

Ask the locals: multi-way local pooling for image recognition

Supplemental material.

Y-Lan Boureau^{1,3,*}

Nicolas Le Roux^{1,†}

Francis Bach^{1,†}

Jean Ponce^{2,*}

Yann LeCun³

¹INRIA

²Ecole Normale Supérieure

³Courant Institute, New York University

Abstract

This supplemental material contains numerical results of experiments plotted in the main paper, and results from the Caltech-256 and Scenes datasets that were omitted due to space constraints.

	Caltech 30 tr.	Scenes
$K = 256, \text{Pre}, P = 1$	70.5 ± 0.8	79.2 ± 0.7
$P = 4$	72.6 ± 1.0	81.7 ± 0.5
$P = 16$	74.0 ± 1.0	82.0 ± 0.7
$P = 64$	75.0 ± 0.8	81.4 ± 0.4
$P = 128$	75.5 ± 0.8	81.0 ± 0.3
$P = 256$	75.1 ± 1.0	–
$P = 512$	74.5 ± 0.7	–
$P = 1024$	73.8 ± 0.8	–
$P = 1 + 16$	74.2 ± 1.1	81.5 ± 0.8
$P = 1 + 64$	75.6 ± 0.6	81.9 ± 0.7
$K = 256, \text{Post}, P = 4$	72.4 ± 1.2	79.6 ± 0.8
$P = 16$	75.1 ± 0.8	80.9 ± 0.6
$P = 64$	76.4 ± 0.8	81.1 ± 0.6
$P = 128$	76.7 ± 0.8	81.1 ± 0.5
$P = 256$	75.9 ± 0.8	–
$P = 512$	75.2 ± 0.8	–
$P = 1024$	74.2 ± 0.6	–
$K = 1024, \text{Pre}, P = 1$	75.6 ± 0.9	82.7 ± 0.7
$P = 4$	76.0 ± 1.2	83.5 ± 0.8
$P = 16$	76.3 ± 1.1	82.8 ± 0.8
$P = 64$	76.2 ± 0.8	81.8 ± 0.7
$P = 128$	–	80.9 ± 0.7
$P = 1 + 16$	76.9 ± 1.0	83.3 ± 1.0
$P = 1 + 64$	77.3 ± 0.6	83.1 ± 0.7
$K = 1024, \text{Post}, P = 4$	75.8 ± 1.5	82.9 ± 0.6
$P = 16$	77.0 ± 0.8	82.9 ± 0.5
$P = 64$	77.1 ± 0.7	82.4 ± 0.7
$P = 128$	76.9 ± 0.5	82.0 ± 0.7
$P = 256$	75.7 ± 0.8	–

Table 1. Accuracy as a function of whether clustering is performed before (Pre) or after (Post) the encoding, K : dictionary size, and P : number of configuration space bins.

*WILLOW project-team, Laboratoire d’Informatique de l’Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548.

†SIERRA project-team, Laboratoire d’Informatique de l’Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548.

p	1	4	16
$k = 256$	32.9 ± 0.4	34.5 ± 0.5	38.0 ± 0.6
$k = 1024$	39.7 ± 0.3	39.2 ± 0.5	40.8 ± 0.6
$k = 256$	32.3 ± 0.8	34.9 ± 0.5	38.0 ± 0.5
$k = 1024$	38.1 ± 0.6	39.8 ± 0.7	41.6 ± 0.6

Table 2. Recognition accuracy on Caltech 256, 30 training examples, as a function of K : size of the codebook for sparse coding, and P : number of clusters extracted on the input data. Macrofeatures extracted every 4 pixels. Top two rows: no resizing of the image. Bottom two rows: resized so that maximal dimension is ≤ 300 pixels.

CI01	$k = 4$	$k = 16$	$k = 64$	$k = 256$	$k = 1024$
Single	38.5 ± 1.0	58.3 ± 1.1	66.7 ± 1.1	72.6 ± 1.0	76.0 ± 1.2
Sep	45.9 ± 1.1	60.0 ± 1.0	68.2 ± 1.3	73.4 ± 0.9	76.7 ± 1.0
Scenes	$k = 4$	$k = 16$	$k = 64$	$k = 256$	$k = 1024$
Single	56.6 ± 0.7	69.3 ± 0.7	76.6 ± 1.0	81.7 ± 0.5	83.5 ± 0.8
Sep	61.5 ± 0.5	70.6 ± 0.8	78.2 ± 0.7	81.8 ± 0.7	83.7 ± 0.8

Table 3. Recognition accuracy on Caltech-101 and Scenes, according to whether a separate dictionary is learned for each of $P = 4$ clusters. Single: shared dictionary. Sep: one dictionary per cluster. Dictionary size K .

Caltech-101	$p = 1$	$p = 4$	$p = 16$	$p = 64$
$k = 4$	28.0 ± 0.7	38.5 ± 1.0	53.9 ± 1.0	62.8 ± 0.8
$k = 16$	53.1 ± 1.0	58.3 ± 1.1	63.2 ± 1.0	68.6 ± 1.0
$k = 64$	62.8 ± 1.0	66.7 ± 1.1	69.7 ± 0.8	73.0 ± 0.6
Scenes	$p = 1$	$p = 4$	$p = 16$	$p = 64$
$k = 4$	40.6 ± 0.7	56.6 ± 0.7	64.5 ± 0.6	70.1 ± 0.8
$k = 16$	63.3 ± 0.8	69.3 ± 0.7	74.0 ± 0.6	76.0 ± 0.6
$k = 64$	72.6 ± 0.6	76.6 ± 1.0	78.9 ± 0.5	79.2 ± 0.5
Caltech-256	$p = 1$	$p = 4$	$p = 16$	$p = 64$
$k = 4$	6.9 ± 0.3	11.7 ± 0.4	17.4 ± 0.3	21.0 ± 0.3
$k = 16$	16.5 ± 0.3	20.9 ± 0.4	24.8 ± 0.3	26.5 ± 0.3
$k = 64$	23.0 ± 0.3	26.5 ± 0.3	29.8 ± 0.3	—

Table 4. Recognition accuracy for smaller dictionaries, as a function of K : size of the codebook for sparse coding, and P : number of clusters for pooling. Preclustering needs larger final global image representations to outperform richer dictionaries. Macrofeatures used for Caltech-101 and Scenes with image resizing, standard features with full-size images for Caltech-256.

w_i	$k = 4$	$k = 16$	$k = 64$	$k = 256$
Caltech-101, max pooling				
1	53.9 ± 1.0	63.2 ± 1.0	69.7 ± 0.8	74.0 ± 1.0
$\sqrt{n_i/n}$	53.9 ± 1.2	61.6 ± 0.6	66.7 ± 0.8	70.9 ± 1.0
n_i/n	47.0 ± 1.1	54.9 ± 1.1	60.9 ± 0.8	65.5 ± 0.9
Scenes, max pooling				
1	64.5 ± 0.6	74.0 ± 0.6	78.9 ± 0.5	81.5 ± 0.8
$\sqrt{n_i/n}$	66.1 ± 1.0	72.7 ± 0.8	77.2 ± 0.6	79.5 ± 0.9
n_i/n	63.4 ± 0.9	69.2 ± 0.7	74.2 ± 0.8	76.9 ± 0.8
Caltech-256, max pooling				
1	17.4 ± 0.3	24.8 ± 0.3	29.8 ± 0.3	–
$\sqrt{n_i/n}$	19.3 ± 0.4	24.5 ± 0.3	28.2 ± 0.3	–
n_i/n	16.4 ± 0.3	20.7 ± 0.2	24.0 ± 0.3	–
Caltech-101, average pooling				
1	44.9 ± 0.8	54.8 ± 0.8	60.7 ± 0.9	64.1 ± 0.9
$\sqrt{n_i/n}$	49.7 ± 0.9	56.6 ± 0.8	62.3 ± 1.1	65.7 ± 1.2
n_i/n	44.6 ± 0.8	51.5 ± 0.9	58.1 ± 1.2	62.8 ± 1.3
Scenes, average pooling				
1	56.4 ± 0.6	66.3 ± 1.0	72.0 ± 0.6	74.5 ± 0.6
$\sqrt{n_i/n}$	62.6 ± 0.8	68.5 ± 0.8	72.4 ± 0.4	74.6 ± 0.8
n_i/n	60.9 ± 1.1	65.7 ± 0.8	70.2 ± 0.4	72.5 ± 0.9
Caltech-256, average pooling				
1	12.6 ± 0.2	17.9 ± 0.3	19.9 ± 0.4	–
$\sqrt{n_i/n}$	16.8 ± 0.4	20.4 ± 0.4	21.9 ± 0.3	–
n_i/n	14.8 ± 0.3	17.8 ± 0.4	20.2 ± 0.3	–

Table 5. Recognition accuracy on Caltech-101, Scenes, and Caltech-256, for different combinations of pooling operator (max or average pooling) and cluster weighting schemes, for $P = 16$ clusters, and dictionary size K . Macrofeatures on resized images are used for Caltech-101 and Scenes, standard features on full-size images for Caltech-256.