Part Localization by Exploiting Deep Convolutional Networks

Marcel Simon, Erik Rodner, and Joachim Denzler

Computer Vision Group, Friedrich Schiller University of Jena, Germany
www.inf-cv.uni-jena.de

Abstract. Deep convolutional neural networks have shown an amazing ability to learn object category models from large-scale data. In this paper, we present a novel approach for part discovery and detection with a pre-trained convolutional neural network. It is based on analyzing gradients of intermediate layer outputs and locating areas containing large gradients. By comparing these with ground-truth part locations, channels in the network related to semantic object parts are identified. On the Caltech Birds CUB200-2011 dataset, our approach achieves state-of-the-art performance in part localization as well as image categorization. An important advantage is that it can be also applied if no bounding box annotation is given during testing.

1 Introduction

In recent years, the concept of *deep learning* [2, 1] has gained tremendous interest in the vision community. A key idea is to jointly train a model for the whole classification pipeline. A successful model especially for classification are convolutional neural networks (CNN) [8]. The very recent work of [6, 10, 9] shows that pre-trained deep models [8] can also be exploited for classification tasks on datasets which they were not trained on. Our work follows a similar line of thought. In particular the questions we were interested in are: *"Can we re-use pre-trained deep convolutional networks for part discovery and detection? Does a deep model learned on ImageNet [5] already include implicit detectors related to common parts found in fine-grained recognition tasks?"*

The answer to both questions is yes and to show this we present a novel part discovery and detection scheme using pre-trained deep convolutional neural networks. Object representations are often part-based and the benefit is especially notable in fine-grained classification tasks [14, 3, 7, 4]. Our technique for providing such a part-based representation is based on computing gradient maps with respect to certain channel outputs and finding clusters of high activation within. This is followed by selecting channels which have their corresponding clusters closest to ground-truth positions of semantic parts. An outline of our approach is given in Fig. 1. The most interesting aspect is that after a simple training step, parts can be reliably detected without much additional computational effort based on the results of the CNN.

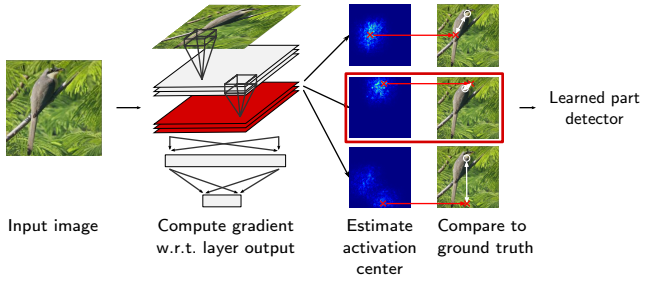

Fig. 1. Outline of our approach during learning: (1) compute gradients of CNN channels with respect to image positions, (2) estimate activation centers, (3) find spatially related semantic parts to select useful channels that act as part detectors later on.

2 Part Discovery in CNNs by Correspondence

Part Discovery by Correspondence Most recent deep learning architectures for vision are based on a single CNN comprised of multiple convolutional and fully connected layers. Important for our approach is that the output of the convolutional layers is organized in channels from which we now want to identify those related to object parts. In the following, we assume that the ground-truth part locations z_i of the training images x_i are given. However, our method can be also provided with the location of the bounding box only, but we leave this for future work. We associate a binary latent variable h_k with each channel k, which indicates whether the channel is related to an object part. Our part discovery scheme can be motivated as a maximum likelihood estimation of these variables. First, let us consider the task of selecting the most related channel corresponding to a part which can be written as (assuming x_i are independent samples):

$$\hat{k} = \underset{1 \leq k \leq K}{\operatorname{argmax}} \; p(\boldsymbol{X} \mid h_k = 1) = \underset{1 \leq k \leq K}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{p\left(h_k = 1 | x_i\right) p\left(x_i\right)}{p\left(h_k = 1\right)} \,. \tag{1}$$

where $\boldsymbol{X}$ is the training data and K is the total number of channels. In the following, we assume a flat prior for $p(h_k = 1)$ and $p(x_i)$. The term $p\left(h_k = 1 | x_i\right)$ expresses the probability that channel k corresponds to the part currently under consideration given a single training example x_i. This is the case when the position p_i^k estimated using channel k equals the ground-truth part position z_i. However, the estimated position p_i^k is likely not perfect, so we assume it to be a Gaussian random variable distributed as $p_i^k \sim \mathcal{N}(\mu_i^k, \sigma^2)$, where μ_i^k is the center of activation extracted from the gradient map of channel k. We therefore have:

$$p(h_k = 1 | x_i) = p(p_i^k = z_i | x_i) = \mathcal{N}(z_i | \mu_i, \sigma^2) \tag{2}$$

Putting it all together, we obtain a very simple scheme for selecting a channel:

$$\hat{k} = \underset{1 \leq k \leq K}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(h_k = 1 | x_i) = \underset{1 \leq k \leq K}{\operatorname{argmin}} \sum_{i=1}^{N} \|\mu_i^k - z_i\|^2 \tag{3}$$

For all channels of all training images, the center of activation μ_i^k is calculated as explained in the subsequent paragraph. These locations are compared to the ground-truth part locations z_i by computing the mean distance. Finally, for each ground-truth part, the channel with the smallest mean distance is selected. The result is a set of channels, which are sensitive to different parts of the object. There does not need to be a one-to-one relationship between parts and channels.

In order to robustly localize the center of activation of a channel, we first calculate the gradient for each channel output with respect to the input image in a similar fashion as done in [11] for full objects. All gradients are summed up in order to obtain a single gradient map and a Gaussian mixture model with two components is fitted to the pixel locations weighted by the normalized absolute gradient values. We then take the mean location of the most prominent cluster in the mixture as the center of activation. In comparison to simply taking the maximum position in the gradient map, this approach is much more robust to noise as can be seen in the experiments.

Why should this work? The results of [13] suggest that at least in the special case of deep CNNs trained on ImageNet, each element of a hidden layer is sensitive to specific patterns in the image. That means the occurrence of a pattern leads to a substantial change of the output. There is an implicit association between certain image patterns and output elements of a particular layer. In higher layers these patterns become increasingly abstract and hence might correspond to a specific part of an object. Our method automatically identifies channels with this property.

3 Experiments

Experimental Setup We evaluate our approach on the challenging Caltech Birds CUB200-2011 [12] dataset. The CNN framework DeCAF [6] and the network learned on the ILSVRC 2012 dataset provided by the authors of [6] is used for all experiments. Out of this network, we use the 256 channels of the last pooling layer for the part detector discovery.

We also apply our part detection approach to the part-based classification system of [7] replacing the SIFT and color name features by the CNN activations of the last hidden layer and the part transfer by the presented approach. At the estimated part positions of the training and test images, squared patches of size $p = \sqrt{n \cdot m \cdot \lambda}$ are extracted, where m and n denote the height and width of the image or the bounding box, depending on whether the bounding box is given or not. We used $\lambda = \frac{1}{9}$ if the bounding box is unknown and $\lambda = \frac{1}{4}$ if known.

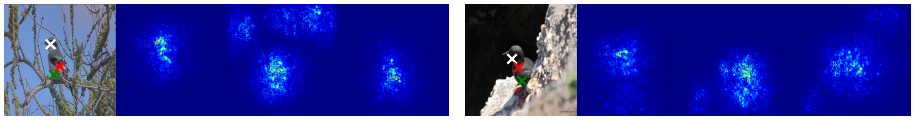

Fig. 2. Part localization results for the head (white), belly (green) and tail (red) along with the corresponding gradient maps for two images from the test set. No bounding box and no geometric constraints for the part locations are used during the localization.

Table 1. Part localization error on the CUB-2011-200 dataset for our method w/ and w/o GMM for finding the activation centers, our method w/ and w/o restricting the localization to the bounding box (BB), and the method of [7] .

Method	Norm. Error
Ours (GMM, BB)	**0.16**
Ours (GMM, Full)	0.17
Ours (MaxG, BB)	0.17
Part Transfer [7] (BB)	0.18

Table 2. Species categorization performance on the CUB200-2011 dataset. The bounding box is either known (BB) or unknown (Full) at test time.

Method	Recognition rate
POOF (BB) [3]	56.78%
Part transfer (BB) [7]	57.84%
Symbiotic (BB) [4]	59.4%
Ours (BB, Est. Parts)	62.53%
DeCAF + DPD (BB) [6]	64.96%
Ours (Full, Est. Parts)	60.17%
Ours (Full, GT Parts)	60.55%

Results Figure 2 presents some examples of our part localization applied to uncropped test images along with the corresponding gradient maps. The first quantitative analysis examines to what extent the learned part detectors relate to semantic parts. After identifying the spatially most related channel for each semantic part, we can apply our method to the test images to predict the location of semantic parts. The normalized localization errors calculated according to [7] are given in Table 1. There are groups of parts that are associated with the same channel. The results of the part-based classification are given in Table 2. In contrast to other methods, our approach can perform fine-grained classification on full images without a manual preselection of the area containing the bird.

4 Conclusions

We very briefly presented a novel approach for object part discovery and detection with pre-trained deep models. We make use of the high-level knowledge of CNNs to discover useful parts for a fine-grained recognition task by analyzing gradient maps of deep models and selecting activation centers related to annotated semantic parts. After this simple learning step, part detection basically comes for free when applying the deep CNN to the image. In contrast to previous work [7], our approach is also suitable for situations when the ground-truth bounding box is not given during testing.

References

1. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (Jan 2009), http://dx.doi.org/10.1561/2200000006
2. Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 35(8), 1798–1828 (2013)
3. Berg, T., Belhumeur, P.: POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 955–962 (June 2013)
4. Chai, Y., Lempitsky, V., Zisserman, A.: Symbiotic segmentation and part localization for fine-grained categorization. In: IEEE International Conference on Computer Vision (ICCV). pp. 321–328 (Dec 2013)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 248–255 (June 2009)
6. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
7. Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2014), preprint, http://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Goring_Nonparametric_Part_Transfer_2014_CVPR_paper.pdf
8. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. vol. 25, pp. 1097–1105 (2012)
9. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. arXiv preprint arXiv:1403.6382 (2014)
10. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR). CBLS (2014), preprint, http://arxiv.org/abs/1312.6229
11. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
12. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
13. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901 (2013)
14. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: IEEE International Conference on Computer Vision (ICCV). pp. 729–736 (Dec 2013)