LEARNING MULTI-ATTENTION CONVOLUTIONAL NEURAL NETWORK FOR FINE-GRAINED IMAGE RECOGNITION

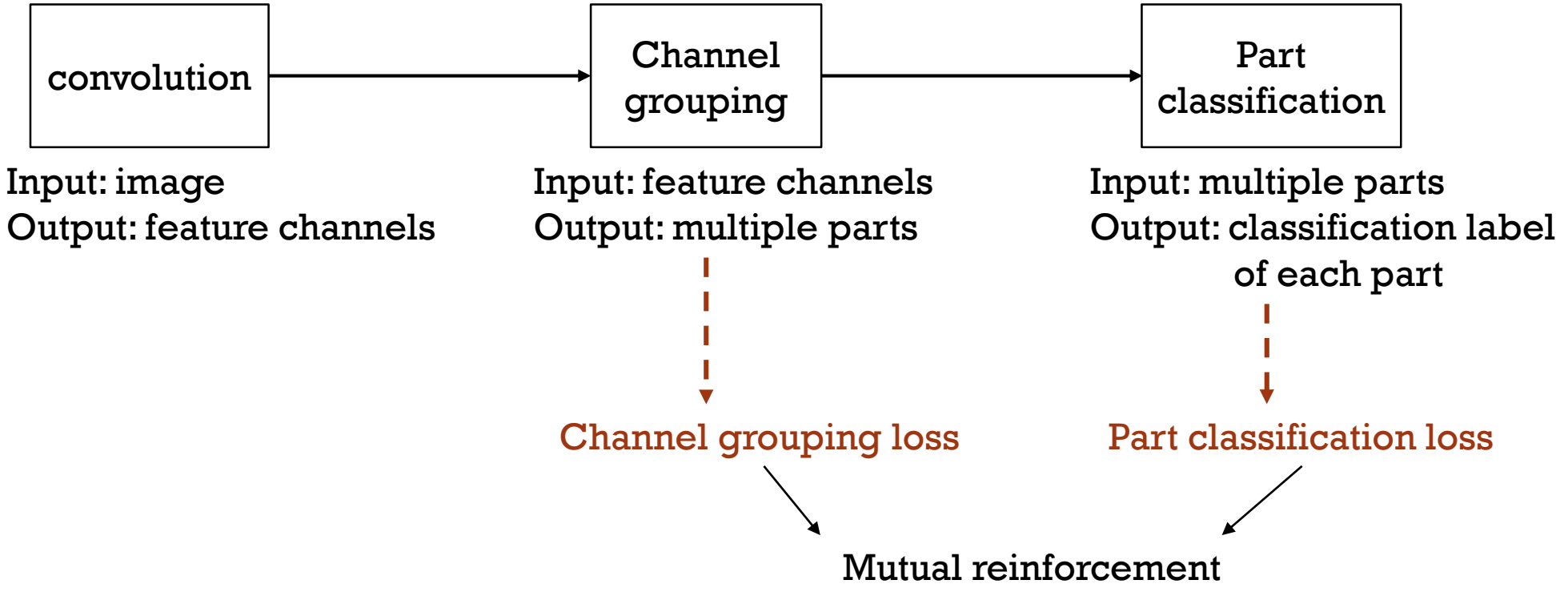
HELIANG ZHENG, JIANLONG FU, TAO MEI, JIEBO LUO

(ICCV, 2017)

7107053120 許庭瑄

ABSTRACT

- Objection: Fine-grained recognition
- Proposed Method: Multi-Attention Convolutional Neural Network（MA-CNN）

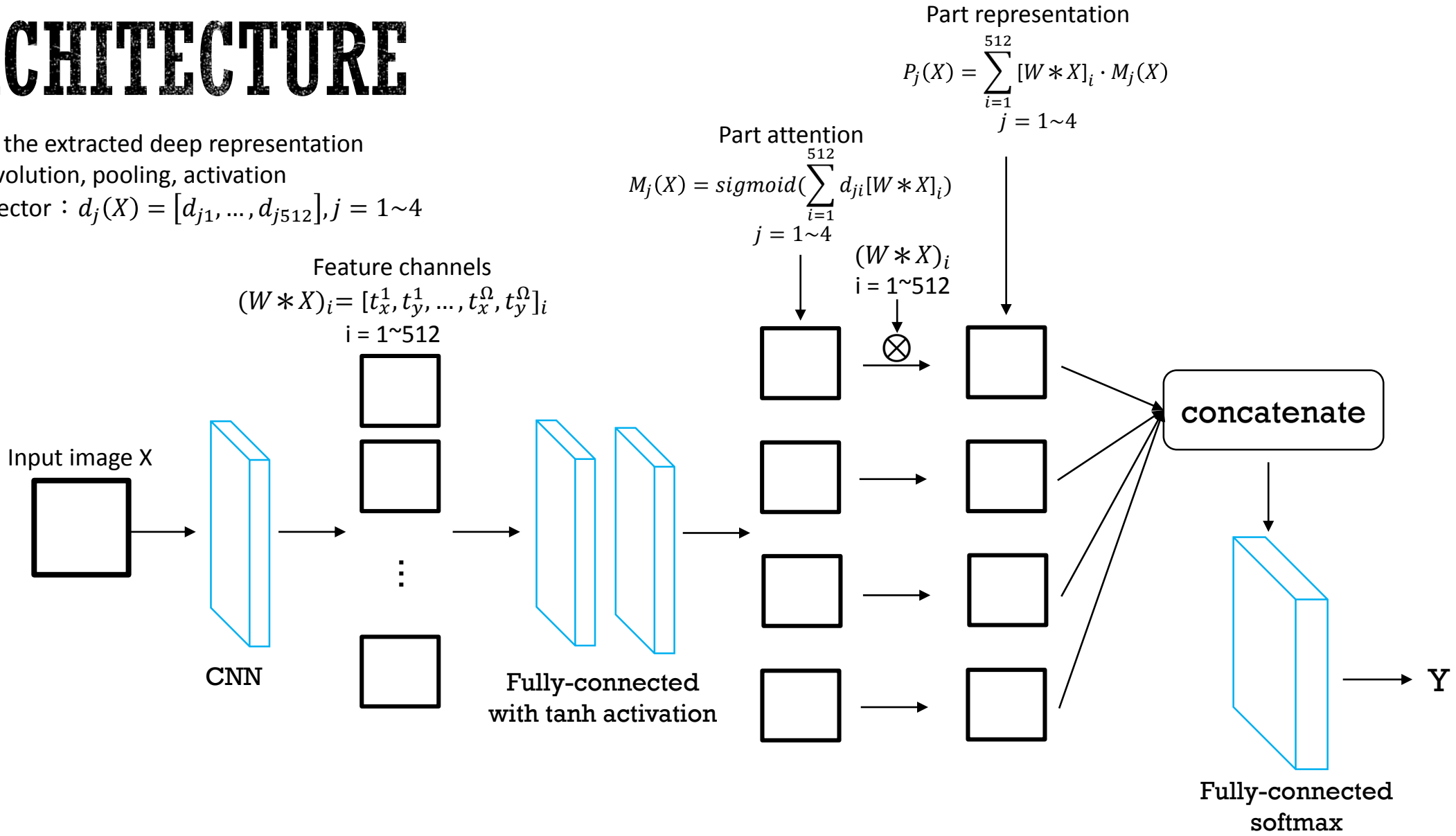


ARCHITECTURE

$W * X$ ：the extracted deep representation

$*$ ：convolution, pooling, activation

Weight vector：$d_j(X) = [d_{j1}, \dots, d_{j512}], j = 1\sim4$



OBJECTIVE FUNCTION

$Y^{(i)}$: predict label vector
Y^* : ground truth label vector
L_{cng} : channel grouping loss
$m_i(x, y)$: the value of $M_i(X)$ at (x, y)
mrg : a margin

- $L(X) = \sum_{i=1}^{4}\left[L_{cls}\left(Y^{(i)}, Y^*\right)\right] + L_{cng}(M_i(X))$
- L_{cls} : classification loss on each of 4 parts
- $L_{cng}(M_i(X)) = Dis(M_i(X)) + \lambda Div(M_i(X))$
- $Dis(M_i(X)) = \sum_{(x,y)\in M_i(X)} m_i(x, y)\left[\left||x - t_x|\right|^2 + ||y - t_y||^2\right]$
- $Div(M_i(X)) = \sum_{(x,y)\in M_i(X)} m_i(x, y)[\max_{k\neq i} m_k(x, y) - mrg]$

RESULT

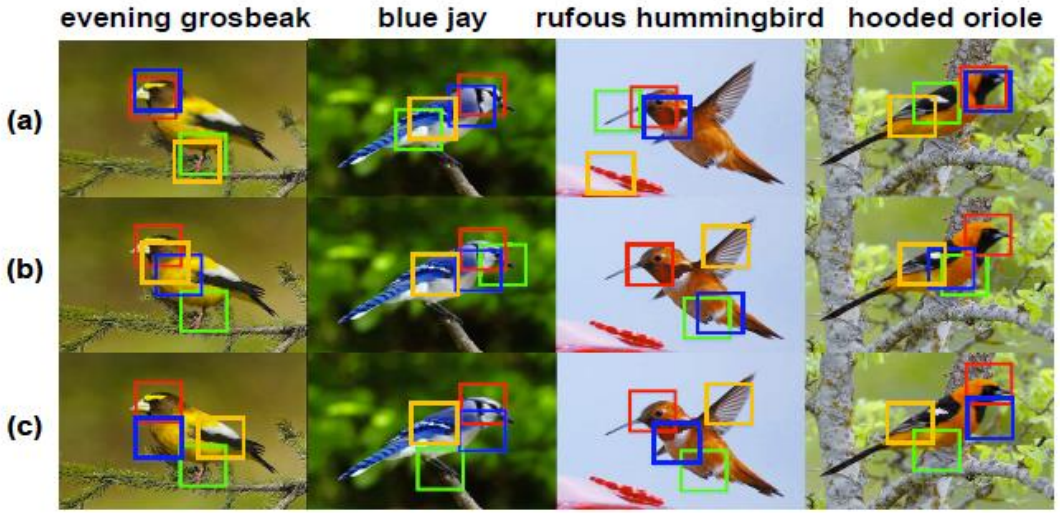


Figure 4: Four bird examples of the visualized part localization results by (a) initial parts by channel clustering, (b) optimizing channel grouping loss L_{cng}, and (c) joint learning $L_{cng} + L_{cls}$.

Table 3: Comparison results on CUB-200-2011 dataset. Train Anno. represents using bounding box or part annotation in training.

Approach	Train Anno.	Accuracy
PN-CNN(AlexNet) [1]	✓	75.7
Part-RCNN(AlexNet) [34]	✓	76.4
PA-CNN [14]	✓	82.8
MG-CNN [27]	✓	83.0
FCAN [18]	✓	84.3
B-CNN (250k-dims) [17]	✓	85.1
Mask-CNN [29]	✓	85.4
TLAN(AlexNet) [31]		77.9
MG-CNN [27]		81.7
FCAN [18]		82.0
B-CNN (250k-dims) [17]		84.1
ST-CNN (Inception net) [10]		84.1
PDFR [35]		84.5
RA-CNN [5]		85.3
MA-CNN (2 parts + object)		85.4
MA-CNN (4 parts + object)		**86.5**

REFERENCE

[1] Very deep convolutional networks for large-scale image recognition.

[2] Caltech-UCSD Birds 200.

[3] Bird species categorization using pose normalized deep convolutional nets.