Proposal Learning for Semi-Supervised Object Detection

Peng Tang[†] Chetan Ramaiah[†] Yan Wang[‡] Ran Xu[†] Caiming Xiong[†]
[†]Salesforce Research [‡]The Johns Hopkins University
{peng.tang,cramaiah,ran.xu,cxiong}@salesforce.com wyanny.9@gmail.com

Abstract

In this paper, we focus on semi-supervised object detection to boost performance of proposal-based object detectors (a.k.a. two-stage object detectors) by training on both labeled and unlabeled data. However, it is non-trivial to train object detectors on unlabeled data due to the unavailability of ground truth labels. To address this problem, we present a proposal learning approach to learn proposal features and predictions from both labeled and unlabeled data. The approach consists of a self-supervised proposal learning module and a consistency-based proposal learning module. In the self-supervised proposal learning module, we present a proposal location loss and a contrastive loss to learn context-aware and noise-robust proposal features respectively. In the consistency-based proposal learning module, we apply consistency losses to both bounding box classification and regression predictions of proposals to learn noise-robust proposal features and predictions. Our approach enjoys the following benefits: 1) encouraging more context information to delivered in the proposals learning procedure; 2) noisy proposal features and enforcing consistency to allow noise-robust object detection; 3) building a general and high-performance semi-supervised object detection framework, which can be easily adapted to proposal-based object detectors with different backbone architectures. Experiments are conducted on the COCO dataset with all available labeled and unlabeled data. Results demonstrate that our approach consistently improves the performance of fully-supervised baselines. In particular, after combining with data distillation [20], our approach improves AP by about 2.0% and 0.9% on average compared to fully-supervised baselines and data distillation baselines respectively.

1. Introduction

With the giant success of Convolutional Neural Networks (CNNs) [13, 14], great leap forwards have been achieved in object detection [7, 8, 21]. However, training accurate object detectors relies on the availability of large scale labeled datasets [5, 16], which are very expensive and time-consuming to collect. In addition, training object detectors only on the labeled datasets may limit their detection performance. By contrast, considering that acquiring unlabeled data is much easier than collecting labeled data, it is important to explore approaches for the Semi-Supervised Object Detection (SSOD) problem, *i.e.*, training object detectors on both labeled and unlabeled data, to boost performance of current state-of-the-art object detectors.

In this paper, we focus on SSOD for proposal-based object detectors (a.k.a. two-stage object detectors) [7, 8, 21] due to their high performance. Proposal-based object detectors detect objects by 1) first generating region proposals that may contain objects and 2) then generating proposal features and predictions (*i.e.*, bounding box classification and regression predictions). Specially, we aim to improve the second stage by learning proposal features and predictions from both labeled and unlabeled data.

For labeled data, it is straightforward to use ground truth labels to get training supervisions. But for unlabeled data, due to the unavailability of ground truth labels, we cannot learn proposal features and predictions directly. To address this problem, apart from the standard fully-supervised learning for labeled data [21] shown in Fig. 1, we present an approach named proposal learning, which consists of a self-supervised proposal learning module and a consistency-based proposal learning module, to learn proposal features and predictions from both labeled and unlabeled data, see Fig. 1.

Recently, self-supervised learning has shown its efficacy to learn features from unlabeled data by defining some pretext tasks [4, 9, 25]. Our self-supervised proposal learning module uses the same strategy of defining pretext tasks, inspired by the facts that context is important for object detection [3, 12] and object detectors should be noise-robust [17, 24]. More precisely, a proposal location loss and a contrastive loss are presented to learn context-aware and noise-robust proposal features respectively. Specifically, the proposal location loss uses proposal location prediction as a pretext task to supervise training, where a small neural network are attached after proposal features for proposal

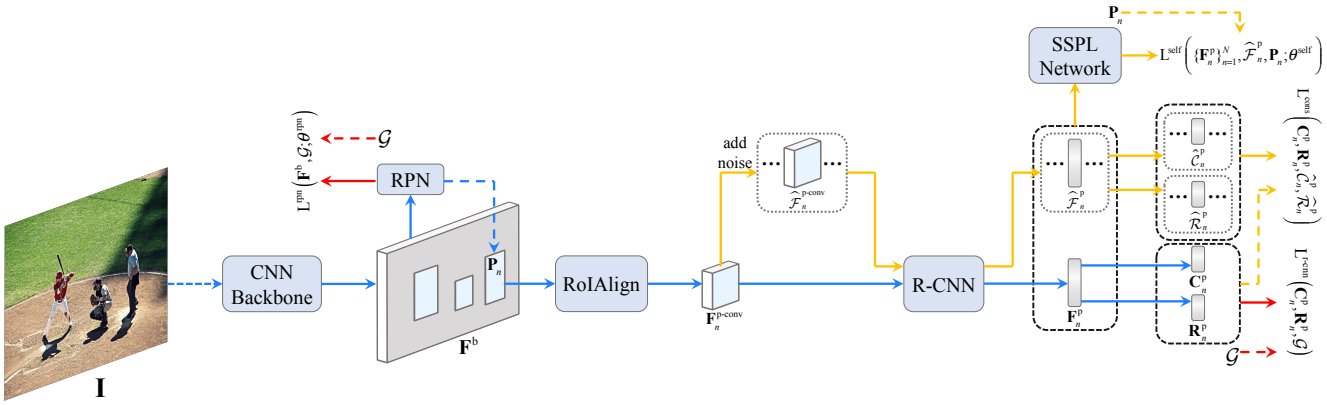

Figure 1. The overall framework of our proposal learning approach. All arrows have forward computations during training, only the solid ones have back-propagation computations, the red ones are only utilized for labeled data, and only the blue ones are utilized during inference. "RPN": Region Proposal Network; "R-CNN": Region-based CNN; "SSPL": Self-Supervised Proposal Learning; $\mathbf{I}$: input image; "$\mathbf{F}^{\text{b}}$": image convolutional feature maps; "$\boldsymbol{\theta}^{\text{rpn}}$": parameters of the RPN; "$\mathbf{P}_n$": a proposal with its location; "$\mathbf{F}_n^{\text{p-conv}}$" and "$\hat{\mathcal{F}}_n^{\text{p-conv}}$": the original and noisy convolutional feature maps of $\mathbf{P}_n$; "$\mathbf{F}_n^{\text{p}}$" and "$\hat{\mathcal{F}}_n^{\text{p}}$": the original and noisy features of $\mathbf{P}_n$; "$\mathbf{C}_n^{\text{p}}, \mathbf{R}_n^{\text{p}}$" and "$\hat{\mathcal{C}}_n^{\text{p}}, \hat{\mathcal{R}}_n^{\text{p}}$": the original and noisy predictions (bounding box classification and regression predictions) of $\mathbf{P}_n$; "$\boldsymbol{\theta}^{\text{self}}$": parameters of the SSPL network; "$\mathcal{G}$": ground truth labels; "$\text{L}^{\text{rpn}}\left(\mathbf{F}^{\text{b}}, \mathcal{G}; \boldsymbol{\theta}^{\text{rpn}}\right)$": RPN loss; "$\text{L}^{\text{r-cnn}}\left(\mathbf{C}_n^{\text{p}}, \mathbf{R}_n^{\text{p}}, \mathcal{G}\right)$": R-CNN loss; "$\text{L}^{\text{self}}\left(\{\mathbf{F}_n^{\text{p}}\}_{n=1}^{N}, \hat{\mathcal{F}}_n^{\text{p}}, \mathbf{P}_n; \boldsymbol{\theta}^{\text{self}}\right)$": SSPL loss; "$\text{L}^{\text{cons}}\left(\mathbf{C}_n^{\text{p}}, \mathbf{R}_n^{\text{p}}, \hat{\mathcal{C}}_n^{\text{p}}, \hat{\mathcal{R}}_n^{\text{p}}\right)$": consistency-based proposal learning loss.

location prediction. This loss helps learn context-aware proposal features, because proposal location prediction requires proposal features understanding some global image information. At the same time, the contrastive loss learns noise-robust proposal features by a simple instance discrimination task [9, 25], which ensures that noisy proposals features are closer to their original proposal features than to other proposal features. In particular, instead of adding noise to images to compute contrastive loss [9, 28], we add noise to proposal features, which shares convolutional feature computations for the entire image between noisy proposal feature computations and the original proposal feature computations for training efficiency [7].

To further train noise-robust object detectors, our consistency-based proposal learning module uses consistency losses to ensure that predictions from noisy proposal features and their original proposal features are consistent. More precisely, similar to consistency losses for semi-supervised image classification [18, 22, 26], a consistency loss for bounding box classification predictions enforces class predictions from noisy proposal features and their original proposal features to be consistent. In addition, a consistency loss for bounding box regression predictions enforces object location predictions from noisy proposal features and their original proposal features also to be consistent. With these two consistency losses, proposal features and predictions are robust to noise.

We apply our approach to Faster R-CNN [21] with feature pyramid networks [15] and RoIAlign [10], using different CNN backbones, where our proposal learning modules are applied to both labeled and unlabeled data, as shown in Fig. 1. We conduct elaborate experiments on the challenging COCO dataset [16] with all available labeled and unlabeled data, showing that our approach outperforms fully-supervised baselines consistently. In particular, when combining with data distillation [20], our approach obtains about 2.0% and 0.9% absolute AP improvements on average compared to fully-supervised baselines and data distillation based baselines respectively.

2. Experiments

2.1. Experimental Setup

We evaluate our approach on the challenging COCO dataset [16]. The COCO dataset contains more than 200K images for 80 object classes. Unlike many semi-supervised object detection works conducting experiments on a simulated setting by splitting a fully annotated dataset into labeled and unlabeled subsets, we use all available labeled and unlabeled training data in COCO as our labeled and unlabeled data respectively, following [20]. More precisely, we use the COCO `train2017` set (118K images) as labeled data and the COCO `unlabeled2017` set (123K images) as unlabeled data to train object detectors. In addition, we use the COCO `test-dev2017` set (20K images) for testing.

We use the standard COCO criterion as our evaluation metrics, including AP (averaged average precision over different IoU thresholds, the primary evaluation metric of COCO), AP_{50} (average precision for IoU threshold 0.5), AP_{75} (average precision for IoU threshold 0.75), AP_S (AP

for small objects), AP_M (AP for medium objects), AP_L (AP for large objects).

In our experiments, we choose Faster R-CNN [21] with feature pyramid networks [15] and RoIAlign [10] as our proposal-based object detectors, which are the foundation of many recent state-of-the-art object detectors. Different CNN backbones, including ResNet-50 [11], ResNet-101 [11], ResNeXt-101-32×4d [27], and ResNeXt-101-32×4d with Deformable ConvNets [29] (ResNeXt-101-32×4d+DCN), are chosen.

We train object detectors on 8 NVIDIA Tesla V100 GPUs for 24 epochs, using stochastic gradient descent with momentum 0.9 and weight decay 0.0001. During each training mini-batch, we randomly sample one labeled image and one unlabeled image for each GPU, and thus the effective mini-batch size is 16. Learning rate is set to 0.01 and is divided by 10 at the 16^{th} and 22^{nd} epochs. We use linear learning rate warm up to increase learning rate from 0.01/3 to 0.01 linearly in the first 500 training iterations. In addition, object detectors are first trained only on labeled data for 6 epochs. We also use fast stochastic weight averaging for checkpoints from the last few epochs for higher performance, following [1].

We add two types of noise, DropBlock [6] with block size 2 and SpatialDropout [23] with dropout ratio 1/64, to proposal convolutional feature maps. Images are resized so that the shorter side is 800 pixels with/without random horizontal flipping for training/testing. Considering that most of the proposals mainly contain backgrounds, we only choose the positive proposals for labeled data and the proposals having maximum object score larger than 0.5 for unlabeled data to compute proposal learning based losses, which ensures networks focusing more on objects than backgrounds.

Our experiments are implemented based on the PyTorch [19] deep learning framework and the MMDetection [2] toolbox.

2.2. Main Results

We report the result comparisons among fully-supervised baselines, Data Distillation (DD) [20], and our approach in Table 1 on the COCO `test-dev2017` set. As we can see, our approach obtains consistently better results compared to the fully-supervised baselines for different CNN backbones. In addition, both DD and our approach obtain higher APs than the fully-supervised baselines, which demonstrates that training object detectors on both labeled and unlabeled data outperforms training object detectors only on labeled data, confirming the potentials of semi-supervised object detection. Using our approach alone obtains comparable APs compared to DD.

In particular, we also evaluate the efficiency of our method by combining with DD. More specifically, we first train object detectors using our approach, then follow DD to label unlabeled data, and finally re-train object detectors using both fully-supervised loss and proposal learning losses. The combination of our approach and DD obtains 39.6% (ResNet-50), 41.3% (ResNet-101), 42.8% (ResNeXt-101-32×4d), and 46.2% (ResNeXt-101-32×4d+DCN) APs, which outperforms the fully-supervised baselines by about 2.0% on average and DD alone by 0.9% on average. The results demonstrate that our approach and DD are complementary to some extent.

Table 1. Experimental result comparisons among fully-supervised baselines (no "✓"), Data Distillation (DD) [20], and our approach (Ours) on the COCO `test-dev2017` set. Different CNN backbones are chosen. Results of DD are reproduced by ourselves and are comparable with or even better than the results reported in the original DD paper.

CNN backbone	DD	Ours	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50			37.7	59.6	40.8	21.6	40.6	47.2
	✓		38.5	60.4	41.7	22.5	41.9	47.4
		✓	38.6	60.2	41.9	21.9	41.4	48.9
	✓	✓	**39.6**	**61.5**	**43.2**	**22.9**	**42.9**	**49.8**
ResNet-101			39.6	61.2	43.2	22.2	43.0	50.3
	✓		40.6	62.2	44.3	23.2	44.4	50.9
		✓	40.4	61.8	44.2	22.6	43.6	51.6
	✓	✓	**41.3**	**63.1**	**45.3**	**23.4**	**45.0**	**52.7**
ResNeXt-101-32×4d			40.7	62.3	44.3	23.2	43.9	51.6
	✓		41.8	63.6	45.6	24.5	45.4	52.4
		✓	41.5	63.2	45.4	23.9	44.8	52.8
	✓	✓	**42.8**	**64.4**	**46.9**	**24.9**	**46.4**	**54.4**
ResNeXt-101-32×4d+DCN			44.1	66.0	48.2	25.7	47.3	56.3
	✓		45.4	67.0	49.6	27.3	49.0	57.9
		✓	45.1	66.8	49.2	26.4	48.3	57.6
	✓	✓	**46.2**	**67.7**	**50.4**	**27.6**	**49.6**	**59.1**

3. Conclusion

In this paper, we focus on semi-supervised object detection for proposal-based object detectors (a.k.a. two-stage object detectors). To this end, we present a proposal learning approach, which consists of a self-supervised proposal learning module and a consistency-based proposal learning module, to learn proposal features and predictions from both labeled and unlabeled data. The self-supervised proposal learning module learns context-aware and noise-robust proposal features by a proposal location loss and a contrastive loss respectively. The consistency-based proposal learning module learns noise-robust proposal features and predictions by consistency losses for both bounding box classification and regression predictions. Experimental results show that our approach outperforms fully-supervised baselines consistently. It is also worth mentioning that we can further boost detection performance by combining our approach and data distillation. In the future, we will explore more self-supervised learning and semi-supervised learning ways for semi-supervised object detection, and explore how to apply our approach to semi-supervised instance segmentation.

References

[1] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. *arXiv preprint arXiv:1806.05594*, 2018. 3

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[3] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271–1278, 2009. 1

[4] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 1

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1

[6] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *NeurIPS*, pages 10727–10737, 2018. 3

[7] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 1, 2

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 1

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 1, 2

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 2, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[12] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018. 1

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 1

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 2, 3

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 2

[17] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 1

[18] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 41(8):1979–1993, 2018. 2

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 3

[20] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, pages 4119–4128, 2018. 1, 2, 3

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2017. 1, 2, 3

[22] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, pages 1163–1171, 2016. 2

[23] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *CVPR*, pages 648–656, 2015. 3

[24] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-Fast-RCNN: Hard positive generation via adversary for object detection. In *CVPR*, pages 2606–2615, 2017. 1

[25] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 1, 2

[26] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2019. 2

[27] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 3

[28] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pages 6210–6219, 2019. 2

[29] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable ConvNets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 3