covered pedestrians) are the most serious problems in real-world applications [9], we focus on the diversification of the predicted class probabilities by introducing multiple independent classification heads.

The upper part of figure 1 shows the **training graph** of our multi-head SSD, exemplified with three classification heads. Each head uses the same features generated by the shared backbone, but makes its own class predictions for each predefined anchor. As in the case of an ensemble, the heads are trained independently, getting their own classification losses to allow diversity. Additionally, regularization is used to further increase the diversity among the classification

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.



(a) Training graph.



(b) Inference graph.

Fig. 1: Training and inference graph of our multi-head detector, exemplified with three classification heads. For each predefined anchor, each head makes its own predictions regarding the expected class probabilities. To increase the diversity, the heads are trained independently, using an additional diversity regularization. During inference, the mean of the predicted class probabilities is calculated for each anchor, which results in better calibrated confidences, causing more objects to be detected by post processing.

heads, replacing the distillation of a pretrained ensemble. Such diversity regularization can be done for example by using Negative Correlation [13], which can be applied to the classification heads as follows:

$$\text{Regularization} = -\frac{1}{N}\sum_{n=1}^{N}(y_n - \overline{y})^2\,, \quad \text{with} \quad \overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n\,.$$

Here, N is the number of classification heads, y_n the output of head n, and $\overline{y}$ is mean output over all heads. Consequently, the training loss aggregates to:

$$\text{Loss} = \frac{1}{N}\sum_{n=1}^{N}\text{Cls-Loss}_n + \text{Loc-Loss} + \lambda\,\text{Regularization}\,,$$

where Cls-Loss_n is the classification loss of head n, Loc-Loss is the localization

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Model	mAP	AP Person	AP car	AP truck	Runtime
SSD	0.58	0.29	0.73	0.72	100%
3-Head SSD	0.66	0.38	0.79	0.80	118%
5-Head SSD	0.71	0.42	0.85	0.84	140%

Table 1: Mean average precision (mAP), class-dependent average prevision (AP), and runtime for a different number of classification heads.

loss, and λ is the regularization parameter. The higher λ, the higher the expected diversity among the classification heads.

The lower part of figure 1 shows the **inference graph** of our multi-head SSD. Here it is important to calculate the mean value over the predictions of the classification heads before processing the anchors into objects. Otherwise, it would no longer be possible to assign the predicted objects of the different heads to each other without applying clustering methods. In contrast, calculating the mean increases the predictive performance on the anchor level, allowing more objects to be detected. Thus, the whole post processing remains the same. In fact, since we only diversify the class predictions on the anchor space - by using a shared backbone and by calculating the mean over the predictions of the classification heads - the computational overhead is very low, especially compared to a Deep Ensemble [5] an MC-dropout model [6].

4 Experiments

We test our multi-head SSD on KITTI, a well-known benchmark dataset for 2D object detection [14]. It consists of 7485 RGB images divided into training, validation, and test data. The labels contain four classes: person, car, truck, and background. We train for 50k iterations, using the SGD optimizer with the Nesterov Momentum set to 0.9 and a batch size of 32. The initial learning rate is 10^{-4}, which is divided by 10 after 30k and 40k iterations. Furthermore, we apply random crops, mirroring, and photometric distortions. For our multi-head models, the regularization parameter λ is set to 10^{-2}, which was found by an ablation study (see table 2).

First, we compare our approach to the standard SSD architecture. Table 1 shows the results with a different number of classification heads. The standard SSD has a mean average precision (mAP) of 0.58, with the average precision (AP) for persons (0.29) being significantly worse than for cars (0.73) and trucks (0.70). The runtime of the standard SSD is used as a reference point at 100%.

Compared to this, our 3-Head SSD increases the mAP by 14% from 0.58 to 0.66. The highest improvement is for the person class, where the performance increases significantly by 31% from 0.29 to 0.38. This can also be seen in figure 2, which shows a test image containing a lot of pedestrians. The upper part shows the persons detected by the standard SSD. Here it is noticeable that both the covered persons (e.g., behind the bicycles) and the persons located further away from the camera are poorly detected. As can be seen in the lower part

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Model	λ	mAP	AP Person	AP car	AP truck
3-Head SSD	0	0.59	0.31	0.73	0.70
3-Head SSD	10^{-3}	0.63	0.33	0.77	0.79
3-Head SSD	10^{-2}	0.66	0.38	0.79	0.80
3-Head SSD	10^{-1}	0.64	0.34	0.78	0.79

Table 2: Performance of our 3-Head SSD with different regularization values. The higher λ, the higher the diversity among the classification heads.

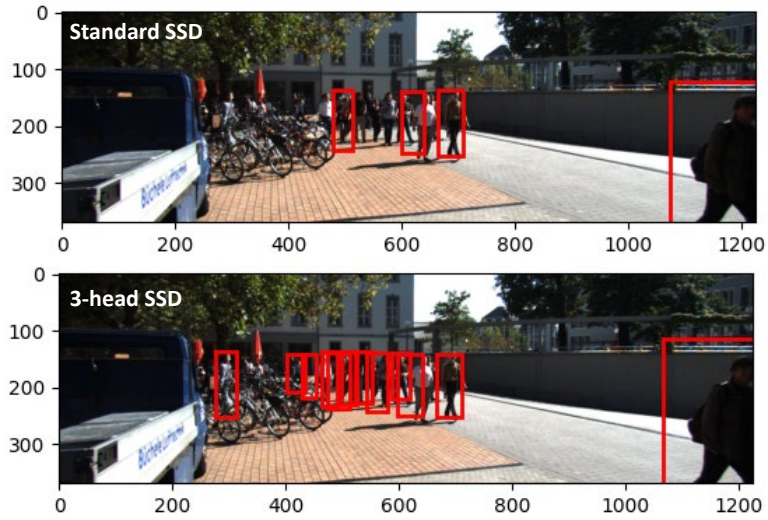


Fig. 2: Compared to the standard SSD, our 3-Head SSD detects significantly more pedestrians in the image, including the ones covered by bicycles.

of figure 2, our 3-Head SSD detects significantly more persons, including the ones located behind the bicycles. Relative to the standard SSD, the runtime increases only slightly by 18%. In comparison, a Deep Ensemble consisting of three independent models increases the runtime by 200%.

Compared to our 3-Head SSD, our 5-Head SSD further increases the AP for each class. The mAP improves from 0.66 to 0.71 and is thus 22% higher compared to the standard SSD. Here, the AP for persons improves by 44% from 0.29 to 0.42. Comparing the runtime of the three networks, it is noticeable that it increases by about 10% for each classification head that is added.

Second, we make an ablation study on the impact of our diversity regularization. Here we claim that using the Negative Correlation regularization increases the diversity of the classification heads, improving the predictive performance at least up to a certain point. Thus, table 2 shows the results of our 3-Head SSD trained with different values of the regularization parameter. The following observations can be made: **1.)** Without regularization, our model is only slightly better than the standard SSD (0.58 vs. 0.59 mAP), **2.)** Increasing λ up

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

to 10^{-2} improves the mAP significantly from 0.59 to 0.66, and **3.)** If λ is too high, the performance drops again. Consequently, the diversity regularization is an essential component and it confirms that a higher diversity among the heads is able to increase the predictive performance.

5 Conclusion

In this paper, we propose an effective and easy-to-implement ensemble method for 2D object detection. Our model uses multiple independent classification heads to combine different predictions for the class probabilities of the anchor space. Furthermore, we apply Negative Correlation regularization to increase the diversity among the classification heads, making ensemble distillation unnecessary. In this way, our method significantly improves the mean average precision of an SSD detector from **0.58** to **0.66** with only **18%** increase in runtime.

References

[1] L. Deng and D. Yu, Deep learning: Methods and applications, *Foundations and Trends in Signal Processing*, 7(34):197?387, 2014.

[2] Y. Lecun, Y. Bengio, and G. Hinton, Deep learning, *Nature*, 521(7553):436?444, 2015.

[3] L. Tran, B. Veeling, K. Roth, J. ?wi?tkowski, J. Dillon, J. Snoek, S. Mandt, T. Salimans, S. Nowozin, R. Jenatton, Hydra: Preserving Ensemble Diversity for Model Distillation, 2020.

[4] J. Lust and A. Condurache, A Survey on Assessing the Generalization Envelope of Deep Neural Networks: Predictive Uncertainty, Out-of-distribution and Adversarial Samples, 2021.

[5] B. Lakshminarayanan, Alexander Pritzel, and C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, *Advances in neural information processing systems*, 6402?6413, 2017.

[6] Y. Gal and Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, *International conference on machine learning*, 1050?1059, 2016.

[7] A. Malinin, B. Mlodozeniec, M. Gales, Ensemble Distribution Distillation, 2019.

[8] Y. Shen, Z. Zhang, Mert R. Sabuncu, L. Sun, Real-Time Uncertainty Estimation in Computer Vision via Uncertainty-Aware Distribution Distillation, *Winter Conference on Applications of Computer Vision*, 707-716, 2021.

[9] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges, *Transactions on Intelligent Transportation Systems*, 1341-1360, 2021.

[10] R. Girshick, Fast R-CNN, *International Conference on Computer Vision (ICCV)*, 2015.

[11] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Advances in Neural Information Processing Systems 28*, 2015.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg, SSD: Single Shot MultiBox Detector, 2015.

[13] C. Shui, A. Mozafari, J. Marek, I. Hedhli, C. Gagne, Diversity Regularization in Ensembles, *International Conference on Learning Representations Workshop*, 2018.

[14] A. Geiger, P. Lenz, R Urtasun, Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suit, *Conference on Computer Vision and Pattern Recognition*, 2012.