Deep Spatial Pyramid Ensemble for Cultural Event Recognition

Xiu-Shen Wei, Bin-Bin Gao and Jianxin Wu*
National Key Laboratory for Novel Software Technology, Nanjing University

Dec. 12, 2015 Santiago, Chile

Outline

- Background
- Deep Spatial Pyramid (DSP) and its ensemble
- Implementation details
- Experimental results

Background

Cultural Event Recognition

Deep Spatial Pyramid (DSP)

CNN:

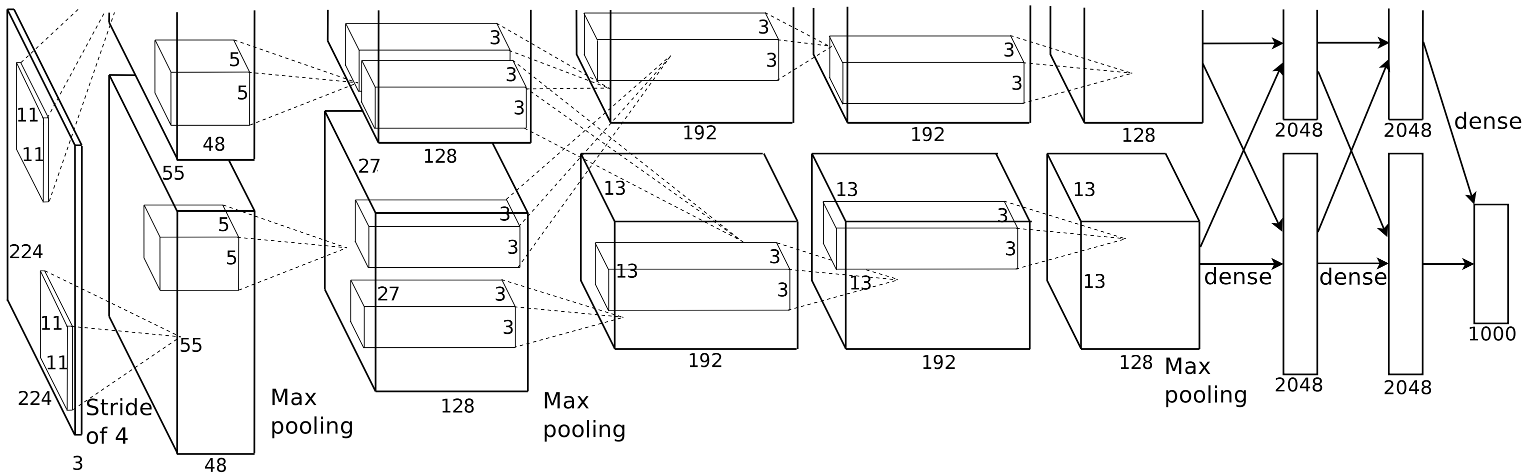
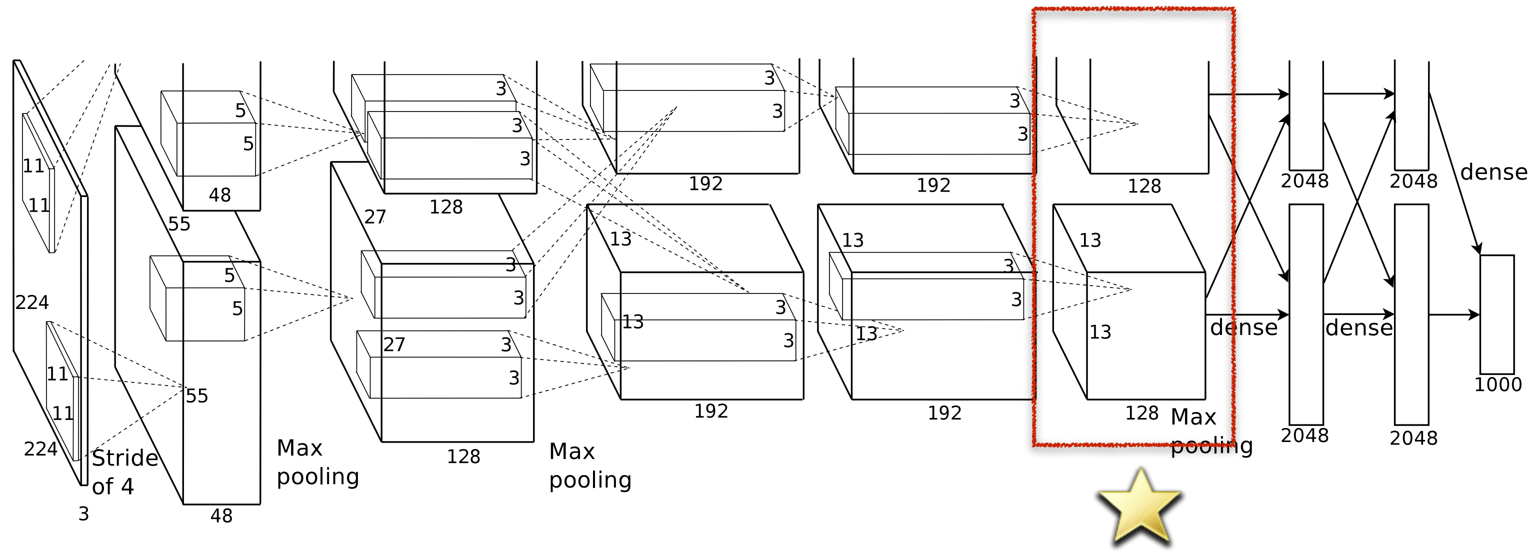


Figure courtesy of [A. Krizhevsky et al., NIPS' 12]

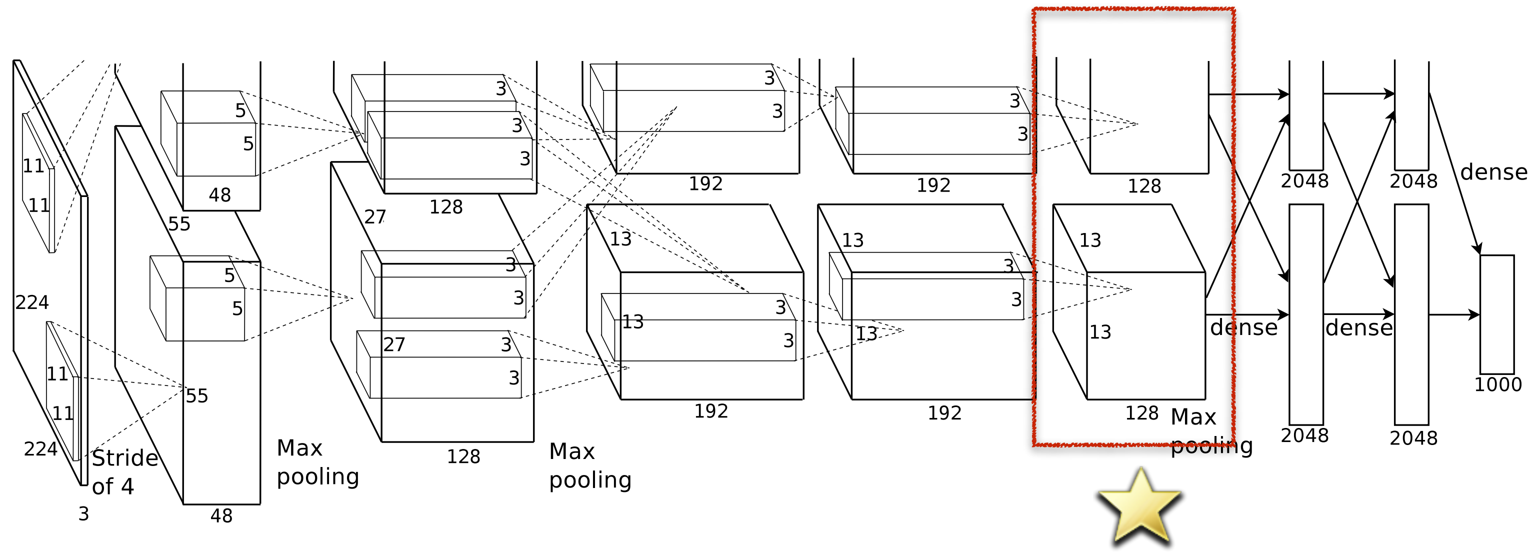
Deep Spatial Pyramid (DSP)

CNN:



Deep Spatial Pyramid (DSP)

CNN:



DSP:

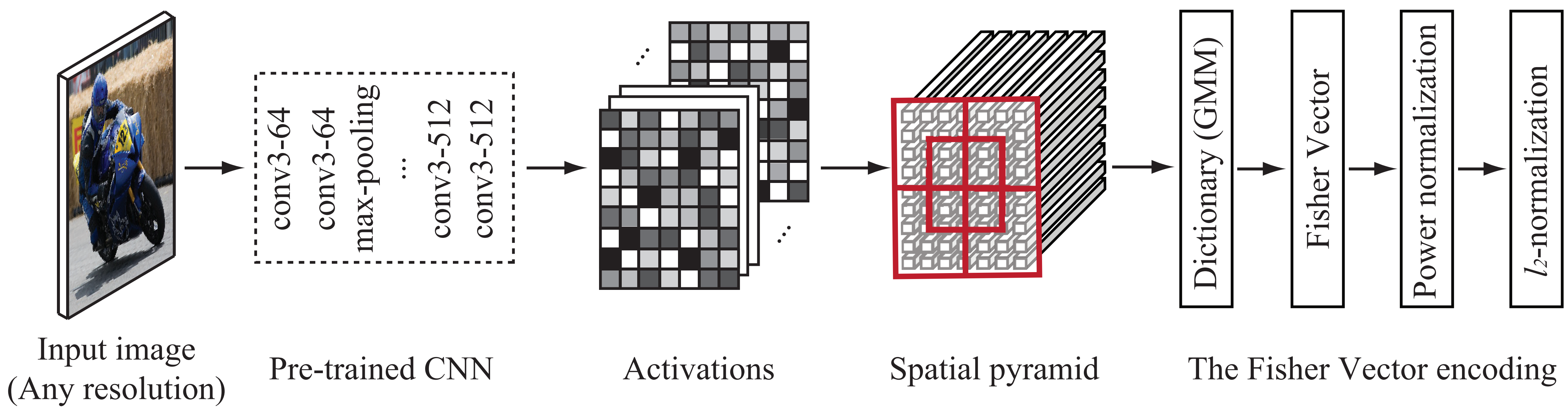
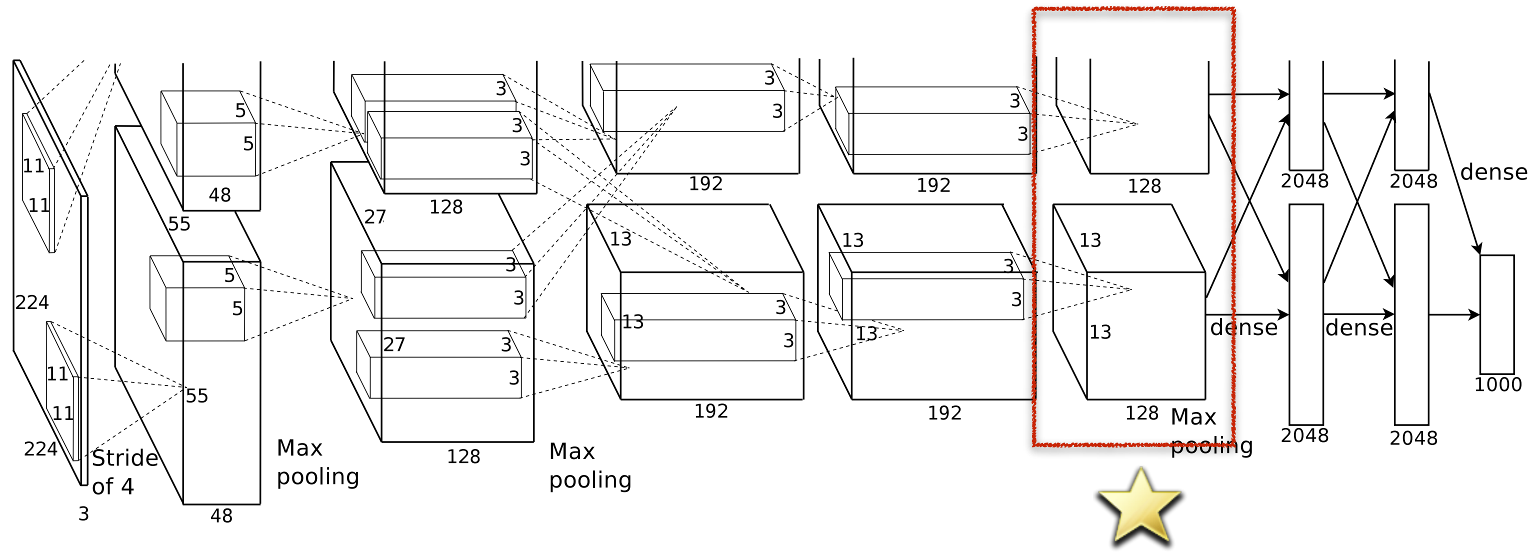


Figure courtesy of [A. Krizhevsky et al., NIPS' 12]

Deep Spatial Pyramid (DSP)

CNN:



DSP:

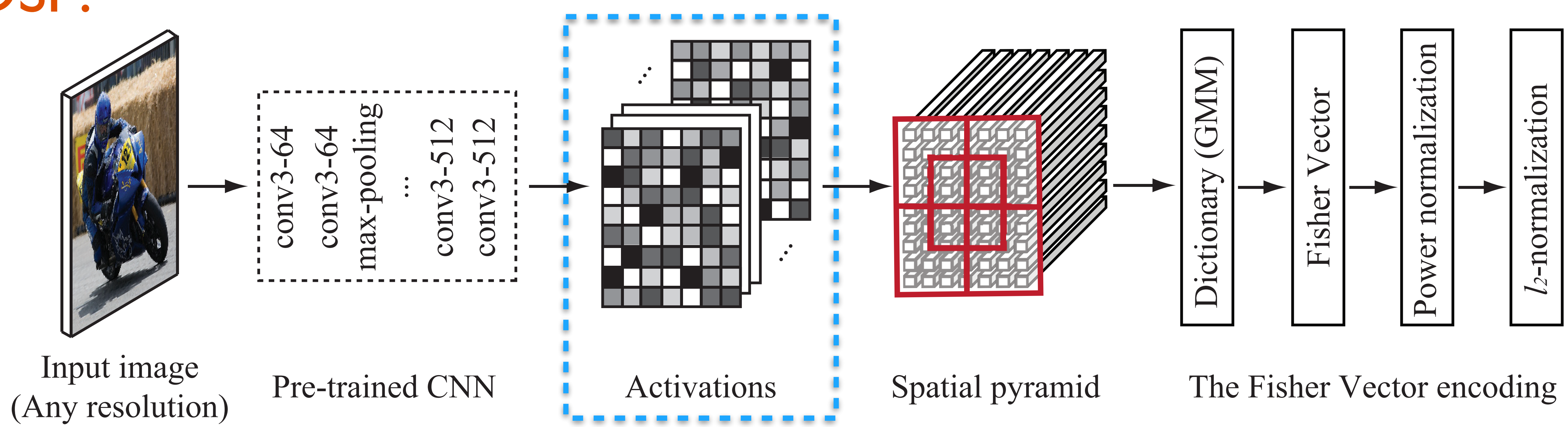
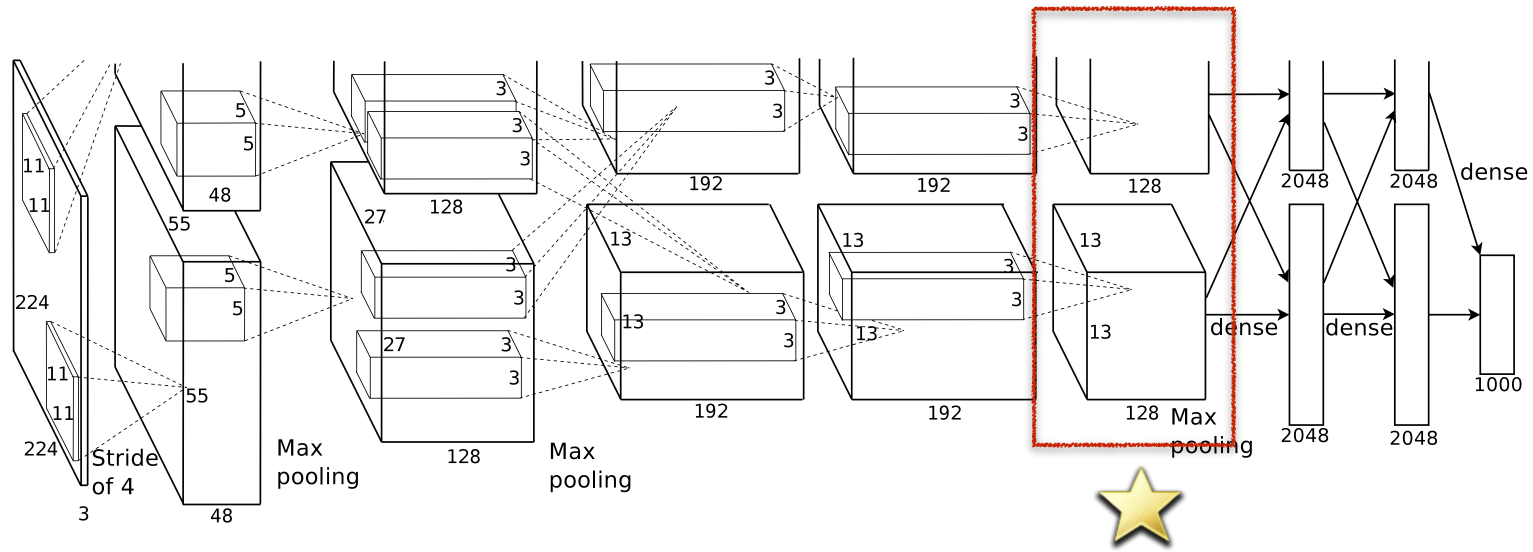


Figure courtesy of [A. Krizhevsky et al., NIPS' 12]

Deep Spatial Pyramid (DSP)

CNN:



DSP:

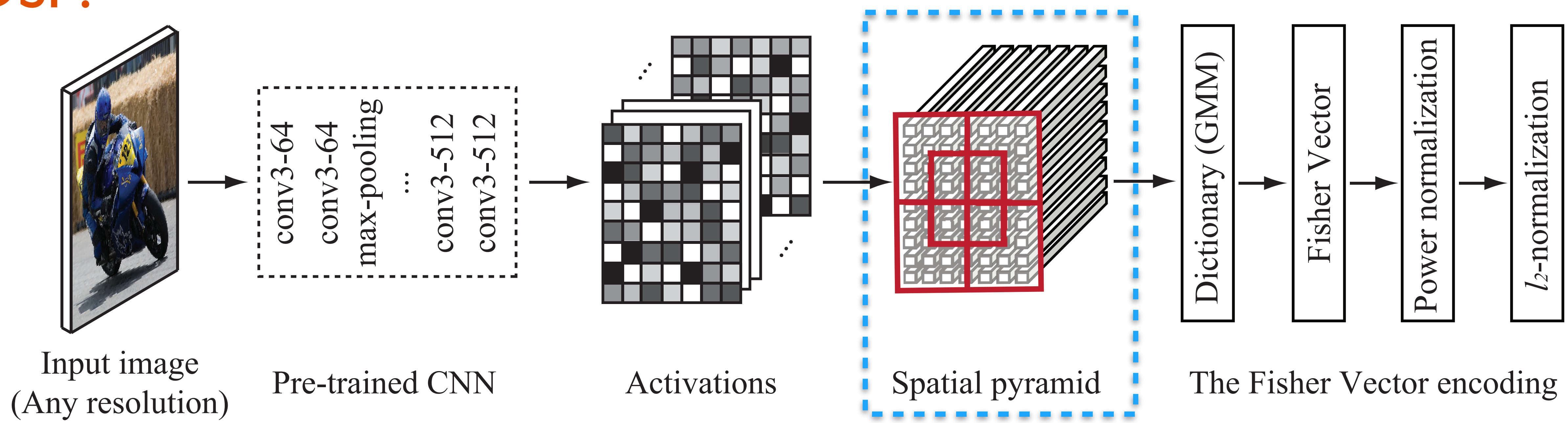


Figure courtesy of [A. Krizhevsky et al., NIPS' 12]

DSP (con't)

l_2 matrix normalization in DSP:

d-dimentional deep desctrptors

$$\boxed{\boldsymbol{x}_t} \leftarrow \boldsymbol{x}_t / \boxed{\|X\|_2}$$

matrix spectral norm

DSP (con't)

l_2 matrix normalization in DSP:

d-dimentional deep desctrptors

$$\boxed{\boldsymbol{x}_t} \leftarrow \boldsymbol{x}_t / \boxed{\|X\|_2}$$

matrix spectral norm

Results of the different normalization methods:

	Caltech101	Stanford40	Scene15	Indoor67
No	90.63	74.84	90.75	71.20
ℓ_2 vector	92.02	73.41	90.92	74.03
ℓ_2 matrix	**92.56**	**78.43**	**90.99**	**74.55**
PCA+ℓ_2 matrix	91.95	75.69	90.22	71.79

DSP (con't)

Encoding deep descriptors by FV:

$$\boldsymbol{f}_{\boldsymbol{\mu}_k}(X) = \frac{1}{\sqrt{\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{\boldsymbol{x}_t - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k} \right) ,$$

$$\boldsymbol{f}_{\boldsymbol{\sigma}_k}(X) = \frac{1}{\sqrt{2\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left[\frac{(\boldsymbol{x}_t - \boldsymbol{\mu}_k)^2}{\boldsymbol{\sigma}_k^2} - 1 \right] .$$

DSP (con't)

Learning And Mining from DatA

Encoding deep descriptors by FV:

$$\boldsymbol{f}_{\boldsymbol{\mu}_k}(X) = \frac{1}{\sqrt{\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{\boldsymbol{x}_t - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k} \right) ,$$

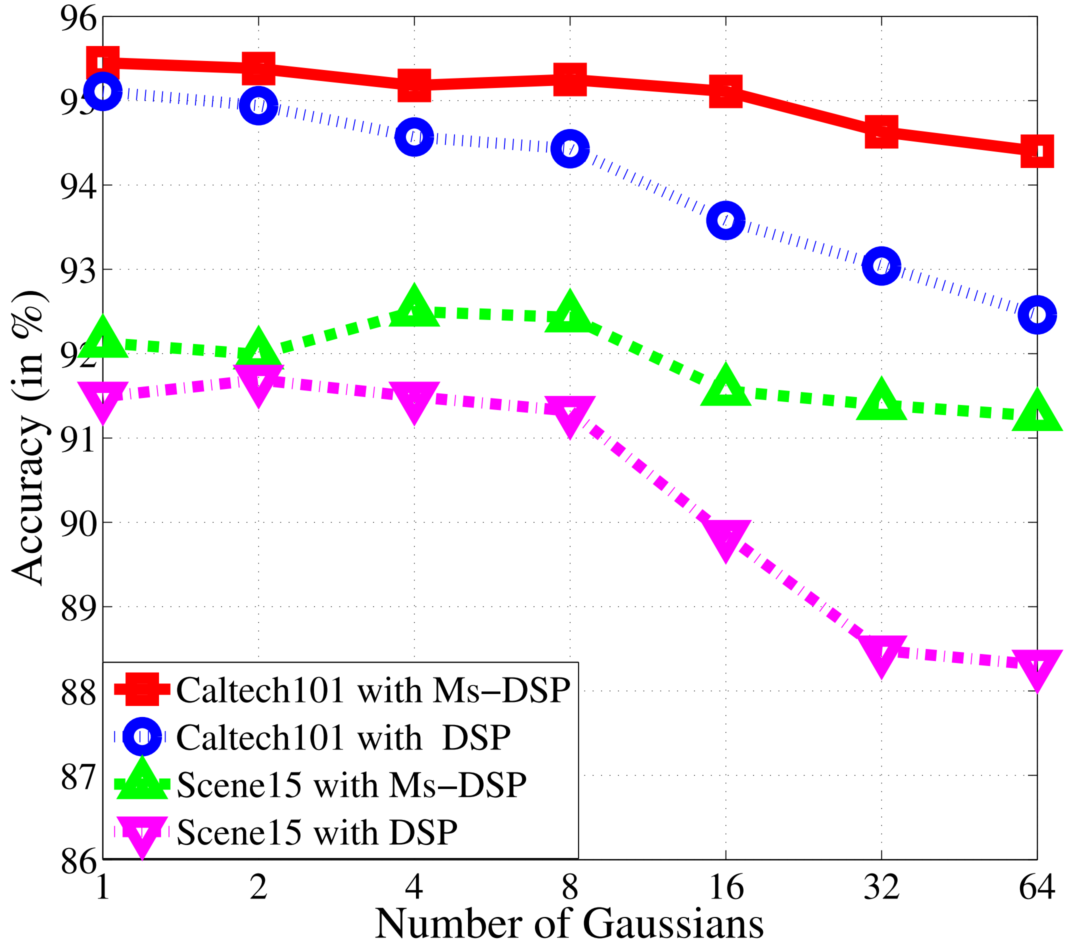
$$\boldsymbol{f}_{\boldsymbol{\sigma}_k}(X) = \frac{1}{\sqrt{2\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left[\frac{(\boldsymbol{x}_t - \boldsymbol{\mu}_k)^2}{\boldsymbol{\sigma}_k^2} - 1 \right] .$$

Multi-scale DSP:

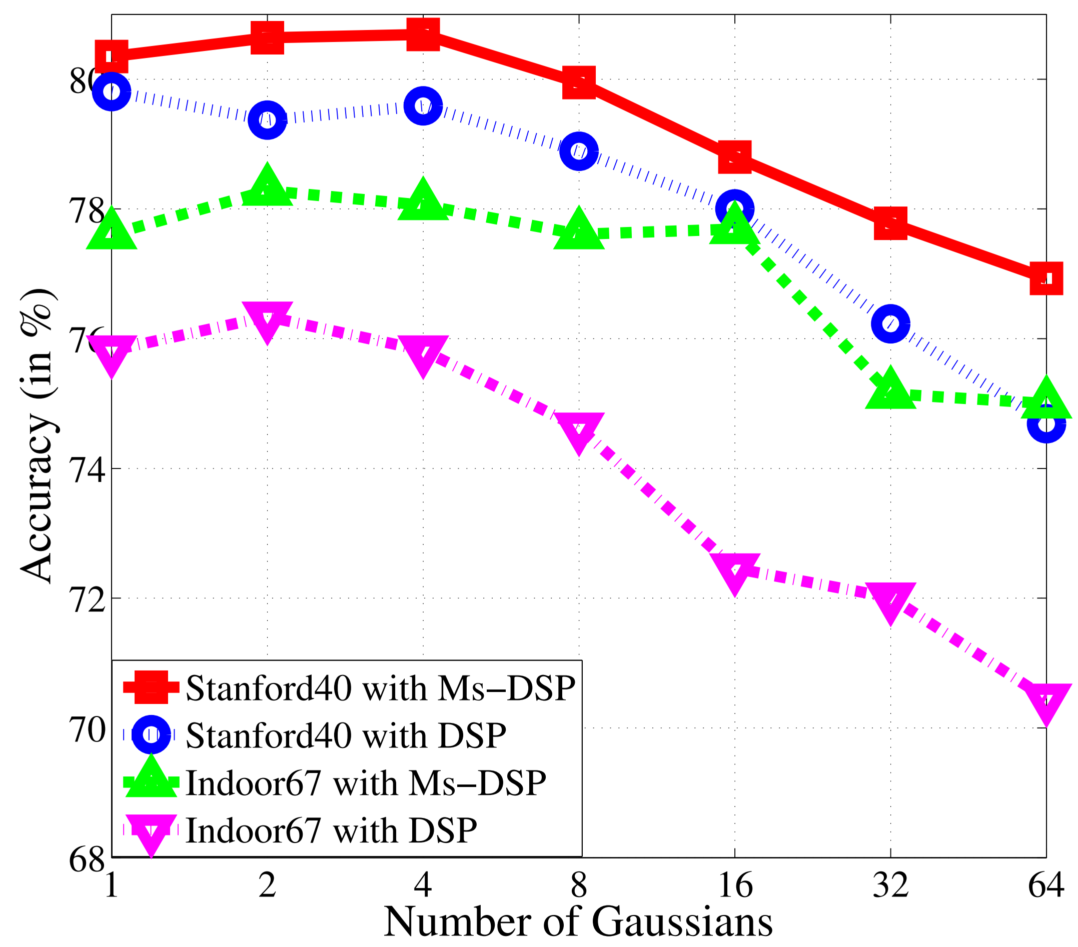
$$\boldsymbol{f}_m = \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{f}_s \qquad S = \{1.4, 1.2, 1.0, 0.8\}$$

DSP (con't)

Classification performance with different *K*:



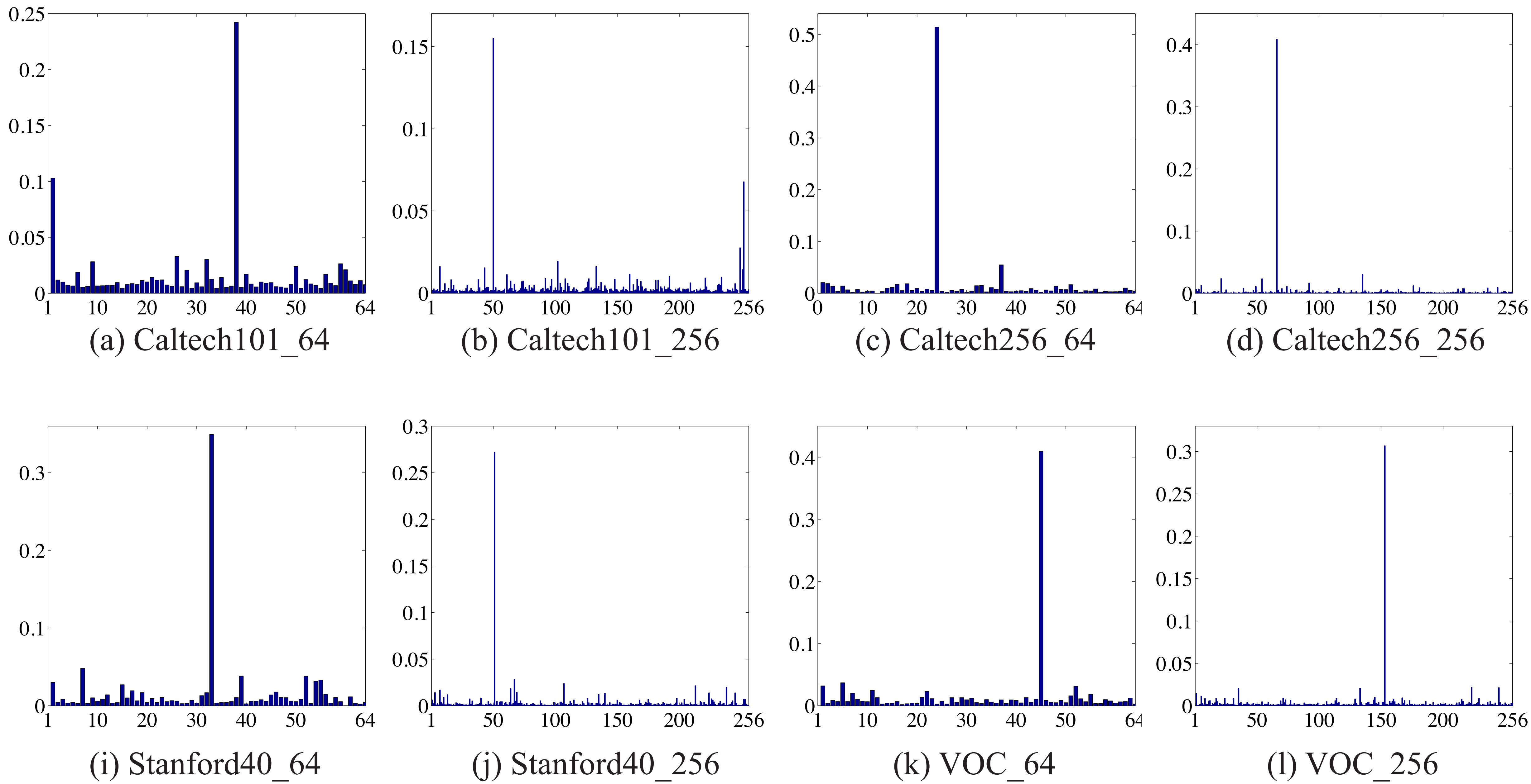
(a) *Caltech-101* and *Scene15*



(b) *Stanford40* and *Indoor67*

DSP (con't)

Plot of *w* values in DSP:



(a) Caltech101_64 (b) Caltech101_256 (c) Caltech256_64 (d) Caltech256_256

(i) Stanford40_64 (j) Stanford40_256 (k) VOC_64 (l) VOC_256

DSP (con't)

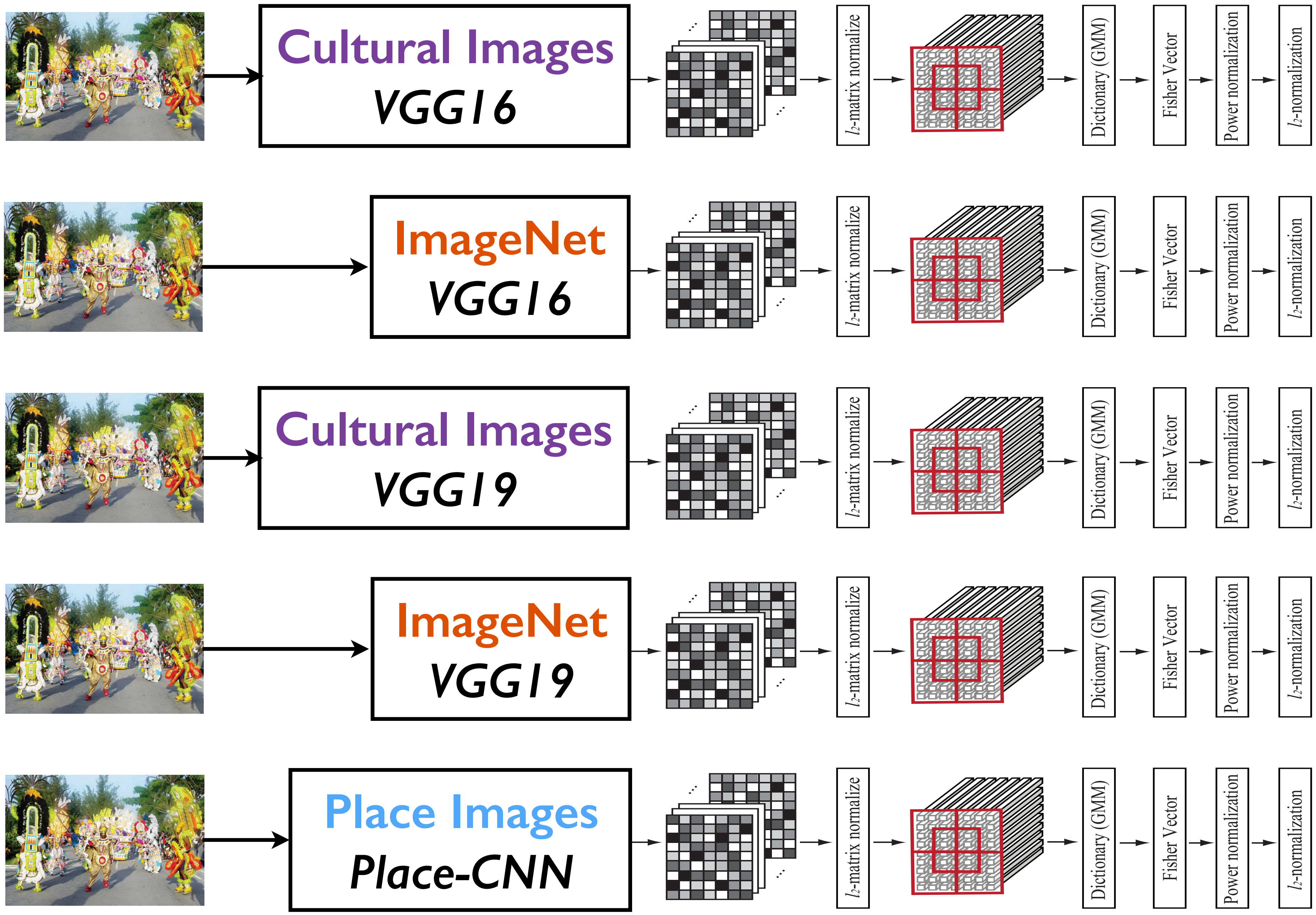
Classification accuracy/MAP comparisons:

Table 3. Recognition accuracy (or mAP) comparisons on seven datasets. The highest accuracy (mAP) of each column is marked in bold. [17]'s results were achieved using VGG Net-D and VGG Net-E, evaluation was measured by mean class recall on *Caltech-101*, *Caltech-256* instead of accuracy .

Methods	Description	Caltech-101	Caltech-256	VOC 2007	Scene15	SUN397	MIT Indoor67	Stanford40
SoA	[9]	**93.42±0.50**	-	82.44	-	-	-	-
	[7]	-	-	-	-	51.98	68.88	-
	[27]	-	-	82.13	-	-	77.56	-
	[30]	84.79±0.66	65.06±0.25	-	91.59±0.48	53.86±0.21	70.80	55.28±0.64
	[1]	88.35±0.56	77.61±0.12	82.4	-	-	-	-
	[17]	92.7±0.5 (*)	**86.2±0.3**(*)	**89.7**	-	-	-	-
Baseline	Fc_8	90.55±0.31	82.02±0.12	84.61	89.88±0.76	53.90±0.45	69.78	71.53±0.34
	$Pool_5$+FV	90.03±0.75	79.48±0.53	88.12	89.00±0.42	51.39±0.51	71.57	73.96±0.52
	DSP	94.66±0.26	84.22±0.11	88.60	91.13±0.77	57.27±0.34	76.34	79.75±0.34
Our	Ms-DSP	**95.11±0.26**	**85.47±0.14**	**89.31**	**91.78±0.22**	**59.78±0.47**	**78.28**	**80.81±0.29**

DSP Ensemble

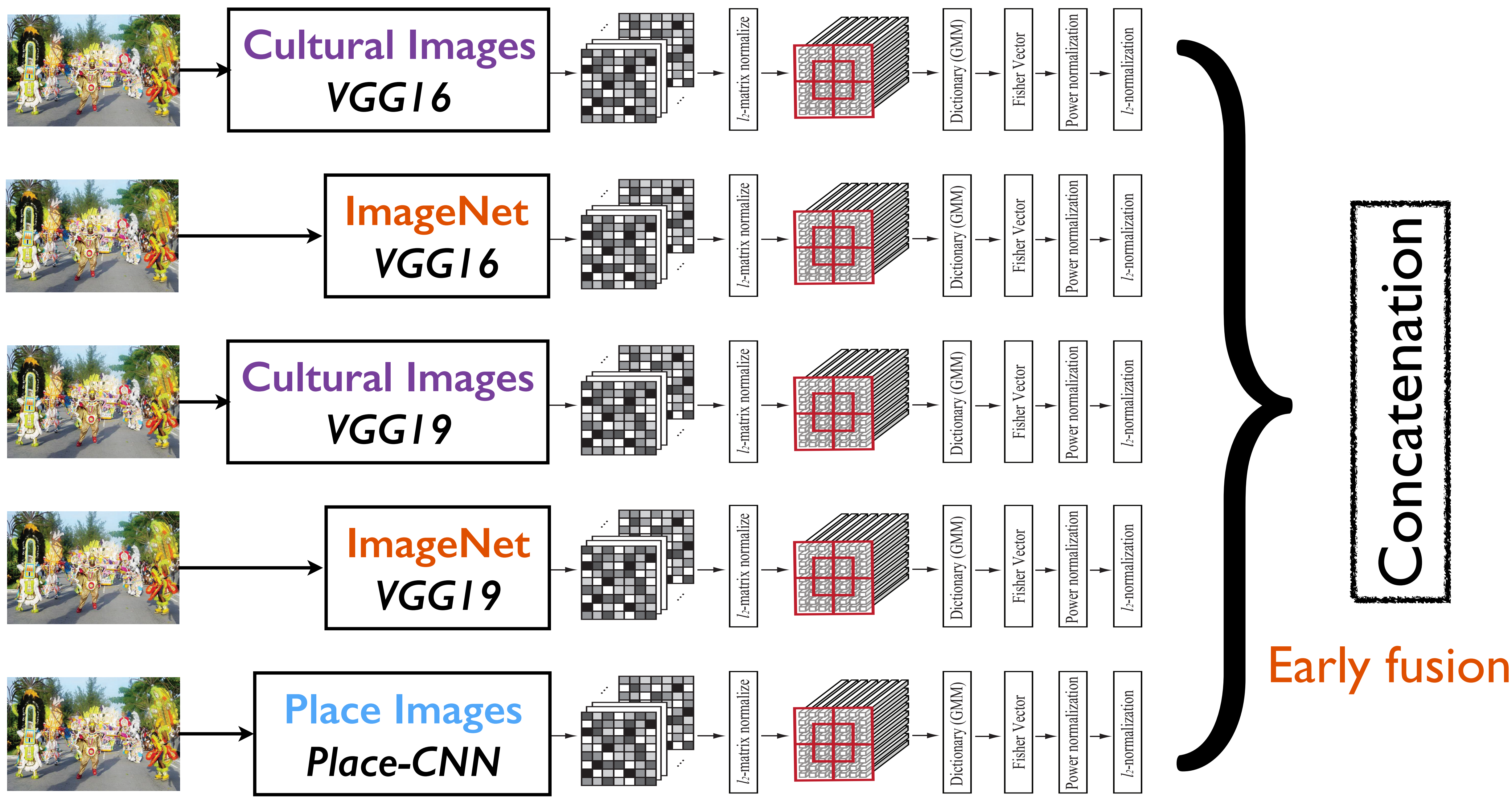
Our framework



DSP Ensemble

Learning And Mining from DatA

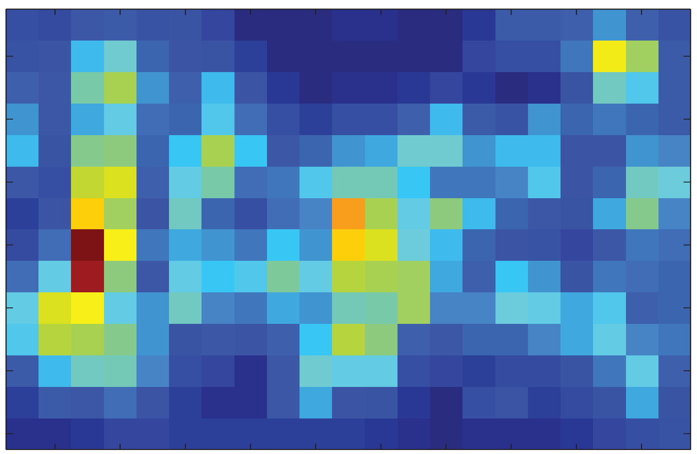
Our framework



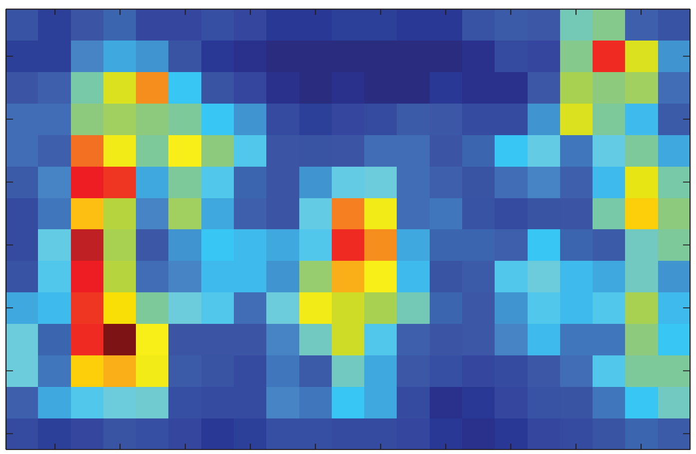
DSP Ensemble (con't)

Different feature maps from different networks:

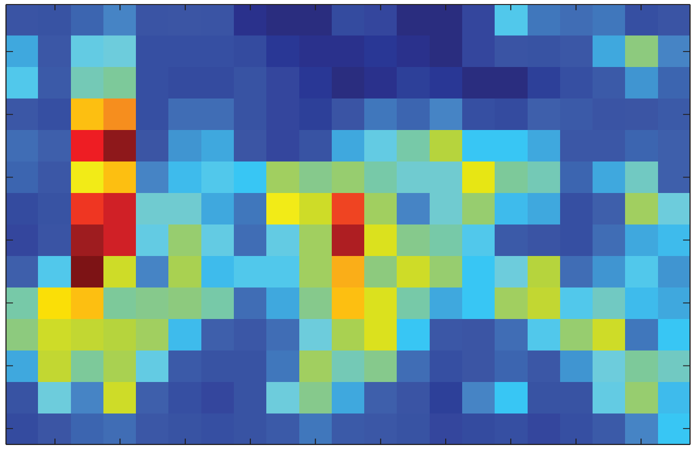
(a) The original image



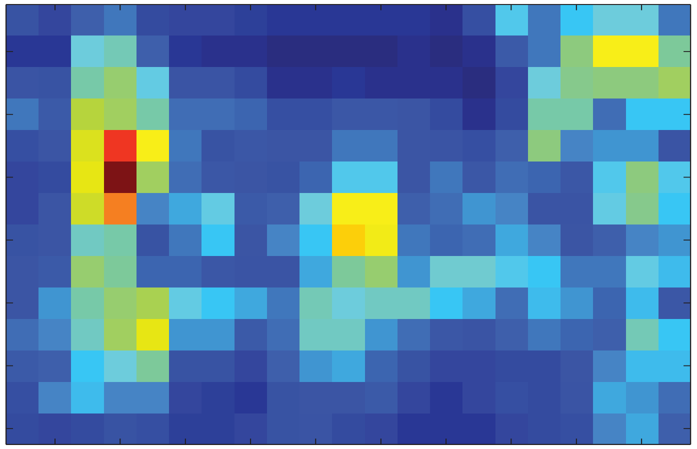
(b) Fine-tuned VGG Net-D



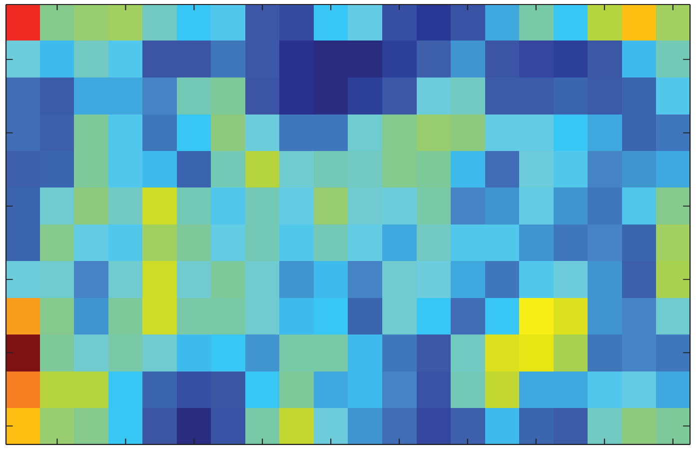
(c) VGG Net-D



(d) Fine-tuned VGG Net-E



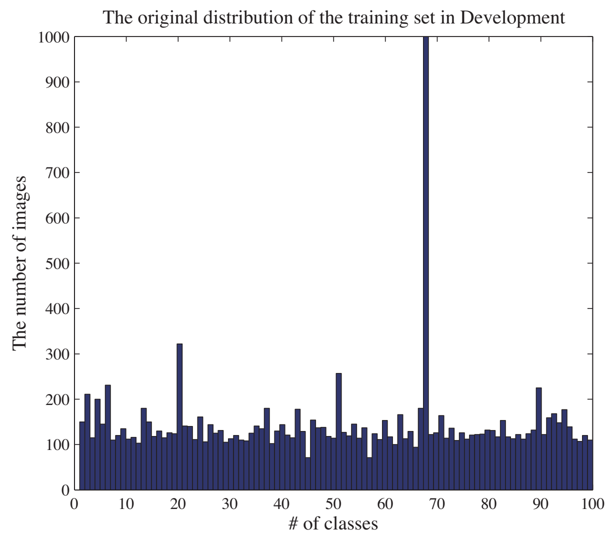
(e) VGG Net-E



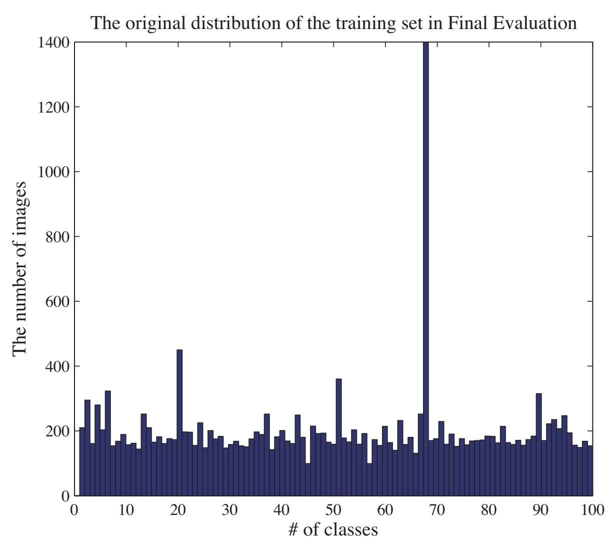
(f) Place-CNN

Implementation details

Distributions of the number of training images in Dev. and Final Evaluation:



(a)

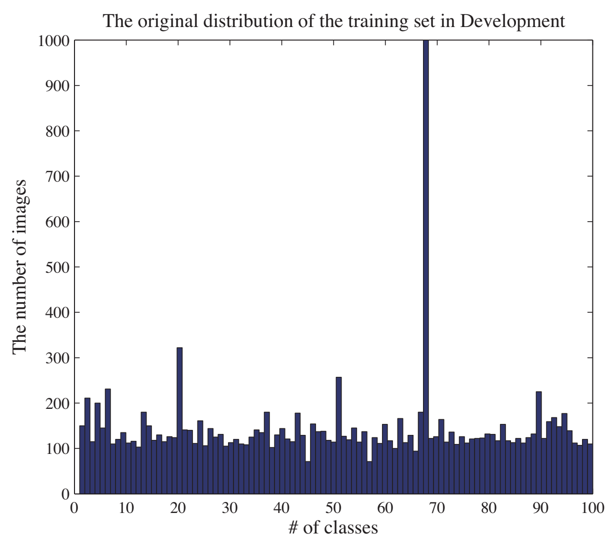


(c)

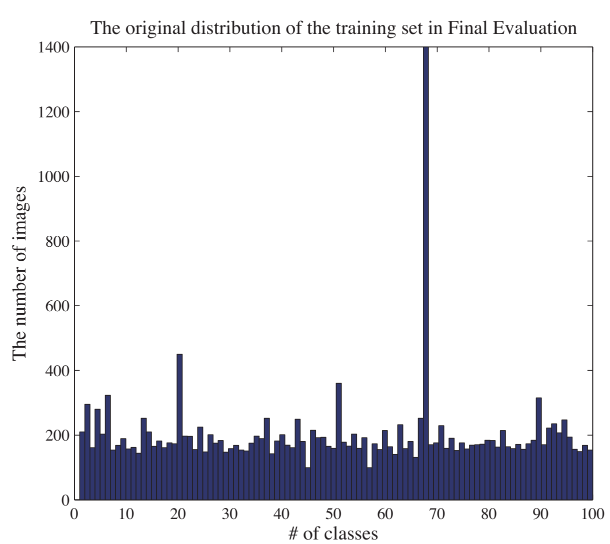
Implementation details

Learning And Mining from DatA

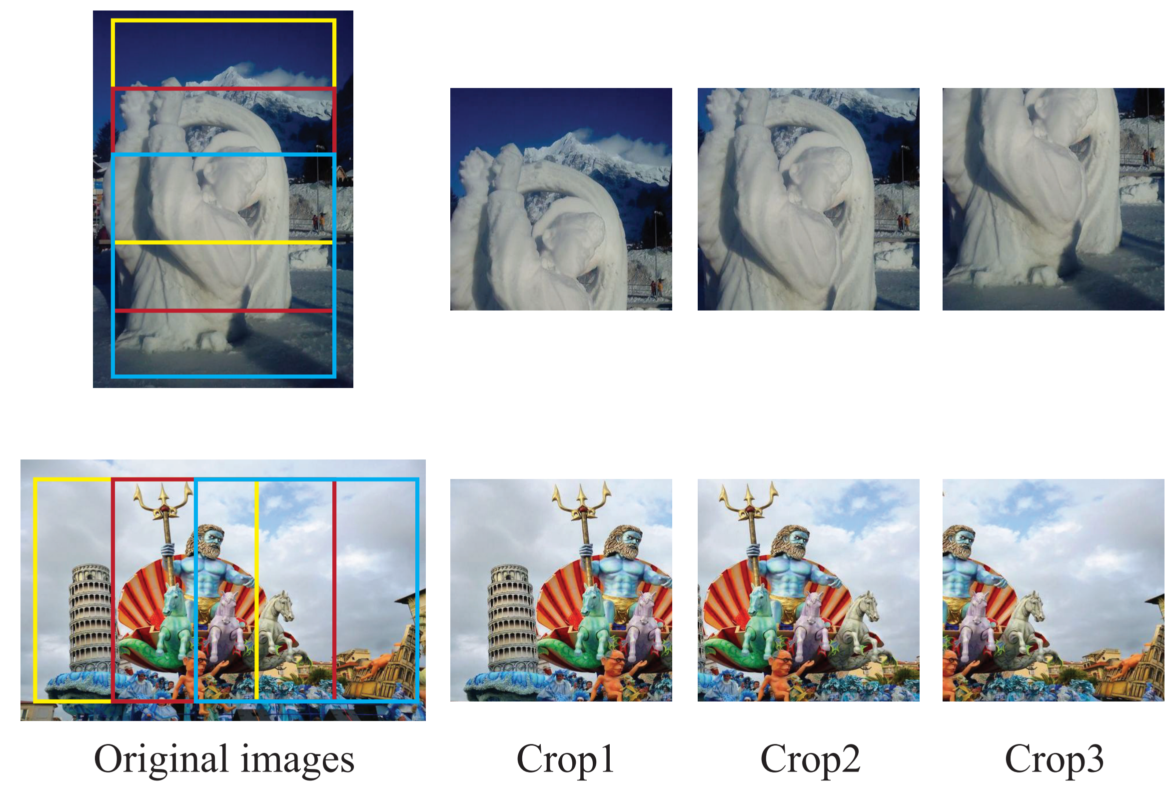
Distributions of the number of training images in Dev. and Final Evaluation:



(a)



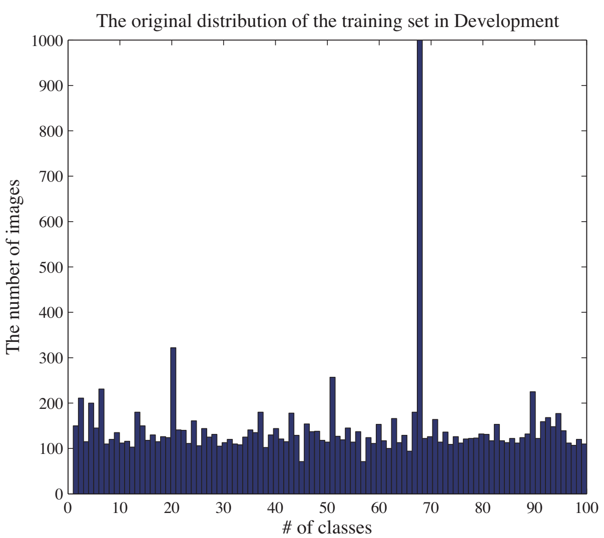
(c)



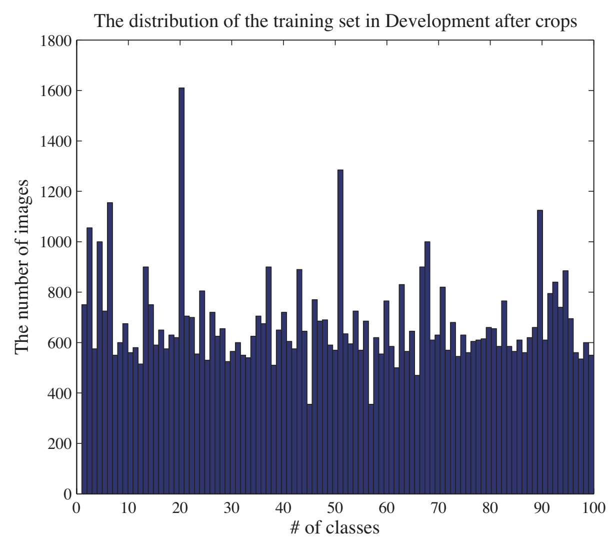
Late fusion

Implementation details

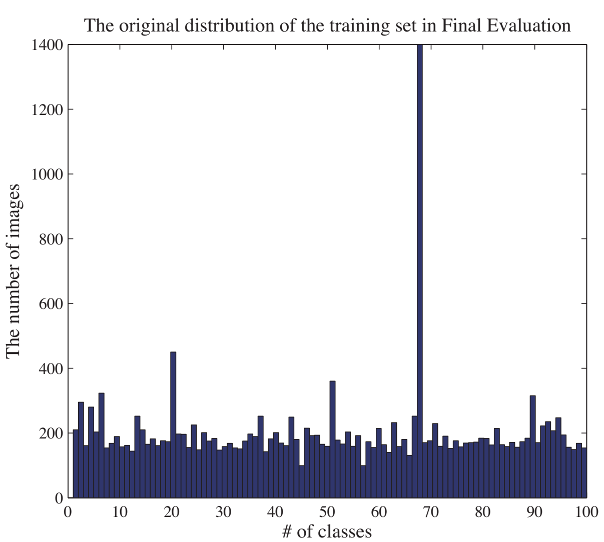
Distributions of the number of training images in Dev. and Final Evaluation:



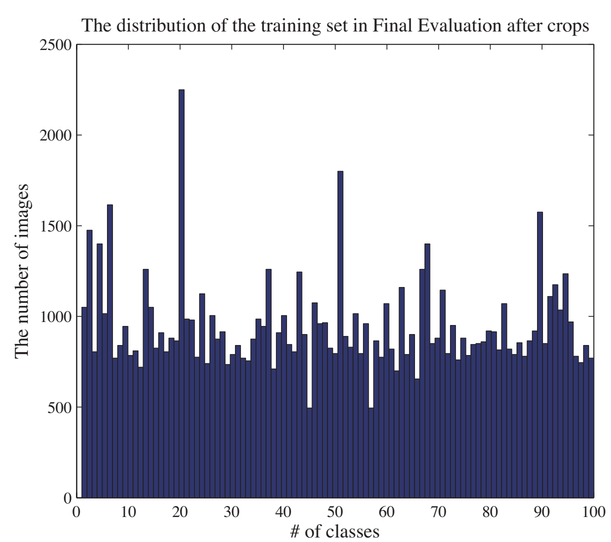
(a)



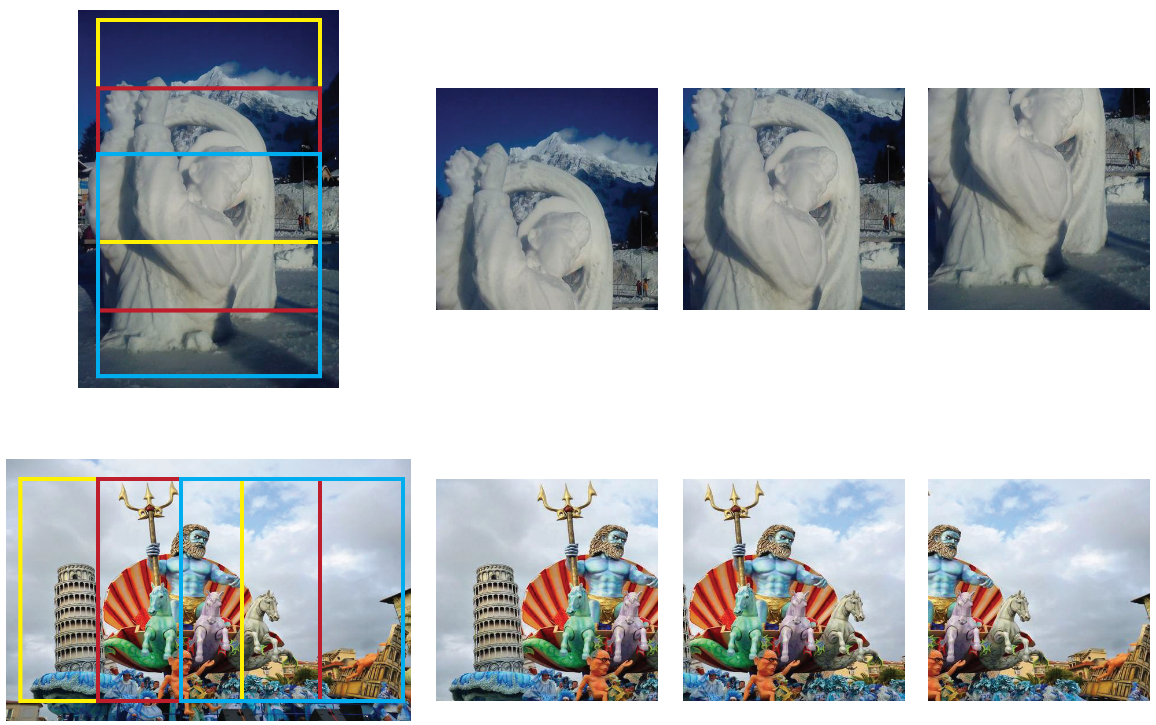
(b)



(c)



(d)


Original images　Crop1　Crop2　Crop3

Late fusion

Experimental results

Recognition mAP comparisons of the Development phase. Note that, “FT” stands for the fine-tuned deep networks; “SS” is for single scale, and “MS” is for multiple scales.

	VGG Net-D	VGG Net-E	FT VGG Net-D	FT VGG Net-E	Place-CNN
SS	0.761	0.762	–	–	–
MS	0.770	0.773	0.779	0.769	0.640
Late fusion	0.782	0.784	0.802	0.791	0.649
Ensemble	0.841				

Experimental results

LAMDA
Learning And Mining from DatA

Recognition mAP comparisons of the Development phase. Note that, "FT" stands for the fine-tuned deep networks; "SS" is for single scale, and "MS" is for multiple scales.

	VGG Net-D	VGG Net-E	FT VGG Net-D	FT VGG Net-E	Place-CNN
SS	0.761	0.762	–	–	–
MS	0.770	0.773	0.779	0.769	0.640
Late fusion	0.782	0.784	0.802	0.791	0.649
Ensemble	0.841				

Rank	Team	Score
1	VIPL-ICT-CAS	0.854
2	**FV (Ours)**	0.851
3	MMLAB	0.847
4	NU&C	0.824
5	CVL_ETHZ	0.798

Thank you!