Depth CNNs for RGB-D Scene Recognition: Learning From Scratch Better Than Transferring From RGB-CNNs

Xinhang Song[1,2], **Luis Herranz**[1], **Shuqiang Jiang**[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, China
{xinhang.song, luis.herranz, shuqiang.jiang}@vipl.ict.ac.cn

Abstract

Scene recognition with RGB images has been extensively studied and has reached very remarkable recognition levels, thanks to convolutional neural networks (CNN) and large scene datasets. In contrast, current RGB-D scene data is much more limited, so often leverages RGB large datasets, by transferring pretrained RGB CNN models and fine-tuning with the target RGB-D dataset. However, we show that this approach has the limitation of hardly reaching bottom layers, which is key to learn modality-specific features. In contrast, we focus on the bottom layers, and propose an alternative strategy to learn depth features combining local weakly supervised training from patches followed by global fine tuning with images. This strategy is capable of learning very discriminative depth-specific features with limited depth images, without resorting to Places-CNN. In addition we propose a modified CNN architecture to further match the complexity of the model and the amount of data available. For RGB-D scene recognition, depth and RGB features are combined by projecting them in a common space and further leaning a multilayer classifier, which is jointly optimized in an end-to-end network. Our framework achieves state-of-the-art accuracy on NYU2 and SUN RGB-D in both depth only and combined RGB-D data.

Introduction

Success in visual recognition mainly depends on the feature representing the input data. Scene recognition in particular has benefited from recent developments in the field. Most notably, massive image datasets (ImageNet and Places(Zhou et al. 2014)) provide the necessary amount of data to train complex convolutional neural networks (CNNs)(Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015), with millions of parameters, without falling into overfitting. The features extracted from models pretrained in those datasets have shown to be generic and powerful enough to obtain state-of-the-art performance in smaller datasets (e.g. MIT indoor 67(Quattoni and Torralba 2009), SUN397(Xiao et al. 2010)), just using an SVM(Donahue et al. 2014) or fine-tuning, outperforming earlier handcrafted paradigms (e.g. SIFT, HOG, bag-of-words).

Low cost depth sensors can capture depth information in addition to RGB data. Depth can provide valuable information to model object boundaries and understand the global layout of objects in the scene. Thus, RGB-D models can improve recognition over mere RGB models. However, in contrast to RGB data, which can be crowdsourced by crawling the web, RGB-D data needs to be captured with a specialized and relatively complex setup(Silberman and Fergus 2011; Song, Lichtenberg, and Xiao 2015). For this reason, RGB-D datasets are orders of magnitude smaller than the largest RGB datasets, also with much fewer categories. This prevents from training deep CNNs properly, and handcrafted features are still a better choice for this modality.

However, the recent SUN RGB-D dataset(Song, Lichtenberg, and Xiao 2015) is significantly larger than previous RGB-D scene datasets (e.g. NYU2(Silberman et al. 2012)). While still not large enough to train from scratch deep CNNs of comparable size to RGB counterparts (10335 RGB-D images compared with 2.5 million RGB images in Places), at least provides enough data for fine tuning deep models (e.g. AlexNet-CNN on Places) without significant overfitting. This approach typically exploit the HHA encoding for depth data(Gupta et al. 2014a), since it also a three channel representation (horizontal disparity, height above ground, and angle with the direction of gravity, see Figure 1 top). Fine tuning is typically used when the target dataset has limited data, but there is another large dataset that covers a similar domain which can be exploited first to train a deep model. Thus, transferring RGB features and fine tuning with depth (HHA) data is the common practice to learn deep representations for depth data(Song, Lichtenberg, and Xiao 2015; Wang et al. 2016; Zhu, Weibel, and Lu 2016; Gupta, Hoffman, and Malik 2016). However, although HHA images resemble RGB images and shapes and objects can be identified, is it really reasonable reusing RGB features in this inter-modal scenario?

In this paper we will focus on the low-level differences between RGB and HHA data, and show that a large number of low-level filters are either useless or ignored during fine tuning the network from RGB to HHA. Figure 1 (middle) shows the average activation ratio (i.e. how often the activation is non-zero) of the 96 filters in the layer conv1 of Places-CNN for different input data (sorted in descending order). When a network is properly designed and trained, it tends to show a balanced activation rate curve (e.g. conv1 activations extracted from the Places validation set, where the curve is almost a constant) meaning that most of the filters

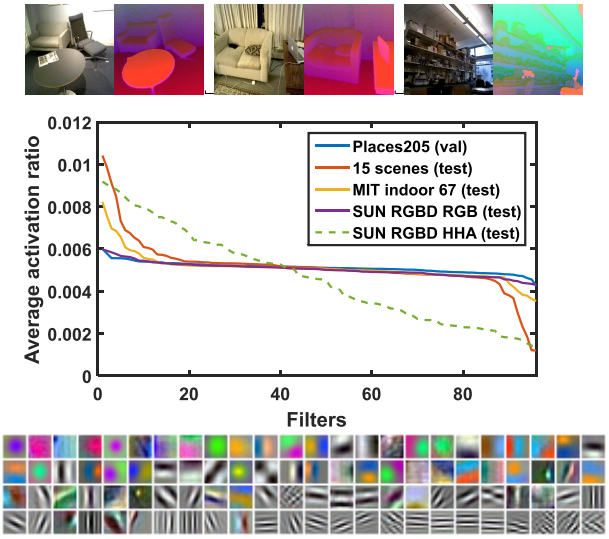


Figure 1: RGB and depth modalities. Top: examples of scenes captured in RGB and depth (HHA encoding). Middle: average nonzero activations of filter in the conv1 layer of Places-CNN for different datasets. Bottom: Conv1 filters ordered by mean activation on SUN RGB-D HHA.

are contributing almost equally to build discriminative representations. When transferred to other RGB scene datasets, the curve is still very similar, showing that the majority of the conv1 filters are still useful for 15 scenes (Lazebnik, Schmid, and Ponce 2006) and MIT Indoor 67. However, the curve for HHA shows a completely different behavior, where only a subset of the filters are relevant, while a large number are rarely activated, because they are not useful for HHA (see Figure 1 bottom). Edges and smooth gradients are common, while, Gabor-like patterns and high frequency patterns are seldom found in HHA data.

Thus, we observe that filters at the very bottom layers are crucial. However, conventional full fine tuning from RGB CNNs can hardly reach them (i.e. vanishing gradient problem), we explore other ways to make better use of the limited data while focusing on bottom layers. In particular, we compare the strategies of fine tuning only top and only bottom layers, and propose a weakly supervised strategy to learn filters directly from the data. In addition, we combine the pretrained RGB and depth networks into a new network, fine tuned with RGB-D image pairs. We show experimentally that these features lead to state-of-the-art performance with depth and RGB-D, and provide some insights and evidences.

Related work

RGB-D scene recognition

Earlier works use handcrafted features, engineered by an expert to capture some specific properties considered representative. Gupta *et al. (Gupta et al. 2015)* propose a method to detect contours on depth images for segmentation, then further quantize the segmentation outputs as local features for scene classification. Banica *et al. (Banica and Sminchisescu 2015)* quantize local features with second order pooling, and use the quantized feature for segmentation and scene classification. More recently, multi-layered networks can learn features directly from large amounts of data. Socher *et al.*(Socher et al. 2012) use a single layer CNN trained unsupervisedly on patches, and combined with a recurrent convolutional network (RNN). Gupta *et al.*(Gupta et al. 2014a) use R-CNN on depth images to detect objects in indoor scenes. Since the training data is limited, they augment the training set by rendering additional synthetic scenes.

Current state-of-the-art relies on transferring and fine tuning Places-CNN to RGB and depth data(Gupta, Hoffman, and Malik 2016; Wang et al. 2016; Zhu, Weibel, and Lu 2016; Song, Lichtenberg, and Xiao 2015). Wang *et al.* (Wang et al. 2016) extract deep features on both local regions and whole images, then combine the features of all RGB and depth patches and images in a component aware fusion method. Some approaches(Zhu, Weibel, and Lu 2016; Gupta, Hoffman, and Malik 2016) propose incorporating CNN architectures to fine-tune jointly RGB and depth image pairs. Zhu *et al.* (Zhu, Weibel, and Lu 2016) jointly fine-tune the RGB and depth CNN models by including a multi-model fusion layer, simultaneously considering inter and intra-modality correlations, meanwhile regularizing the learned features to be compact and discriminative. Alternatively, Gupta *et al. (Gupta, Hoffman, and Malik 2016)* propose to transfer RGB CNN model to the depth data according to the RGB and depth image pairs.

In this paper we avoid relying on large yet biased RGB models to obtain depth features, and train depth CNNs using weak supervision directly from depth data, learning truly depth-specific and discriminative features, compared with those transferred and adapted from biased RGB models.

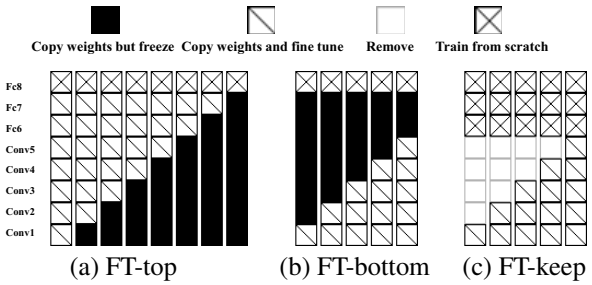
Weakly-supervised CNNs

Recently, several works propose weakly supervised frameworks(Durand, Thome, and Cord 2016; Bilen and Vedaldi 2016; Oquab et al. 2015), specially for object detection (object labels are known but not the bounding boxes). Oquab *et al.* (Oquab et al. 2015) propose an object detection framework to fine tune pretrained CNNs with multiple regions, where a global max-pooling layer selects the regions to be used in fine tuning. Durand *et al.* (Durand, Thome, and Cord 2016) extend this idea by selecting both useful (positive) and "useless" (negative) regions with a maximum and minimum mixed pooling layer. The weakly supervised detection network in (Bilen and Vedaldi 2016) uses a region proposal method to select regions.

These works rely on CNNs already pretrained on large datasets, and weak supervision is used in a subsequent fine tuning or adaptation stage to improve the final features for a particular task. In contrast, our motivation is training when data is very scarce, with a weakly supervised CNN that does not rely on any pretrained CNNs. In fact it is used to pretrain convolutional layers prior to fine tuning with full images.

Transferring from RGB to depth

Fine tuning Places-CNN with depth data. Transferring RGB CNNs and fine tuning with RGB data (intra-modal



(a) FT-top (b) FT-bottom (c) FT-keep

Figure 2: Strategies to fine tune Places-CNN withdepth data, (a) only top layers, bottom layers are frozen, (b) only bottom layers, top layers are frozen; (c) bottom layers (top layers retrained and some convolutional layers removed). Each column represents a particular setting.

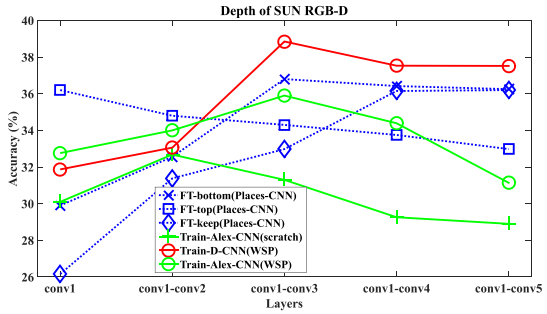
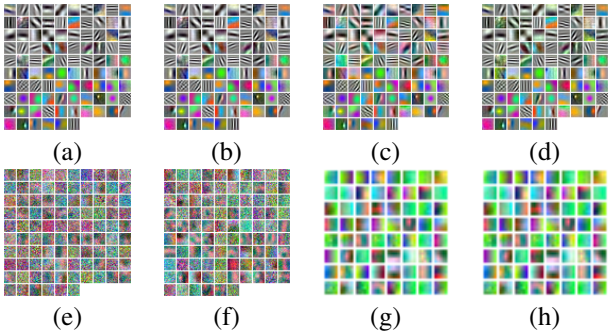


Figure 3: Comparisons of different learning strategies on depth images in accuracy (%) obtained with softmax, the fine tuning strategies are illustrated in Fig. 2 and training strategies are in Fig. 6.

transfer) has been well studied. In general, low-level filters from bottom layers capture generic patterns from the real visual world, and can be reused effectively in datasets with the same modality and similar characteristics. Thus, fine tuning one or two top layers (e.g. fc6, fc7) is often enough. These layers are more dataset-specific and need to be rewired to the new target categories(Agrawal, Girshick, and Malik 2014). In contrast, RGB and depth images have significantly different low-level visual patterns and regularities (e.g. depth patterns in HHA encoding are typically smooth variations, contrasts and borders, but without textures and high frequency patterns). Since the bottom convolutional layers are essential to capture this modality-specific visual appearances and regularities, only fine tuning top layers seem insufficient to adapt the RGB CNN properly to depth data.

Since we want to focus on bottom layers, we compare conventional fine tuning with other strategies that reach bottom layers better (see Fig. 2, each column represents a particular setting). The strategies differ basically on which layers are fine tuned, trained and which remain unaltered. Using the AlexNet architecture and the pretrained Places-CNN, we first compare three strategies: a) *FT-top (Places-CNN),* the conventional method where only a few top layers are fine tuned, b) *FT-bottom (Places-CNN),* where a few bottom layers are frozen, and c) *FT-keep (Places-CNN)*, top layers are directly removed. Note that fc8 is always trained, since it must be resized according to the target number of categories.

The classification accuracy on depth data (from SUN RGB-D dataset) with different strategies is shown in Fig. 3. Fine tuning top layers (*FT-top*) does not help significantly until including bottom convolutional layers, which is the opposite to fine tuning for RGB(Agrawal, Girshick, and Malik 2014), where fine tuning one or two top layers is almost enough to reach the maximum gain, and further extending to bottom layers helps very marginally. In contrast, fine tuning only the three bottom layers (*FT-bottom*) achieves 36.5% which is higher than fine tuning all layers. Furthermore, fine tuning after removing top layers (*FT-keep*) is also comparable to fine tuning all layers. All these results support our intuition that bottom layers are much more important than top layers when transferring RGB to depth, and that conventional transfer learning and adaptation tools used in RGB modality may not be effective in this inter-modal case.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4: Visualizing the first convolutional layer (conv1): (a) Places-CNN; (b) full fine tuned Places-CNN; (c) FT-bottom (Places-CNN); (d) FT-keep (Places-CNN), conv1; (e) Train-Alex-CNN (scratch); (f) Train-Alex-CNN (WSP), training with patches (99×99 pixels); (g) WSP-CNN, kernel size 5×5 pixels, training with patches (35×35 pixels); (f) Train-D-CNN (WSP). All methods are trained/fine tuned using only the depth data from SUN RGB-D.

More insight from *conv1* layer. To provide complementary insight of why bottom layers are more important we focus on the filters from the first convolutional layer, i.e. conv1, shown in Fig. 4. Although there is some gain in accuracy, it can be observed that only a few particular filters have noticeable changes during the fine tuning process. This suggest that the CNN is reusing RGB filters, and thus trying to find RGB-like patterns in depth data. Additionally, Fig. 1 middle shows that a large number of filters from Places-CNN are significantly underused on depth data (while they are properly used on RGB data). These observations suggest that reusing Places-CNN filters for conv1 and other bottom layers may not be a good idea. Moreover, since filters also represent tunable parameters, this results in a model with too many parameters that is difficult to train with limited data.

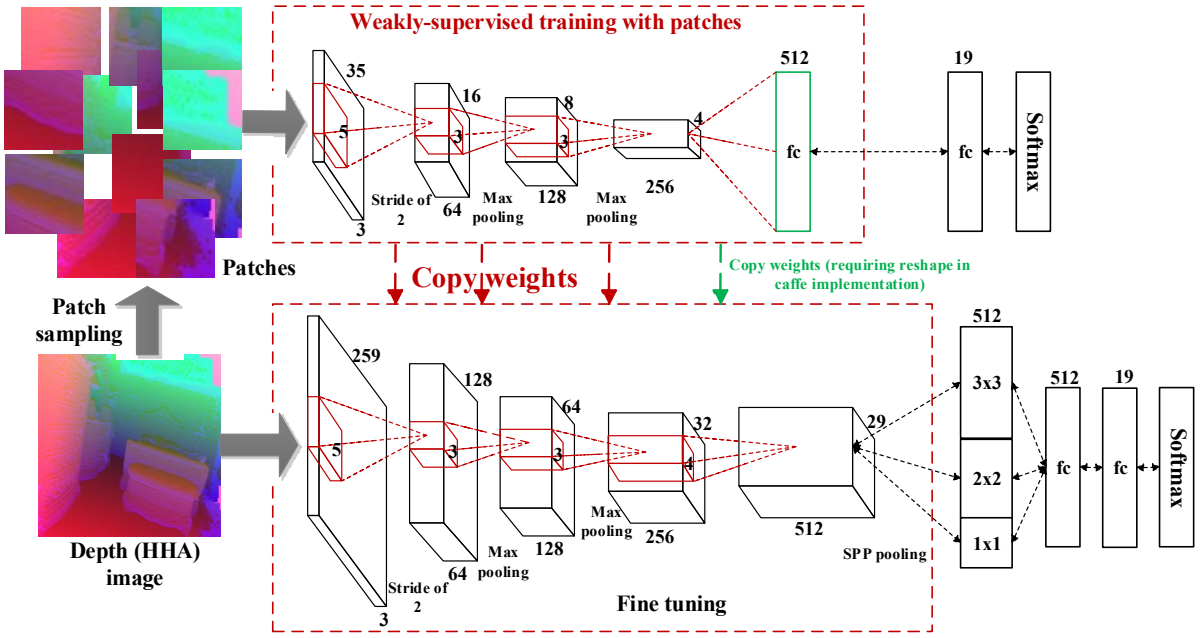


Figure 5: Two-step learning of depth CNNs combining weakly supervised pretraining and fine tuning.

Weakly supervised pretrained CNN

It is difficult to learn deep CNN from scratch with depth images, due to the lack of enough training data. We alternatively train CNN from scratch with different architectures (see Fig. 6a), from shallow to deep. The visualization of conv1 (architecture with two convolutional layers) is illustrated in the top row of Fig. 4e. However, we cannot observe any visual regularities from the visualizations, even though this is a more shallow network.

So in order to adapt the amount of training data and the complexity (i.e. number of parameters) of the model, we modify slightly the training procedure. First, we reduce the support of the CNN by working on patches rather than on the whole image. Thus networks with simpler architectures with fewer trainable parameters are capable. Additionally, we can extract multiple patches from a single image, increasing the amount of training data. Second, we include weakly supervised training. Patches typically cover objects or parts, not the whole scene. However, since we do not have those local labels we use the scene category as weak label, since we know that all patches in a given image belong to the same scene. We refer to this network as weakly supervised patch-CNN (WSP-CNN).

Weak supervision on patches

The weakly-supervised strategy can be used for Alex-CNN training. We first sample a grid of 4×4 patches with 99×99 pixels for weakly-supervised pretraining, and fine tune it with full images. When switching the architecture from WSP-CNN to full Alex-CNN, the amount of connections in fc6 changes. Thus, only the weights of the convolutional layers of the pretrained WSP-CNN are copied for fine tuning, similarly as other weakly supervised methods(Durand, Thome, and Cord 2016; Bilen and Vedaldi 2016). Fig. 3 shows that using weak supervision on patches (WSP) significantly outperforms training with full image (compare *Train-Alex-CNN (WSP)* vs *Train-Alex-CNN (scratch)*). Furthermore, in the conv1 filters shown in Fig. 4f (WSP) the depth specific-patterns are much more evident than in Fig. 4e (full image). Nevertheless, they still show a significant amount of noise, which suggests that AlexNet is still too complex, and perhaps the kernels may be too big for depth data.

Since the complexity of depth images is significantly lower than that of RGB images (e.g. no textures), we reduced the size of the kernels in each layer to train our WSP-CNN, which consists of C (C=3 works best after evaluation) convolutional layers (see Fig 5 top row for the architecture detail). The sizes of the kernelsare 5×5 (stride 2), 3×3

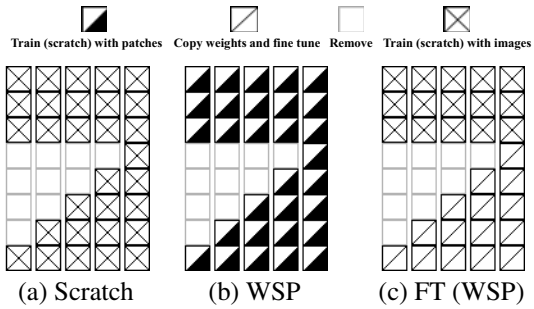


Figure 6: Training strategies for Alex-CNN variants with depth images, (a) from scratch, (b) weakly-supervised with patches, and (c) fine-tuned after weakly supervised trainig with patches.

and 3×3, and the size of max pooling is 2×2, stride 2. We sample a grid of 7×7 patches with 35×35 pixels for weakly-supervised training.

Depth-CNN

Fig 5 bottom shows the full architecture of our depth-CNN (D-CNN). After training the WSP-CNN, we transfer the weights of the convolutional layers. The output of conv4 in D-CNN is $29 \times 29 \times 512$, almost 50 times larger than the output of pool5 (size of $6 \times 6 \times 256$) in Alex-CNN, which leads to 50 times more parameters in this part. In order to reduce the parameters, we include three spatial pyramid pooling (SPP) (He et al. 2014) layers with size of 29×29,15×15, 10×10. SPP also captures spatial information and allows us to train the model end-to-end. The proposed *Train-D-CNN (WSP)* outperforms both fine tuning and weakly-supervised training of Alex-CNN (see in Fig. 3). Comparing the visualizations in Fig. 4, the proposed WSP-CNN and D-CNN learn more representative kernels, supporting the better performance. This also suggests that a smaller kernel size is suitable for depth data.

RGB-D fusion

Most previous works use two independent networks for RGB and depth, optimizing in different independent stages the fusion parameters, fine tuning and classification(Zhu, Weibel, and Lu 2016; Wang et al. 2016; 2015). In contrast, we integrate both RGB-CNN, depth-CNN and the fusion procedure into an integrated RGB-D-CNN, which can be trained end-to-end, jointly learning the fusion parameters and fine tuning both RGB layers and depth layers of each branch. As fusion mechanism, we use a network with two fully connected layers followed by the loss (i.e. fusion network), on top of the concatenated feature $[F_{rgb}, F_d]$, where $F_{rgb} \in R^{D_r \times 1}$ and $F_{depth} \in R^{D_d \times 1}$ are the RGB and depth features, respectively. The first layer of the fusion network learns modality-specific projections $\bar{W} = [W_{rgb}, W_{depth}]$ to a common feature space as

$$F_{rgbd} = [W_{rgb}\, W_{depth}] \begin{bmatrix} F_{rgb} \\ F_{depth} \end{bmatrix} = \bar{W} \begin{bmatrix} F_{rgb} \\ F_{depth} \end{bmatrix} \quad (1)$$

where $W_{rgb} \in R^{D_{rgbd} \times D_r}$ and $W_{depth} \in R^{D_{rgbd} \times D_d}$ are the modality-specific projection matrices.

Recent works exploit metric learning(Wang et al. 2015), Fisher vector(Wang et al. 2016) and correlation analysis(Zhu, Weibel, and Lu 2016) to reduce the redundancy in the joint RGB-D representation. It is important to note that this step is particularly effective when RGB and depth features are significantly correlated. This is likely to be the case in recent works when both RGB and depth feature extractors are fine tuned versions of the same CNN model (e.g. Places-CNN). In our case depth models are learned directly from the data and independently from RGB, so they are already much less correlated, even without explicit multimodal analysis.

Experiments

Dataset. We evaluate our approach in two datasets: NYU depth dataset second version (NYUD2) (Silberman et al. 2012) and SUN RGB-D(Song, Lichtenberg, and Xiao 2015). The former is a relatively small dataset with 27 indoor categories, but only a few of them are well represented. Following the split in (Silberman et al. 2012), all 27 categories are reorganized into 10 categories, including 9 most common categories and an 'other' category consisting of the remaining categories. The training/test split is 795/654 images. SUN RGB-D contains 40 categories with 10335 RGB-D images. Following the publicly available split in (Song, Lichtenberg, and Xiao 2015; Wang et al. 2016), the 19 most common categories are selected, consisting of 4,845 images for training and 4,659 images for test.

Table 1: Ablation study for different models (accuracy %).

Arch.	Alex-CNN				D-CNN
Weights	Places-CNN		Scratch		Scratch
Layer	-	FT	Train	WSP	WSP
pool1	17.2	20.3	22.3	23.5	**25.3**
pool2	25.3	27.5	26.8	30.4	**33.9**
conv3	27.6	29.3	29.8	**35.1**	34.6
conv4	29.5	32.1	-	-	**38.3**
pool5	30.5	35.9	-	-	-
fc6	30.8	36.5	30.7	36.1	-
fc7	30.9	37.2	32.0	36.8	**40.5**
fc8	-	37.8	32.8	37.5	**41.2**

Table 2: Comparison of depth on SUN RGB-D

	Method	Acc.(%)
Proposed	D-CNN	41.2
	D-CNN (wSVM)	**42.4**
State-of-the-art	R-CNN+FV(Wang et al. 2016)	34.6
	FT-PL(Wang et al. 2016)	37.5
	FT-PL+SPP	37.7
	FT-PL+SPP (wSVM)	38.9

FT: Fine tuned, PL: Places-CNN

Classifier and features. Since we found that training linear SVM classifiers with the output of the fully connected layer increases slightly performance, all the following results use SVMs, if not specified.

- (wSVM): this variant uses category-specific weights during SVM training to compensate the imbalance in the training data. The weight $w = \{w_1 ... w_K\}$ of each category k is computed as $w_k = \left(\frac{\min_{i \in K} N_i}{N_k}\right)^p$, where N_k is the number of training images of the k_{th} category. We select $p = 2$ by cross validation.

Evaluation metric. Following (Song, Lichtenberg, and Xiao 2015; Wang et al. 2016), we report the average precision over all scene classes for both datasets.

Evaluations on SUN RGB-D

Ablation study. We compare D-CNN and Alex-CNN on SUN RGB-D depth data. The outputs of different layers are

Table 3: Comparisons of RGB-D data on SUN RGB-D

Method		CNN models RGB	CNN models Depth	Accuracy (%) RGB	Accuracy (%) Depth	Accuracy (%) RGB-D
Baseline	Concate.	PL	PL	35.4	30.9	39.1
	Concate.	FT-PL	FT-PL	41.5	37.5	45.4
	Concate. (wSVM)	FT-PL	FT-PL	42.7	38.7	46.9
Proposed	RGB-D-CNN	FT-PL	FT-PL	41.5	37.5	48.5
	RGB-D-CNN	FT-PL	D-CNN	41.5	41.2	50.9
	RGB-D-CNN (wSVM)	FT-PL	D-CNN	42.7	42.4	**52.4**
State-of-the-art	(Zhu, Weibel, and Lu 2016)	FT-PL	FT-PL	40.4	36.5	41.5
	(Wang et al. 2016)	FT-PL + R-CNN	FT-PL + R-CNN	40.4	36.5	48.1

FT: Fine tuned, PL: Places-CNN

Table 4: Comparisons on NYUD2 in accuracy(%)

Method	Features RGB	Features Depth	Acc.
Baseline methods			
RGB	FT-PL		53.4
Depth		FT-PL	51.8
Concate.	FT-PL	FT-PL	59.5
Proposed methods			
Depth		D-CNN	56.4
RGB-D-CNN	FT-PL	D-CNN	**65.8**
State-of-the-art			
(Gupta et al. 2014b)			45.4
(Wang et al. 2016)			63.9

FT: Fine tuned, PL: Places-CNN

used as features to train the SVM classifiers. We select 5 different models for comparison in Table 1. We use the Alex-CNN architecture with only 3 bottom convolutional layers for training from scratch. With Alex-CNN architecture, the bottom layers (pool1 to conv3) trained from scratch perform better than the transferred from Places-CNN, even though the top layers are worse. Using weakly-supervised training on patches (WSP), the performance is comparable to fine tuned Places-CNN for top layers and better for bottom layers, with a smaller model and without relying on Places data. D-CNN consistently achieves the best performance.

Comparisons with depth data. We compare to related methods on depth recognition in Table 2. For the fair comparison, we also implement SPP on Places-CNN for the fine tuning. Our D-CNN outperforms FT-Places-CNN+SPP with 3.5% in accuracy. When both models using weighted SVM for training, our D-CNN works even better.

RGB-D fusion. We compare to the state-of-the-art works (Zhu, Weibel, and Lu 2016; Wang et al. 2016) of RGB-D indoor recognition in Table 3, where a discriminative RGB-D fusion CNN is proposed in (Zhu, Weibel, and Lu 2016) and a joint feature fusion of RGB-D, scene and objects is proposed in (Wang et al. 2016). The proposed RGB-D-CNN outperforms the RGB-D fusion method in (Wang et al. 2016) with 2.7% with linear SVM, 4.3% with weighted SVM, without including external training of R-CNN (Gupta et al. 2014a) as in that approach.

Comparisons on NYUD2

We compare our RGB-D-CNN to the state-of-the-art on NYUD2 in Table 4. Gupta *et al. (Gupta et al. 2014b)* propose to encode segmentation responses as features for scene recognition. Our RGB-D-CNN largely outperforms this work, where the segmentation is based on the hand-crafted features. Comparing to RGB-D fusion in (Wang et al. 2016), we achieves a gain of 1.9% in accuracy, without including the R-CNN models as in that work.

Conclusion

Transferring deep representations within the same modality (e.g. Places-CNN fine tuned on SUN397) works well, since low-level patterns have similar distributions, and bottom layers can be reused while adjusting the more dataset-specific top layers. However, fine tuning is not that effective in inter-modal transfer, such as Places-CNN to depth in the HHA space, where low-level features require modality-specific filters. In this paper, we focus on the bottom layers, because they are more critical to represent depth data properly. By reducing the number of parameters of the network, and using weakly supervised learning over patches, the complexity of the model matches better the amount of data available. This depth representation is not only more discriminative than those fine tuned from Places-CNN but also when combined with RGB features the gain is higher, showing that both are complementary. Notice also, that we do not depend (for depth) on large datasets such as Places.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under 61322212, Grant 61532018, and Grant 61550110505, in part by the National High Technology Research and Development 863 Program of China under Grant 2014AA015202, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

References

Agrawal, P.; Girshick, R.; and Malik, J. 2014. Analyzing the performance of multilayer neural networks for object recognition. In *ECCV*.

Banica, D., and Sminchisescu, C. 2015. Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bilen, H., and Vedaldi, A. 2016. Weakly supervised deep detection networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*.

Durand, T.; Thome, N.; and Cord, M. 2016. Weldon: Weakly supervised learning of deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gupta, S.; Girshick, R.; Arbelaez, P.; and Malik, J. 2014a. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*.

Gupta, S.; Arbelaez, P.; Girshick, R.; and Malik, J. 2014b. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *Int J Comput Vis* 112:133–149.

Gupta, S.; Arbeláez, P.; Girshick, R.; and Malik, J. 2015. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision* 112(2):133–149.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1106–1114.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Quattoni, A., and Torralba, A. 2009. Recognizing indoor scenes. In *CVPR*.

Silberman, N., and Fergus, R. 2011. Indoor scene segmentation using a structured light sensor. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 601–608.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, 746–760. Berlin, Heidelberg: Springer-Verlag.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Socher, R.; Huval, B.; Bath, B.; Manning, C. D.; and Ng, A. Y. 2012. Convolutional-recursive deep learning for 3d object classification. In *NIPS*.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 567–576.

Wang, A.; Lu, J.; Cai, J.; Cham, T.-J.; and Wang, G. 2015. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Trans. on Multimedia* 17.

Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2016. Modality and component aware feature fusion for rgb-d scene classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao, J.; Hayes, J.; Ehringer, K.; Olivia, A.; and Torralba, A. 2010. SUN database: Largescale scene recognition from abbey to zoo. In *CVPR*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *NIPS*, 487–495.

Zhu, H.; Weibel, J.-B.; and Lu, S. 2016. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.