Deep Co-occurrence Feature Learning for Visual Object Recognition

Ya-Fang Shih*, Yang-Ming Yeh*, Yen-Yu Lin, Ming-Fang Weng, Yi-Chang Lu, Yung-Yu Chuang
*: equal contribution

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Get codes!

Fine-grained recognition

Highly relies on discriminative parts to capture subtle inter-class variations

Part-based methods

- Appearance of parts + Spatial relationship between parts
- Robust to deformations and occlusions
- More discriminative in the case of subtle inter-class variations

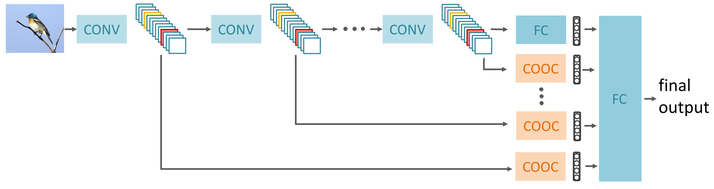
Limitations of previous work of integrating part-based representation into CNN

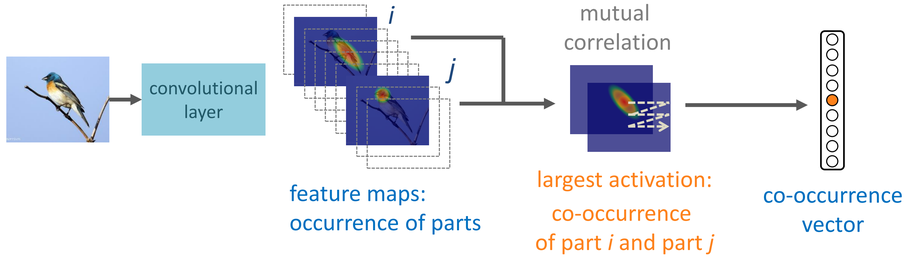
- Pre-defined parts and part number
- High annotation cost
- Complicated models

Co-occurrence layer

$$X_{ij} = \max_{o} \sum_{p} A_{p,i} A_{p+o,j}$$

A: feature maps
$A_{p,i}$: location p (2D) on feature map i
o: the offset (2D) with the maximal response
X: co-occurrence vector
X_{ij}: the co-occurrence value of feature map i and feature map j

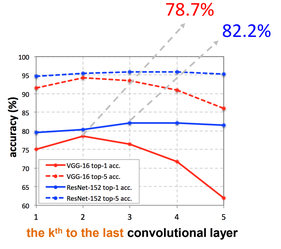




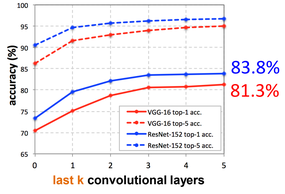
- No pre-defined parts
- Part annotation free
- Easily applicable as a building block in CNN

Properties
- Differentiable
- Nonlinear across neurons
- Rotation- and translation-invariant
- Coarse-to-fine information
- No extra parameters introduced

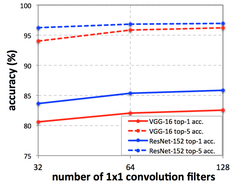
Experiments



Baseline VGG: 70.4%
Baseline ResNet-152: 73.3%
Co-occurrence Feature: 1024d



Complementary between layers



Apply the co-occurrence layer to the last 3 convolutional layers with 128 1x1 filters

Quantitative results

method	network	part annotation	accuracy (%)
Liu et al. [CVPR'15]	Caffe		73.5
Zhang et al. [ECCV'14]	Caffe	✓	73.9
Branson et al. [BMVC'14]	Caffe	✓	75.7
Simon et al. [ICCV'15]	VGG		81.0
Krause et al. [CVPR'15]	VGG		82.0
Huang et al. [CVPR'16]	Caffe	✓	76.6
Xiao et al. [CVPR'15]	AlexNet+VGG		77.9
Wang et al. [ICCV'15]	VGGx3		81.7
Lin et al. [ICCV'15]	VGGx2		84.1
Jaderberg et al. [NIPS'15]	Inceptionx4		84.1
Ours	**VGG**		**83.6**
Ours	**ResNet-152**		**85.8**

Visualization

The detected most influential pair of parts in each class:
- interpretable, meaningful, and consistent in a class