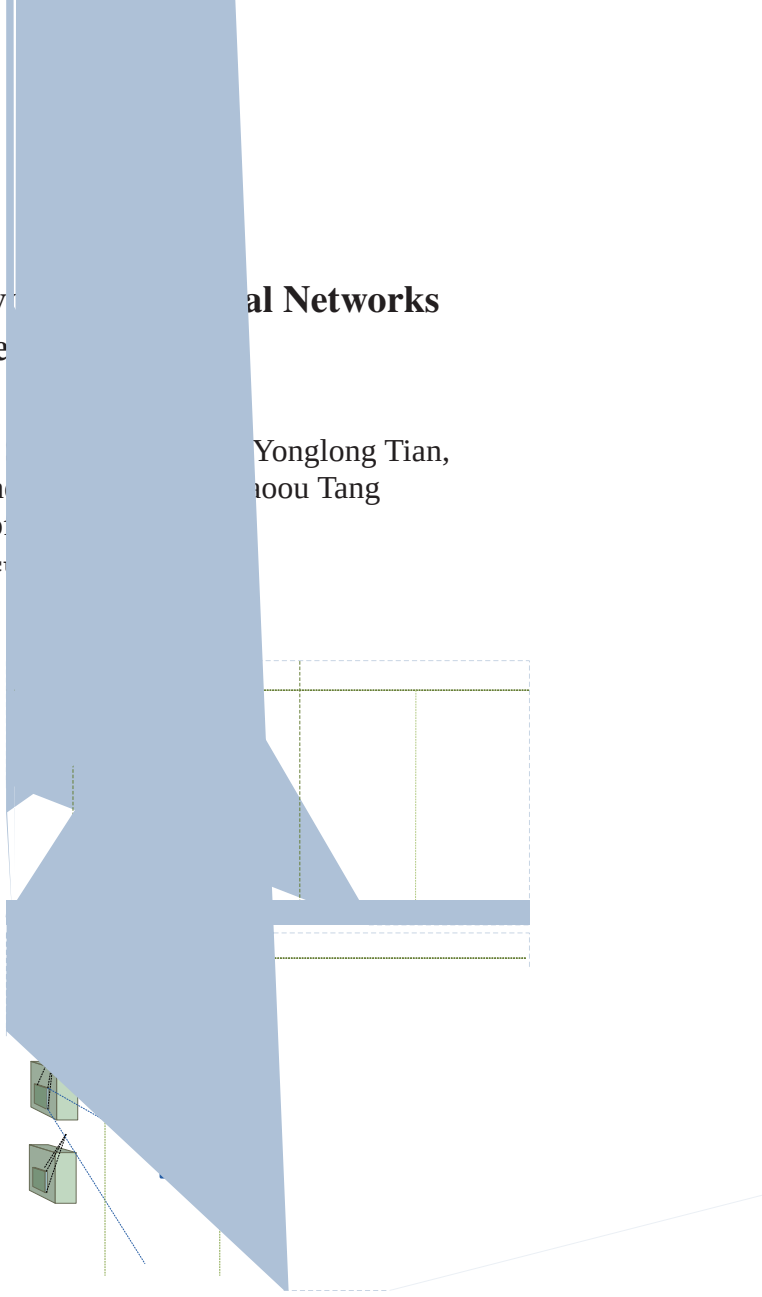
DeepID-Net: Deformable Deep Convi al Networks for ●bje t Dete

Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Yonglong Tian,
Hongsheng Li, Shuo Yang, Zhe Wang, Ch hoou Tang
The Chinese University o
wlouyang, xgwang@ee.c



Abstra t

In this paper, we propose deformable deep convolutional neural networks for generic object detection. This new deep learning object detection framework has innovations in multiple aspects. In the proposed new deep architecture, a new deformation constrained pooling (def-pooling) layer models the deformation of object parts with geometric constraint and penalty. A new pre-training strategy is proposed to learn feature representations more suitable for the object detection task and with good generalization capability. By changing the net structures, training strategies, adding and removing some key components in the detection pipeline, a set of models with large diversity are obtained, which significantly improves the effectiveness of model averaging. The proposed approach improves the mean averaged precision obtained by RCNN [14], which was the state-of-the-art, from 31% *to* 50.3% *on the ILSVRC2014 detection test set. It also outperforms the winner of ILSVRC2014, GoogLeNet, by 6.1%. Detailed component-wise analysis is also provided through extensive experimental evaluation, which provide a global view for people to understand the deep learning object detection pipeline.*

1. Introdu tion

Object detection is one of the fundamental challenges in computer vision. It has attracted a great deal of research interest [5, 39, 11, 19]. Intra-class variation in appearance and deformation are among the main challenges of this task.

Because of its power in learning features, the convolutional neural network (CNN) is being widely used in recent large-scale object detection and recognition systems [44, 38, 19, 22]. Since training deep models is a non-convex optimization problem with millions of parameters, the choice of a good initial point is a crucial but unsolved problem, especially when deep CNN goes deeper [44, 38, 22]. It is also easy to overfit to a small train-

the image classification task and fine-tuning for the object detection task. For image classification, the input is a whole image and the task is to recognize the object within this image. Therefore, learned feature representations have robustness to scale and location change of objects in images. Taking Fig. 1(a) as an example, no matter how large and where a person is in the image, the image should be classified as person. However, robustness to object size and location is not required for object detection. For object detection, candidate regions are cropped and warped before they are used as input of the deep model. Therefore, the positive candidate regions for the object class person should have their lo-

existing deep CNN models, max pooling and average pooling are useful in handling deformation but cannot learn the deformation penalty and geometric models of object parts. The deformation layer was first proposed in [29] for pedestrian detection. In this paper, we extend it to general object detection on ImageNet. In [29], the deformation layer was constrained to be placed after the last convolutional layer, while in this work the def-pooling layer can be placed after all the convolutional layers to capture geometric deformation at all the information abstraction levels. In [29], it was assumed that a pedestrian only has one instance of a body part, so each part filter only has one optimal response in a detection window. In this work, it is assumed that an object has multiple instances of a part (e.g. a car has many wheels), so each part filter is allowed to have multiple response peaks in a detection window. Moreover, we allow multiple object categories to share deformable parts and jointly learn them with a single network. This new model is more suitable for general object detection.

Context gains attentions in object detection. The context information investigated in literature includes regions surrounding objects [5, 8, 13], object-scene interaction [9, 20], and the presence, location, orientation and size relationship among objects [2, 48, 49, 7, 31, 13, 40, 9, 53, 8, 50, 30, 6, 35, 45]. In this paper, we use whole-image classification scores over a large number of classes from a deep model as global contextual information to refine detection scores.

Besides feature learning, deformation modeling, and context modeling, there are also other important components in the object detection pipeline, such as pretraining [14], network structures [36, 54, 21], refinement of bounding box locations [14], and model averaging [54, 21, 19]. While these components were studies individually in different works, we integrate them into a complete pipeline and take a global view of them with component-wise analysis under the same experimental setting. It is an important step to understand and advance deep learning based object detection.

3. Method

3.1. ●verview of our approa h

An overview of our proposed approach is shown in Fig. 2. We take the ImageNet object detection task as an example. The ImageNet image classification and localization dataset with 1,000 classes is chosen to pretrain the deep model. Its object detection dataset has 200 object classes. In the experimental section, the approach is also applied to the PASCAL VOC. The pretraining data keeps the same, while the detection dataset only has 20 object classes. The steps of our approach are summarized as follows.

1. Selective search proposed in [39] is used to propose candidate bounding boxes.

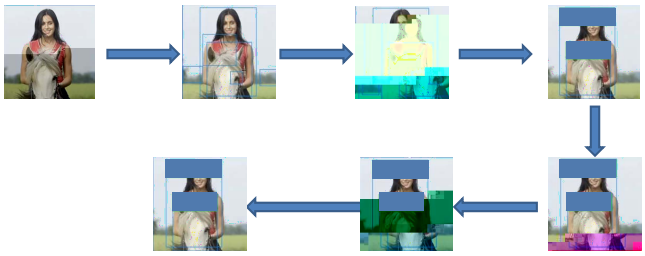
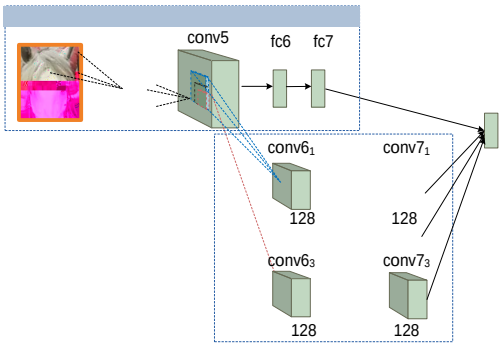


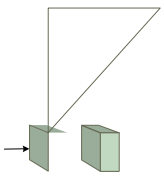
Figure 2. Overview of our approach. Find detailed description in the text of Section 3.1. Texts in red highlight the steps that are not present in RCNN [14].

2. An existing detector, RCNN [14] in our experiment, is used to reject bounding boxes that are most likely to be background.
3. An image region in a bounding box is cropped and fed into the DeepID-Net to obtain 200 detection scores. Each detection score measures the confidence on the cropped image containing one specific object class. Details are given in Section 3.2.
4. The 1000-class whole-image classification scores of a deep model are used as contextual information to re-


conv5
fc6
fc7
conv6₁
conv7₁
128
128
conv6₃
conv7₃
128
128

...

...



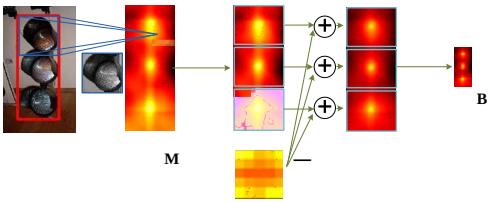


Figure 4. Def-pooling layer. The part detection map and the deformation penalty are summed up. Block-wise max pooling is then performed on the summed map to obtain the output $\mathbf{B}$ of size $\frac{H}{\_} \times \frac{W}{\_}$ (3×1 in this example).

provides the ability to learn the map that implicitly decides the kernel size for max-pooling.

Example 2. The deformation layer in [29] is a special case of the def-pooling layer by enforcing that $\mathbf{z}_{\delta\ ,\delta}$ in (1) covers all the locations in $\text{conv}_{l-1,i}$ and only one output with a pre-defined location is allowed for the def-pooling layer (i.e. $R = \infty$, $s_x = W$, and $s_y = H$). The proof can be found in Appendix 1. To implement quadratic deformation penalty used in [11], we can predefine $\{d^{\delta\ ,\delta}_{c,n}\}_{n=1,2,3,4} = \{\delta_x, \delta$

bulbs co-exist in a traffic light in Fig. 4.

3. As shown in Fig. 3, the def-pooling layer is a shared representation for multiple classes and therefore the learned visual patterns in the def-pooling layer can be shared among these classes. As examples in Fig. 6, the learned circular visual patterns are shared as different object parts in traffic lights, cars, and ipods.

The layers proposed in [29, 16] does not have these advantages, because they can only be placed after the final convolutional layer, assume one instance per object part, and does not share visual patterns among classes.

3.5. Fine-tuning the deep model with hinge-loss

In RCNN, feature representation is first learned with the softmax loss in the deep model after fine-tuning. Then in a separate step, the learned feature representation is input to a linear binary SVM classifier for detection of each class. In our approach, the softmax loss is replaced by the 200 binary hinge losses when fine-tuning the deep model. Thus the deep model fine-tuning and SVM learning steps in RCNN are merged into one step. The extra training time required for extracting features (2.4 days with one Titan GPU) is saved.

3.6. Contextual modeling

The deep model learned for the image classification task (Fig.

Table 1. Detection mAP (%

Table 4. Ablation study of the two pretraining schemes in Section 3.3 for different net structures on ILSVRC2014 val . Scheme 0 only

Referen es

[1] www.ee.cuhk.edu.hk/∼wlouyang/projects/imagenetdeepid. 6

[2] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *CVPR*, 2010. 3

[3] L. Bourdev and J. Malik. Poselets: body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 2

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 2, 8

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 3

[6] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 3

[7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 3

[8] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *CVPR*, 2012. 3

[9] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 3

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014. 1

[11] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. PAMI*, 32:1627–1645, 2010. 1, 2, 4, 5, 6, 8

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, 2005. 2

[13] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *CVPR*, pages 113–120. IEEE, 2010. 3

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2, 3, 4, 6, 7

[15] R. Girshick, P. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://www.cs.berkeley.edu/rbg/latent-v5/. 6, 7

[16] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *arXiv preprint arXiv:1409.5403*, 2014. 5, 6, 7

[17] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. *arXiv preprint arXiv:1403.1840*, 2014. 2

[18] B. Hariharan, C. L. Zitnick, and P. Dollár. Detecting objects using deformation dictionaries. In *CVPR*, 2014. 2

[19] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. 2014. 1, 2, 3, 6, 7

[20] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, pages 30–43. 2008. 3

[21] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3, 6, 7

[22] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 1, 2

[23] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014. 2

[24] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012. 2

[25] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013. 2

[26] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional neural network. In *ICCV*, 2013. 2

[27] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, pages 2337–2344. IEEE, 2014. 2

[28] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012. 2

[29] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013. 3, 5, 6

[30] W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 2013. 2, 3

[31] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 3

[32] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014. 1

[33] X. Ren and D. Ramanan. Histograms of sparse codes for object detection. In *CVPR*, pages 3246–3253. IEEE, 2013. 6, 7

[42] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for computing face similarities. In *ICCV*, 2013. 2

[43] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 2

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 1, 2, 4, 7