An Analysis of Pre-Training on Object Detection

Hengduo Li* Bharat Singh Mahyar Najibi Zuxuan Wu Larry S. Davis
University of Maryland, College Park

Abstract

We provide a detailed analysis of network pre-training on the task of object detection. To this end, we train detectors on large datasets like OpenImagesV4, ImageNet Localization and COCO. We analyze how well their features generalize to tasks including image classification, semantic segmentation and object detection on small datasets. Some important conclusions from our analysis include — 1) Pre-training on large detection datasets is crucial for fine-tuning on small detection datasets, especially when precise localization is needed. 2) Detection pre-training also benefits other localization tasks like semantic segmentation but adversely affects image classification. 3) Features for images which are similar in the object detection feature space are likely to be similar in the image classification feature space but the converse is not true. 4) Visualization of features reveals that activation of detection networks typically covers the entire object, while activation of classification networks focus on parts. Therefore, detection networks are poor at classification when multiple instances are present in an image or when an instance only covers a small fraction of an image.

1. Introduction

For several computer vision problems like object detection, image segmentation and image classification, pre-training on large scale datasets is common [30, 13, 9], since it leads to better results and faster convergence [60, 22, 9, 33, 17]. However, the effect of pre-training in computer vision is often evaluated by training networks for the task of *image classification*, on datasets like ImageNet [8], Places [60], JFT [48], Instagram [33] *etc*., but rarely for object detection. It can be argued that the task of object detection subsumes image classification, so a network good at object detection should learn richer features than one trained for classification. After all, this network has access to an orthogonal semantic information, like the spatial extent of an object. However, it can also be argued that forcing a network to learn position sensitive information may affect its spatial invariance properties which help in recognition. To this end, we provide a comprehensive analysis and compare network pre-training on object detection and image classification.

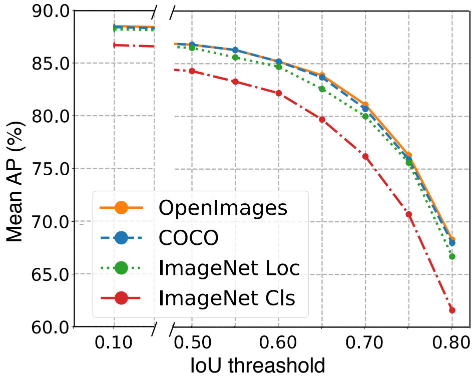


Figure 1: Detection performance (mAP %) at different IoUs on PASCAL-VOC 2007 [11] test set of detectors pre-trained on different datasets.

We pre-train the network on OpenImagesV4 [25] (OPENIMAGES) dataset on the object detection task and fine-tune it on tasks like semantic segmentation, object detection and classification on datasets like PASCAL-VOC [11], COCO [28], CALTECH-256 [14], SUN-397 [51] and OXFORD-102 FLOWERS [36]. For a stronger evaluation, we also pre-train on the ImageNet classification dataset [8] with bounding-box annotations on 3,130 classes [45] (IMAGENET-LOC, as opposed to IMAGENET-CLS for ImageNet Classification dataset without bounding boxes) and the COCO dataset [28] which helps us in evaluating the importance of the number of training samples. We then design careful experiments to understand the differences in properties of features which emerge by pre-training on detection *vs*. classification.

Our experimental analysis reveals that pre-training on object detection can improve performance by more than 5% on PASCAL-VOC for object detection (especially at high IoUs) and 3% for semantic segmentation. However, detection features are significantly worse at performing classification compared to features from IMAGENET-CLS pre-trained networks (∼ 8% on CALTECH-256). We also find

*hdli@umd.edu

Method / Pre-trained Dataset	mAP@**0.5**	mAP@**0.7**
DCNv1 [7]	81.9	68.2
DCNv2 [62]	84.9	73.5
IMAGENET-CLS [8]	84.6	76.3
IMAGENET-LOC [8, 45]	86.5	80.0
COCO [28]	86.8	80.7
OPENIMAGES [25]	**86.8**	**81.1**

Table 1: Baseline and our results on PASCAL-VOC 2007 [11] object detection dataset.

that if features (like average pooled Conv5) are similar in the object detection feature space, they are likely to be similar in the image classification feature space, but the converse is not true. Visualization of activations for object detection shows that they often cover the entire extent of an object, so are poor at recognition when an object is present in a small part of an image or when multiple instances are present.

2. Analysis

We perform pre-training on multiple detection datasets and compare it with IMAGENET-CLS pre-training for different computer vision tasks like object detection, image classification and semantic segmentation. For detection pre-training, our experimental setup is as follows. All our detection networks are pre-trained first on IMAGENET-CLS if not mentioned otherwise. They are then trained on detection datasets like OPENIMAGES [25], IMAGENET-LOC [8, 45] and COCO [28]. The SNIPER [46] detector is trained on all the datasets. We use multiple pre-training datasets for two reasons - 1) To thoroughly evaluate our claims about pre-training for the detection task 2) Since the datasets contain different number of classes and training examples, it also provides an indication of the magnitude of improvement one can expect by pre-training on detection datasets of different sizes.

Datasets For the *object detection* task, we fine-tune on the PASCAL-VOC dataset [11]. We use the VOC 07+12 trainval set for training and the VOC 07 test set for evaluation. For the *semantic segmentation* task, we follow [7, 16, 31, 3] and use VOC 2012 plus additional annotations provided in [15]. For *image classification*, we fine-tune on CALTECH-256 [14], SUN-397 [51] and OXFORD-102 FLOWERS [36]. We use the official split of trainval and test sets for CALTECH-256 and OXFORD-102 FLOWERS; for SUN-397 we follow [22] and use the first split for training and evaluation.

Architecture We briefly describe the architecture of the detection heads. On OPENIMAGES detector after Conv5 (2048,14,14), *i.e.* the last layer of the ResNet backbone before FC layers, we have the following layers: ConvProj (256,14,14), FC1 (1024), FC2 (1024), Output (501), Regression (4). The same architecture is used for the COCO detector, except that the Output layer is 81-dimensional. For the IMAGENET-LOC detector, the architecture is the same as described in [45].

2.1. Object Detection

Baseline Configuration and Results For our object detection experiments, we train our detectors (SNIPER with ResNet-101) on 3 datasets: OPENIMAGES, COCO and IMAGENET-LOC. Our OPENIMAGES model obtains 45% mAP (at 0.5 overlap) on the validation set. It is trained at 2 scales, (480, 512) and (768, 1024) without negative chip mining. Inference is also performed at these two scales only. For the COCO model, training and inference is performed at 3 scales (480, 512), (800, 1280) and (1400,2000) and the detector obtains an mAP of 46.1% (COCO metric) on the test-dev set. The IMAGENET-LOC model obtains 37.4% mAP (at 0.5 overlap) on the ImageNet *Detection* dataset (not IMAGENET-LOC). This detector was only trained at a single scale of (512, 512) on IMAGENET-LOC without any negative chip mining. Inference is also performed only at a scale of (512, 512) as compared to others, this dataset contains relatively bigger objects.

Fine-tuning on PASCAL-VOC We fine-tune these pre-trained models on PASCAL-VOC [11] using the same set of scales as COCO for both training and inference. Detection heads of the models pre-trained on detection datasets are re-initialized before fine-tuning. Training is performed for 7 epochs with learning rate step-down at the end of epoch 5. Horizontal flipping is used for data augmentation. The results are shown in Table 1.

Pre-training helps at Higher IoU As shown in Table 1, our baseline network pre-trained on IMAGENET-CLS obtains an mAP of 76.3% at 0.7 overlap, while the OPENIMAGES, COCO and IMAGENET-LOC models obtain 81.1%, 80.7% and 80% mAP, improving performance on PASCAL by as much as 4.8%. However, such large improvements do not translate to lower overlap thresholds. For example, the difference in mAP between IMAGENET-CLS and the OPENIMAGES model at an overlap of 0.5 is only 2.2%. We plot the mAP for all the detection models at different overlap thresholds in Fig 1. This shows that pre-training for detection helps to a large extent in improving localization performance. We also observe this phenomenon on the COCO dataset: when OPENIMAGES pre-training is used, the performance at 0.5 improves by 0.7%, but results at 0.75 improve by 1.4%.

Relationship between dataset size and performance improvement Another pattern we observe is that the number of samples in the pre-training dataset did not affect the fine-tuning performance to a large extent. The important factor was whether the network was pre-trained on a reasonably large detection dataset ($>$ 1M training instances) or not.

Pre-training Improves General Localization Ability Despite the performance improvement, it is unclear whether

Pre-training Dataset / Category	**Vehicle**	**Animal**
IMAGENET-LOC (Cls)	84.52	82.98
IMAGENET-LOC without Category	86.83	83.95
IMAGENET-LOC	87.20	84.83

Table 2: Results on PASCAL-VOC (mAP@0.7) after removing certain category during pre-training.

Method / Pre-trained Dataset	**mIoU**
DCNv1 [7]	75.2
IMAGENET-CLS [8]	75.7
IMAGENET-LOC [8, 45]	78.3
OPENIMAGES [25]	**78.6**

Table 4: Baseline and our fine-tuning results on PASCAL-VOC 2012 [11] semantic segmentation dataset.

% missed object	**occluded Low**	**occluded Medium**
IMAGENET-CLS [8]	14.7%	15.7%
OPENIMAGES [25]	**10.1%**	**10.8%**

Table 3: Missed objects under different occlusion levels in PASCAL-VOC 2007 [11] test set. Results obtained with the tool in [19].

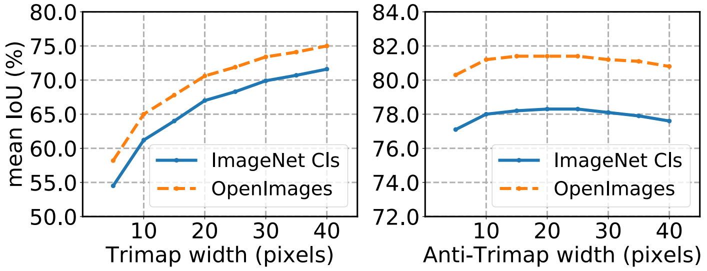


Figure 2: Trimap (**left**) and Anti-Trimap(**right**) experiments.

detection pre-training improves detector's localization ability in general or only on certain seen classes. To this end, we remove " Vehicle" or "Animal" category from IMAGENET-LOC during detection pre-training and compare fine-tuning results on PASCAL-VOC. To be more fair, we pre-train both classifier and detector on IMAGENET-LOC samples, where detector is trained as usual while classifier is trained with only class labels, denoted as IMAGENET-LOC (Cls). As shown in Table 2, detection pre-training still improves fine-tuning performance on these categories after removing them, suggesting that it improves localization ability in general.

2.2. Semantic Segmentation

Baseline Configuration and Results We fine-tune detection networks for the semantic segmentation task on PASCAL-VOC 2012. We use Deformable ConvNets [7] as our backbone in DeepLab [3] the same as [7] in our experiments. Results are shown in Table 4.

Detection Pre-Training Helps Segmentation The results after fine-tuning are shown in Table 4. These results show that networks pre-trained for object detection obtain a 3% performance improvement compared to image classification. We evaluate this for the OpenImages dataset and also for IMAGENET-LOC dataset.

Error Analysis We also perform experiments to understand where these improvements occur. Specifically, we study if the improvements from detection pre-trained networks are due to better segmentation at boundary pixels or not. For this we evaluate the accuracy at boundary pixels and non-boundary pixels. The boundary pixels are obtained by applying morphological dilation on the "void" labeled pixels which often occurs at object boundaries.

In particular, we perform two types of evaluations: 1) Accuracy at pixels which are within a distance x from an object boundary ("trimap experiment" [21, 23, 3, 4]) 2) Accuracy at pixels of an object or background as opposed to boundary pixels ("anti-trimap experiment"). The first evaluation compares the accuracy at boundary pixels and the second one compares the accuracy for pixels which are not at the boundary. The results for these experiments are shown in Fig. 2. These results show that the improvement in performance is not due to better classification at boundary pixels but the whole extent of object.

2.3. Image Classification

We also compare the effect of pre-training for image classification by evaluating multiple pre-trained detection backbones like IMAGENET-LOC and COCO apart from OPENIMAGES. Diverse classification datasets like CALTECH-256, SUN-397 and OXFORD-102 FLOWERS are considered. Apart from fine-tuning for image classification, we also evaluate off-the-shelf features from detection and classification backbones.

Fine-Tuning on Classification Results for fine-tuning different pre-trained networks on classification datasets are shown in Table 8. These results show that pre-training on IMAGENET-CLS outperforms IMAGENET-LOC, OPENIMAGES, and COCO by a significant margin on all three classification datasets. Therefore, pre-training for object detection hurts performance for image classification. It is a bit counter-intuitive that a network which also learns about the spatial extent of an object is worse at classification. To get a better understanding of the possible reasons, we evaluate features which are extracted from the pre-trained image classification networks without any fine-tuning.

Feature	Top-1 Acc
Conv5	76.7
ConvProj blob (256,14,14)	69.7
ConvProj blob (256,4,4)	72.4
ConvProj blob (256,2,2)	73.3
ConvProj blob (256)	74.1
FC1 (1024)	71.6
FC2 (1024)	70.0

Table 5: Linear classification results on Caltech-256 [14] using different features from the detection head of OpenImages [25] pre-trained object detection network.

Pre-trained Dataset	ImageNet-Cls [8, 45]	OpenImages [25]
Caltech-256 [14]	**84.7**	76.7
Sun-397 [51]	**57.3**	51.1
Oxford-102 Flowers [36]	**87.4**	83.1

Table 6: Linear classification results (Top-1 Accuracy) using Conv5 features from ImageNet-Cls and OpenImages pre-trained networks.

Pre-trained Dataset	Pascal-Voc		Caltech-256
	mAP@0.5	mAP@0.7	Top-1 Acc
ImageNet-Loc (Cls)	84.5	76.6	85.8
ImageNet-Loc	86.5	80.0	82.3

Table 7: Fine-tuning results on detection and classification with identical pre-training samples.

Conv5 features We average pool the Conv5 features extracted from networks pre-trained on OpenImages and ImageNet-Cls. Then we add a linear classifier followed by a softmax function to perform image classification. The results for different datasets are presented in Table 6. This shows that without fine-tuning, there exists a large performance gap between the features which are good for object detection *vs.* those which are trained for the task of image classification. The performance of the average pooled Conv5 features of ImageNet-Loc and OpenImages pre-trained networks is the same for classification on Caltech-256. For Coco, the performance drops further by 2%, possibly because of the smaller number of classes in object detection.

Intermediate Detection Features Table 5 compares features extracted from different layers in the detection head of the OpenImages pre-trained object detection network. We present results for classification on the Caltech-256 [14] dataset when a linear classifier is applied to different features, including avg pooled Conv5 (2048), ConvProj blob (256,14,14), avg pooled ConvProj blob (256,4,4), avg pooled ConvProj blob (256,2,2), avg pooled (ConvProj) (256), FC1 (1024) and FC2 (1024) features. We find that FC1 is better than FC2. The avg pooled (ConvProj) (256) is better than avg pooled ConvProj blob (256,2,2), which is better than ConvProj blob (256,4,4). Therefore, it is evident that preserving spatial information hurts image classification. Although averaging is an operation which can be learned from a higher dimensional representation (like ConvProj blob (256,14,14)), it is also easily possible to overfit to the training set in a higher-dimensional feature space. We also find that as we approach the Output layer of detection, the performance for image classification deteriorates.

Pre-trained Dataset	Caltech-256 [14]	Sun-397 [51]	Oxford-102 Flowers [36]
ImageNet-Loc [8, 45]	82.3	58.3	90.9
Coco [28]	79.8	57.8	91.4
OpenImages [25]	82.2	59.5	92.6
ImageNet-Cls [8]	**86.3**	**61.5**	**95.0**

Table 8: Results (Top-1 accuracy) for fine-tuning different pre-trained networks on classification datasets.

Comparison with Identical Pre-training Samples To further verify our empirical results, we pre-train base networks with identical training samples in ImageNet-Loc for classification and detection, then compare the fine-tuning results accordingly. The detector is pre-trained as usual and classifier is pre-trained on ImageNet-Loc with only class label, denoted as ImageNet-Loc (Cls). As shown in Table 7, the patterns are inline with the results and analysis discussed above, suggesting that task difference is the main factor behind the observed patterns.

3. Qualitative Analysis and Visualization

We also present extensive qualitative analysis and visualization for the downstream tasks. Due to page limit we refer the readers to Appendix 6.1 and 6.2.

4. Related Work

Related work is discussed in Appendix 6.3.

5. Conclusion

We present an extensive study on object detection pre-training. When fine-tuning on small detection datasets, we show that pre-training on large detection datasets is beneficial when higher degree of localization is desired. Typically, detection pre-training is beneficial for tasks where spatial information is important such as detection and segmentation, but when spatial invariance is needed, like classification, it can hurt performance. We also conduct feature level analysis and visualization analysis to provide a deeper understanding on the internal difference between detection and classification pre-training.

References

[1] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation. *TPAMI*, 38(9):1790–1802, 2016. 9

[2] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019. 10

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. 2, 3, 7

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3, 7

[5] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018. 9

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016. 7, 10

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2, 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 3, 4, 8, 9

[9] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 1, 9

[10] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 10

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88, 2010. 1, 2, 3

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. 2019. 10

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

[14] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. California Institute of Technology, 2007. 1, 2, 4

[15] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*. IEEE, 2011. 2

[16] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2

[17] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019. 1, 10

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 7

[19] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 3, 7

[20] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 9

[21] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009. 3

[22] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, 2019. 1, 2, 9

[23] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 3

[24] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016. 10

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 1, 2, 3, 4, 8, 9

[26] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 10

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 7, 10

[28] 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*, 2014. 1, 2, 4, 7

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 7, 10

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 7

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 7, 8

[33] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1, 7

[34] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *IJCV*, 2016. 10

[35] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for

neurons in neural networks via deep generator networks. In *NeurIPS*, 2016. 10

[36] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 1, 2, 4

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 7, 10

[38] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *ICML*, 2017. 10

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 10

[40] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshop*, 2014. 9

[41] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Dsod: Learning deeply supervised object detectors from scratch. In *ICCV*, 2017. 10

[42] Yosuke Shinya, Edgar Simo-Serra, and Taiji Suzuki. Understanding the effects of pre-training for object detectors via eigenspectrum. In *ICCV Workshop*, 2019. 9

[43] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 10

[44] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *CVPR*, 2018. 7, 10

[45] Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *CVPR*, 2018. 1, 2, 3, 4, 9

[46] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *NeurIPS*, 2018. 2

[47] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 10

[48] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 1, 7, 9

[49] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 7

[50] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 7

[51] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 2, 4

[52] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 7

[53] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 9

[54] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *ICML Workshop*, 2015. 10

[55] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 9

[56] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 10

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 7

[58] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 7

[59] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 10

[60] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2018. 1, 7

[61] Rui Zhu, Shifeng Zhang, Xiaobo Wang, Longyin Wen, Hailin Shi, Liefeng Bo, and Tao Mei. Scratchdet: Training single-shot object detectors from scratch. In *CVPR*, 2019. 10

[62] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 2

6. Appendix

6.1. Visualization

Semantic and Feature Analysis on Image Classification. In Fig 3 we show the most similar images in `Conv5` feature space for IMAGENET-CLS and OPENIMAGES pre-trained networks on CALTECH-256. As can be seen, similar images from IMAGENET-CLS features can have multiple objects; however for OPENIMAGES, the most similar image pairs typically match in shape and size. To understand the relationship between OPENIMAGES and IMAGENET-CLS features, we perform K-means clustering with different numbers of clusters (from 2 to 1000). Then, given an image pair in the same cluster in an embedding (like OPENIMAGES), we check if the same image pair belongs to the same cluster in another embedding (like ImageNet) or not. We plot this probability in Fig 6. This plot shows that if features are similar in the OPENIMAGES space, they are likely to be similar in the IMAGENET-CLS space; however the converse is not true. Some example images which are close in the IMAGENET-CLS space but distant in the OPENIMAGES space are shown in the middle of Fig 3. This shows that objects of different scale and similar texture can be close in the IMAGENET-CLS space but far away in the OPENIMAGES space. We briefly describe how we define close and distant. An image pair is considered to be close if it is part of the same cluster when the number of clusters is large ($>$ 1000). An image pair is considered to be distant if it not part of the same cluster when the number of clusters is small ($<$ 5).

We also show the t-SNE [32] visualization of `Conv5` features from IMAGENET-CLS and OPENIMAGES pre-trained networks before fine-tuning. Results in Fig 7 show that features from the same class are clustered and close to each other in the IMAGENET-CLS space; however, OPENIMAGES features are fragmented.

Activation Visualization To illustrate the differences in the learned representations between networks pre-trained on detection and classification datasets, we visualize the activation maps (`Conv5`) and investigate which part of input images contribute more. As shown in Fig 4(a-b), the IMAGENET-CLS pre-trained activation map tend to focus on discriminative parts. On the other hand, OPENIMAGES pre-trained models emphasize on the entire spatial extent of the objects. Moreover, the latter exhibits an instance-level representation, especially when multiple objects are present such as Fig 4(c-e).

Mask-out visualization Besides visualizing activation maps, we further conduct "Mask-out" visualization to reveal the relationship between image parts and the final class prediction. Specifically, we shift a 60x60 blank mask over the input image and measure the output confidence of the correct class. The classification layer for IMAGENET-CLS and the detection head for OPENIMAGES is replaced with a linear classification layer. In Fig 5, we show the classification probability at each pixel assuming that the center of the mask is placed at that location. We can see that for many locations (like the head of the camel), the classification score of the IMAGENET-CLS classifier drops to zero, which is not the case for OPENIMAGES. This is because detector relies more on the entire spatial extent of an object to make a prediction so the classification score is not sensitive to minor structural changes in the image, while the classifier focuses more on discriminative parts.

6.2. Qualitative Results and Error Analysis

Qualitative Results and Error Analysis on Object Detection. We show qualitative results on the PASCAL-VOC dataset for OPENIMAGES and IMAGENET-CLS pre-training. Fig 8 shows that localization for the OPENIMAGES model is better. The small gap between mAP@0.1 (where localization errors are typically ignored) and higher IoUs like 0.5 shown in Fig 1 indicates that large localization errors are rare. We also observe in Fig 8 that the OPENIMAGES model handles occlusion cases better. To further verify this observation, we analyze the errors using the analysis tools in [19, 28]. Quantitative results are mentioned in Table 3 which demonstrate that the OPENIMAGES pre-trained network is indeed better under occlusion.

Qualitative and Semantic Analysis on Semantic Segmentation We provide some qualitative examples for segmentation predictions in Fig 9 (using IMAGENET-LOC and IMAGENET-CLS). From these examples, we find that the network pre-trained on classification is unable to cover entire objects as it is weak at understanding instance boundaries - like in the case of the cow in Fig 9. Detection pre-training provides a better prior about the spatial extent of an instance which helps in recognizing parts of an object. It also helps more for object classes like sheep (+7.5%), cow (+6.5%), dining-table (+5.6%). These classes typically have a multi-modal distribution in appearance (like color and shape distribution). On the other hand, classes like Potted Plant which have a consistent shape and appearance, obtain no improvement in performance when detection pre-training is used.

6.3. Related Work

Large Scale Pre-training ImageNet pre-training was crucial to obtain improvements over state-of-the-art results on a wide variety of tasks such as object detection [27, 29, 37, 44, 6, 27], semantic segmentation [3, 18, 4, 30, 57], action/event recognition [49, 58, 52, 50] *etc*. Due to the importance of pre-training, the trend continued towards collecting progressively larger classification datasets such as JFT [48], Places [60] and Instagram [33] to obtain better performance. While the effect of large-scale classification

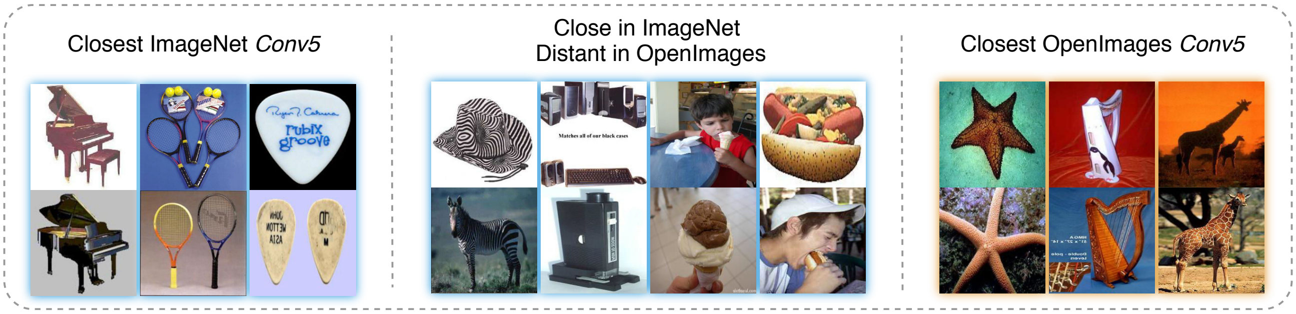


Figure 3: Qualitative results of feature space analysis. **Left/Right**: Image pairs that are closest in feature space of IMAGENET-CLS/OPENIMAGES pre-trained network. **Middle**: Image pairs that are close in feature space of IMAGENET-CLS pre-trained network but distant in that of OPENIMAGES pre-trained network.

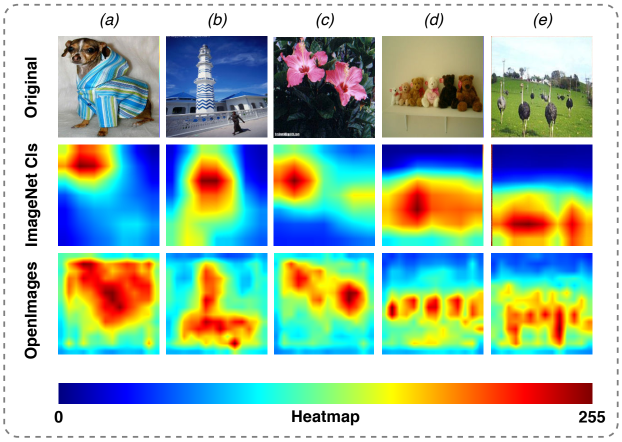
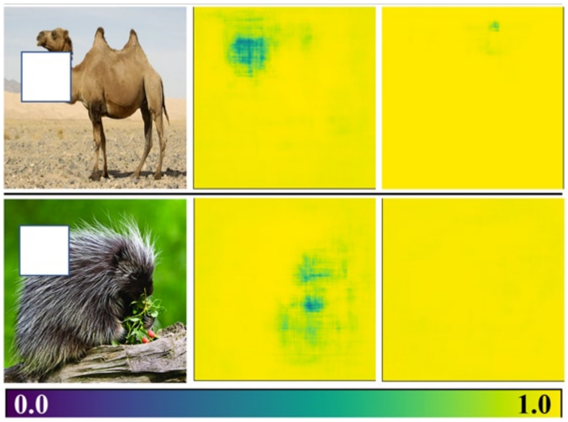


Figure 4: Activation visualization of networks pre-trained on IMAGENET-CLS [8] and OPENIMAGES [25].



(a) Image + Mask (b) ImageNet Model (c) OpenImages Model

Figure 5: Mask-out visualization. The probability of the correct class at each blank mask position is shown.

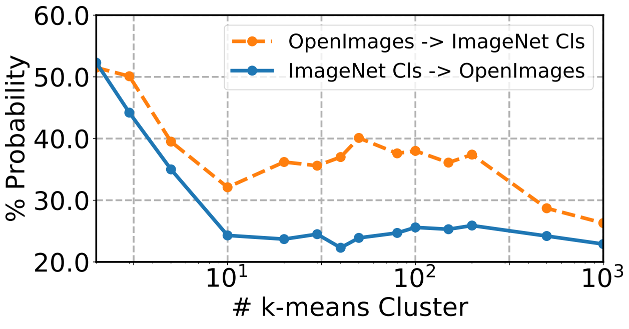


Figure 6: Feature analysis. Similar features in OPENIMAGES space are more likely to be similar in IMAGENET-CLS space, but the reverse is not true.

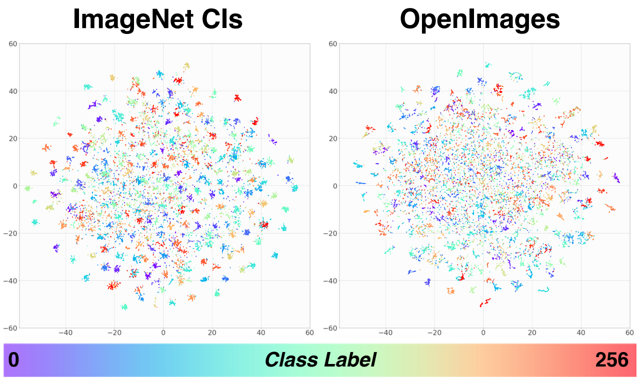


Figure 7: t-SNE [32] visualization of `Conv5` features from IMAGENET-CLS [8] and OPENIMAGES [25] pre-trained networks.

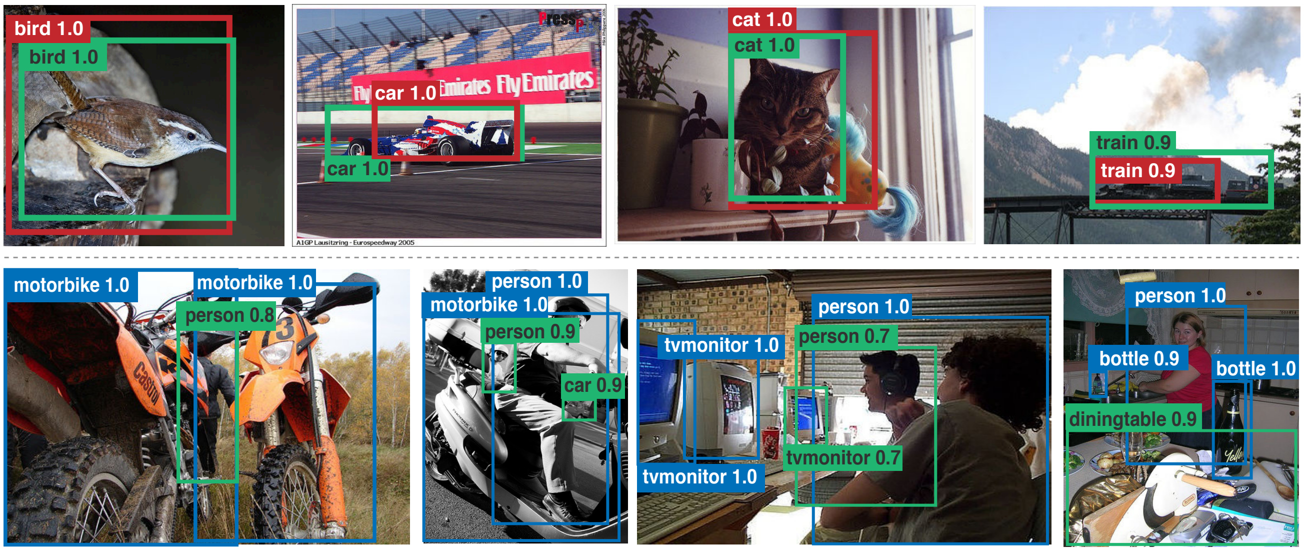


Figure 8: Qualitative results from detectors pre-trained on ImageNet-Cls [8] and OpenImages [25] **Above**: OpenImages pre-trained detector shows better localization ability. Green and red boxes are from OpenImages pre-trained and ImageNet-Cls pre-trained detectors respectively. **Below**: OpenImages pre-trained detector handles occusion cases better. Blue boxes are correct predictions from both detectors while green boxes are occluded objects successfully detected only by the OpenImages pre-trained detector.

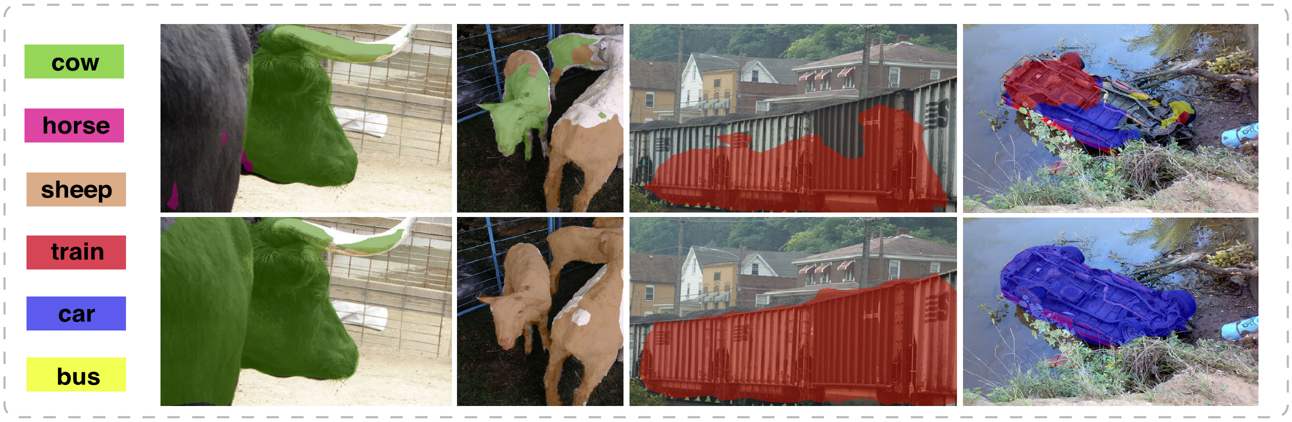


Figure 9: Qualitative results of semantic segmentation from networks pre-trained on ImageNet-Cls [8] (**Above**) and ImageNet-Loc [8, 45] (**Below**). The ImageNet-Loc model is better at covering entire objects while the classification pre-trained model is more likely to mis-classify pixels on some parts of an object.

is extensively studied [40, 9], there is little work on understanding the effect of pre-training on object detection.

Transfer Learning The transferability of pre-trained features has been well studied [1, 53, 20, 22, 5, 48, 42]. For example, [1] measured the similarity between a collection of tasks with ImageNet classification; [5] studied how to transfer knowledge learned on large classification datasets to small fine-grained datasets; [22] addressed relationship among ImageNet pre-training accuracy, transfer accuracy and network architecture; [55] proposed a computational approach to model relationships among visual tasks of various abstract levels and produced a computational taxonomic map. However, the visual tasks in [55] did not involve object detection although object detection is one of the few tasks other than image-classification for which large-scale pre-training can be performed. We study the transferability, generalizability, and internal properties of networks pre-trained for object detection.

Understanding CNNs Towards understanding the superior performance of CNNs on complex perceptual tasks, various qualitative [43, 59, 39, 54, 47, 35, 56, 10, 34] and quantitative [38, 24, 12, 2] approaches have been proposed. A number of previous works explain the internal structure of CNNs by highlighting pixels which contribute more to the prediction using gradients [43], Guided BackPropogation [56, 39], deconvolution [47], *etc*. Other methods adopt an activation maximization based approach and synthesize the preferred input for a network neuron [35, 34, 54, 10]. Attempts have also been made to interpret the properties of CNNs empirically by investigating what it learns and is biased towards [38, 24, 12, 2], such as shape and texture of objects.

Training From Scratch While most modern detectors are pre-trained on the ImageNet classification dataset [27, 44, 6, 27, 29, 37], effort has also been made to deviate from the conventional paradigm and train detectors from scratch [41, 26, 61]. [17] demonstrated that with a longer training schedule, detectors trained from scratch can be as good as ImageNet pre-trained models on large datasets (like COCO). However, pre-training is still crucial when the training dataset is small (like PASCAL-VOC).

6.4. Acknowledgement

This work was supported in part by the Office of Naval Research (ONR) under Grant N000141612713: Visual Common Sense Reasoning for Multi-agent Activity Prediction and Recognition and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of ONR, IARPA, DOI/IBC or the U.S. Government.