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

Xinhang Song, Luis Herranz, Shuqiang Jiang
{xinhang.song,luis.herranz,shuqiang.jiang}@vipl.ict.ac.cn
The Institute of the Computing Technology (ICT) of Chinese Academy of Sciences (CAS)

The AAAI Conference on Artificial Intelligence

Introduction

Limitations of RGB-D scene recognition

- Depth images are **difficult** to capture, **lacking** depth images for CNN training;
- **Transferring/fine tuning** from RGB to depth may not capture the **depth-specific** visual patterns, due to the large **differences** between the RGB and depth modality.

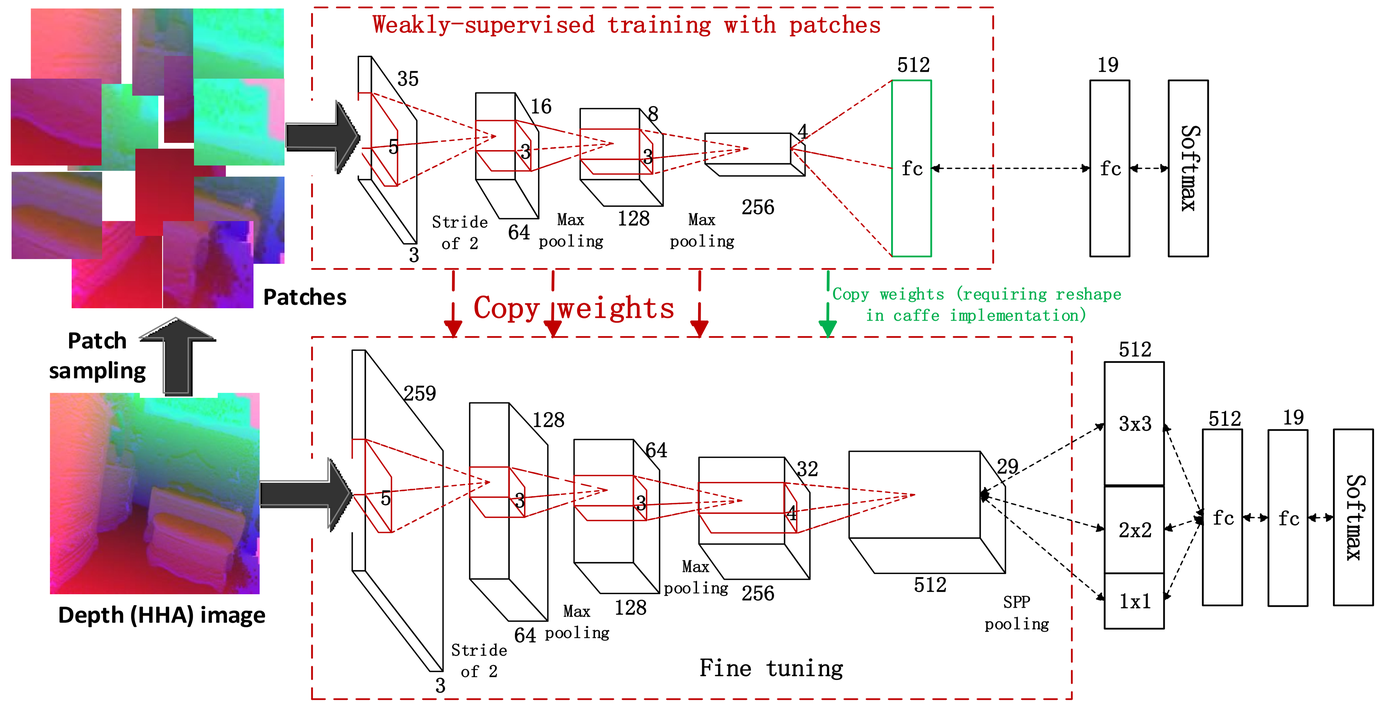
Motivation

- Training **depth-specific** CNN model with the **limited** depth training images rather than transferring/fine tuning from RGB pre-trained CNN model.

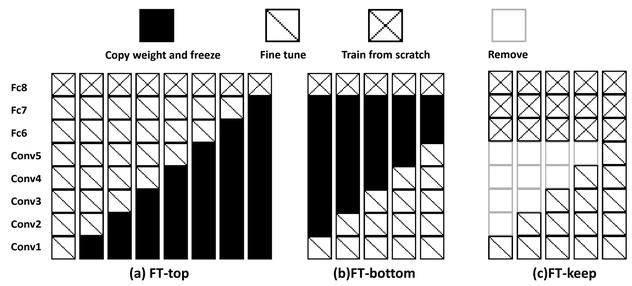
Contributions

- Analyze the large differences between RGB and depth modalities in CNN training;
- Train depth-specific CNN from scratch with **weakly-supervised** pre-training, outperforming transferring from RGB.

Two-step learning of depth CNNs combining weakly supervised pre-training and fine tuning

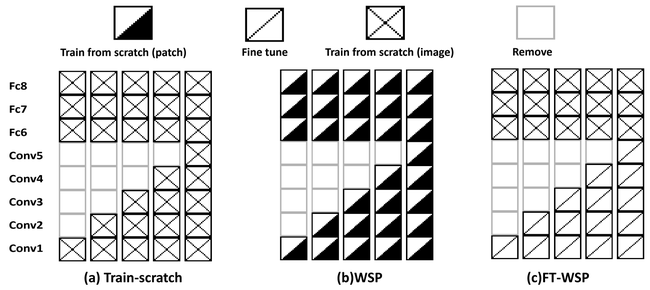


Fine tuning from RGB to depth



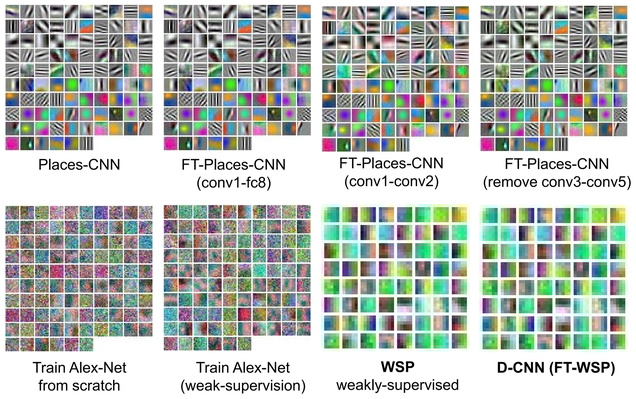
(a)FT-top: only **top** layers, bottom layers are frozen;
(b)FT-bottom: only **bottom** layers, top layers are frozen;
(c)FT-keep: bottom layers (top layers retrained and some convolutional layers **removed**).
Each column represents a particular setting.

Weakly supervised pre-trained CNN



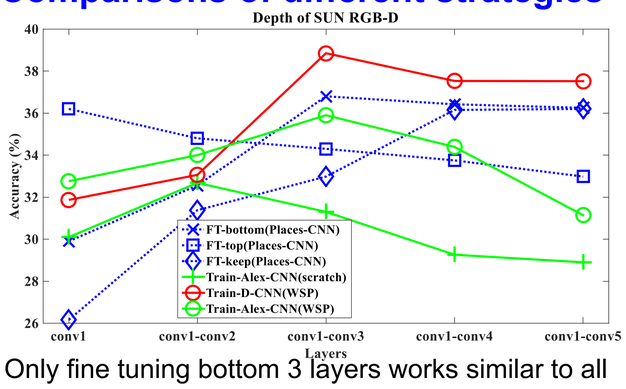
(a)Train-scratch: train from **scratch**;
(b)WSP: **Weakly-Supervised** training with **Patches;**
(c)FT-WSP: fine-tuned with images after weakly supervised training with patches.

Insight from conv1 layer



- Only a few particular filters have noticeable changes during the fine tuning process;
- Training from scratch results noisy filters.

Comparisons of different strategies



- Only fine tuning bottom 3 layers works similar to all layers;
- Architectures with 3 layers work better.

Experimental results

Table.1 Accuracy of depth recognition 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

Table.2 Comparisons of RGB-D data on SUN RGB-D

	Method	CNN models		Acc.(%)
		RGB	Depth	
Baseline	Cat	PL	PL	39.1
	Cat	FT-PL	FT-PL	45.4
	Cat(wSVM)	FT-PL	FT-PL	46.9
Proposed	Cat	FT-PL	D-CNN	50.9
	Cat(wSVM)	FT-PL	D-CNN	**52.4**
State-of-the-art	(Zhu, Weibel, and Lu 2016)			41.5
	(Wang et al. 2016)			48.1

FT: Fine tuned, PL: Places-CNN, Cat: concatenation

- We release our (SUN RGB-D dataset) pre-trained models of WSP-CNN and D-CNN in https://github.com/songxinhang/D-CNN. Note that the WSP-CNN can be efficiently fine tuned to other RGB-D datasets, e.g., NYU2.

[1] Zhu, H.; Weibel, J.-B.; and Lu, S. 2016. Discriminative multimodal feature fusion for rgbd indoor scene recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[2] 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).