Semantic-Guided Multi-Attention Localization for Zero-Shot Learning

Yizhe Zhu[1], Jianwen Xie[2], Zhiqiang Tang[1], Xi Peng[3], Ahmed Elgammal[1]
[1] Rutgers University, [2] Hikivision Research Institute, [3] University of Delaware

Motivation

Existing zero-shot learning approaches predominantly focus on learning the proper mapping function for visual-semantic embedding, while neglecting the effect of learning discriminative visual features.
We observe that multiple discriminative part areas are key points to recognize objects, especially fine-grained objects. For instance, the head and tail are crucial to distinguish bird species.

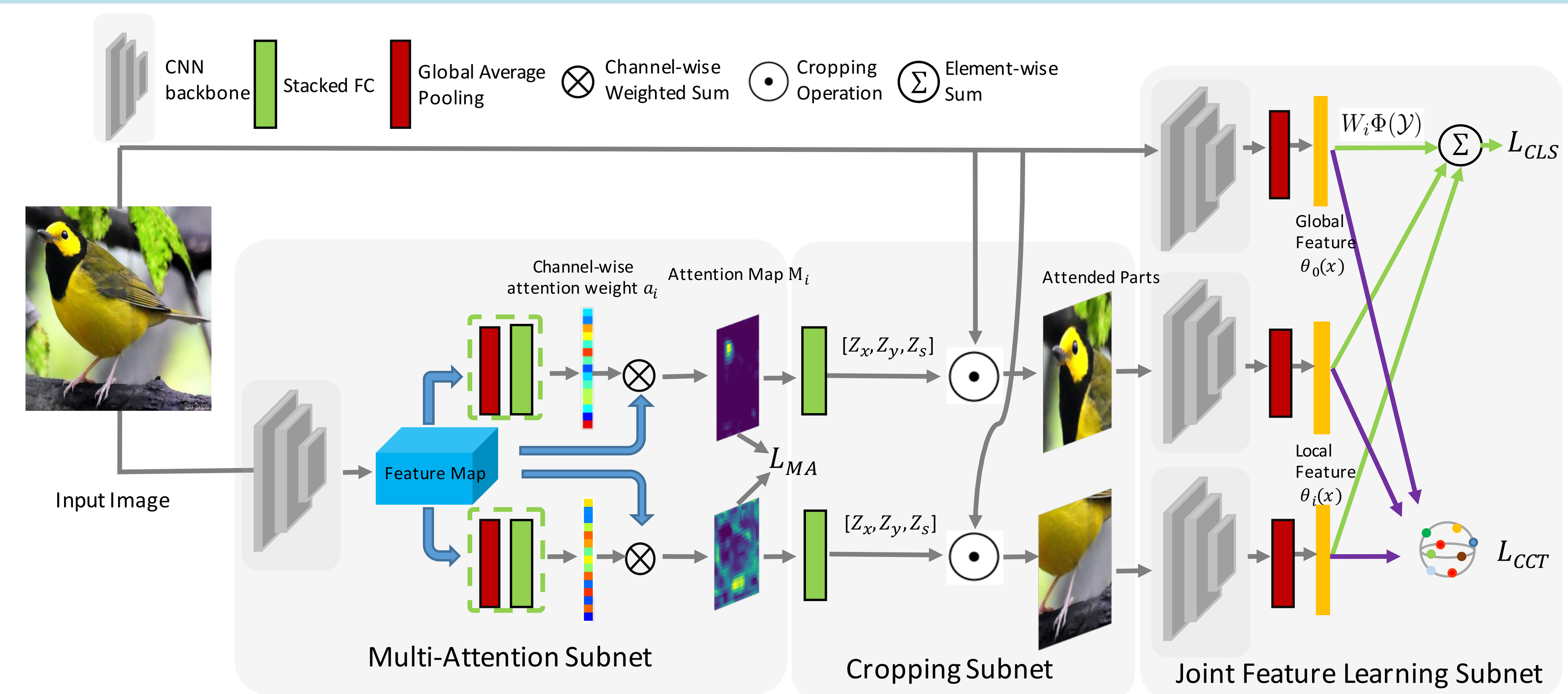
Contribution

- We present a weakly-supervised multi-attention localization model for zero-shot recognition, which jointly discovers the crucial regions and learns feature representation under the guidance of semantic descriptions.
- We propose a multi-attention loss to encourage compact and diverse attention distribution by applying geometric constraints over attention maps.
- We jointly learn global and local features under the supervision of embedding softmax loss and class-center triplet loss to provide an enhanced visual representation for ZSL.
- We conduct extensive experiments and analysis on three zero-shot learning datasets and demonstrate the excellent performance of our proposed method on both part detection and zero-shot learning.

Zero-shot learning Problem

Assume there are N labeled instances from C^s seen classes $\mathcal{D}^s = \{(x_i^s, y_i^s, z_i^s)\}_{i=1}^N$ as training data, where $x_i^s \in \mathcal{X}$ denotes the image, $y_i^s \in \mathcal{Y}^s$ is the corresponding class label , $z_i^s = \varphi(y_i^s) \in \mathcal{S}$ is the semantic representation of the corresponding class. Given an image x_i^u from an unseen class and a set of semantic representations of unseen classes $\{z_i^u = \varphi(y_i^u)\}_{i=1}^{C^u}$, where C^u denotes the number of unseen classes, the task of zero-shot learning is to predict the class label $y^u \in \mathcal{Y}^u$ of the image, where $\mathcal{Y}^s$ and $\mathcal{Y}^u$ are disjoint.

Method



Our model takes as input the original image and produces 2 part attention maps. The part images from the cropping subnet and the original images are fed into different CNNs in the joint feature learning subnet for semantic description-guided object recognition.

Objective function:

$$\mathcal{L}_{SGMA} = \mathcal{L}_{MA} + \alpha_1 \mathcal{L}_{CLS} + \alpha_2 \mathcal{L}_{CCT}$$

Multi-Attention Loss

To discover compact and diverse regions over attention maps, we design $\mathcal{L}_{MA}$:

$$\mathcal{L}_{MA} = \sum_i^{N_a} [\mathcal{L}_{CPT}(M_i) + \lambda \mathcal{L}_{DIV}(M_i)], \quad (1)$$

where N_a is the of attention maps; M_i is i^{th} attention map.
For intra-map compactness: we expect each attention map to concentrate on a small range:

$$\mathcal{L}_{CPT}(M_i) = ||M_i - \widetilde{M}_i||_2^2 \quad (2)$$

where $\widetilde{M}_i$ is an ideal concentrated attention map, created as a Gaussian blob centering on the peak activation of M_i.
For inter-map diversity: we also expect the attention maps to attend different discriminative parts.

$$\mathcal{L}_{DIV}(M_i) = \sum_{z \in \mathcal{Z}} m_i^z \max\{0, \widehat{m}^z - mrg\}, \quad (3)$$

where $\widehat{m}^z = \max_{k \neq i} m_k^z$ represents the maximum of other attention maps at location z and mrg denotes a margin.

Zero-Shot Leaning Loss

We employ two cooperative losses: the embedding softmax loss to encourage **a higher inter-class distinction**, and the class-center triplet loss to force the learned feature of each class to be concentrated with **a lower intra-class divergence**.
Embedding Softmax Loss
Let $\theta(x)$ and $\varphi(y)$ denote the visual and semantic features respectively. The compatibility score $s_j = \theta(x)^T W \varphi(y_j)$, $y_j \in \mathcal{Y}_s$, where W is a trainable transform matrix.

$$\mathcal{L}_{CLS} = -\frac{1}{N} \log \frac{\exp(s_j)}{\sum_{\mathcal{Y}_s} \exp(s_j)}, \quad (4)$$

where N is the number of training samples.
Class-Center Triplet Loss

$$\mathcal{L}_{CCT} = \max\{0, mrg + ||\widehat{\varphi}_i - \widehat{C}_i||_2^2 - ||\widehat{\varphi}_i - \widehat{C}_k||_2^2\}_{i \neq k}, \quad (5)$$

where i, k be the class indices, mrg is the margin, φ_i is the mapped visual feature in semantic feature space (i.e., $\varphi_i = \theta(x)^T W_i$), C_i denotes the "center" of each class that are trainable parameters, $\widehat{\ }$ means L_2 normalization operation.

Zero-Shot Recognition Results

Zero-shot learning results on CUB, AWA, FLO datasets. The best scores and second best ones are marked bold and underline respectively.

Method	CUB		AWA		FLO
	SS	PS	SS	PS	
LATEM (2016)	49.4	49.3	74.8	55.1	40.4
ALE (2015)	53.2	54.9	78.6	59.9	48.5
SJE (2015)	55.3	53.9	76.7	65.6	53.4
ESZSL (2015)	55.1	53.9	74.7	58.2	51.0
SYNC (2016)	54.1	55.6	72.2	54.0	-
SAE (2017)	33.4	33.3	80.6	53.0	45.6
DEM (2017)	51.8	51.7	80.3	<u>65.7</u>	41.6
GAZSL (2018)	57.5	55.8	77.1	63.7	60.5
SCoRe (2017)	59.5	62.7	82.8	61.6	<u>60.9</u>
LDF (2018)	<u>67.1</u>	<u>67.5</u>	<u>83.4</u>	65.5	-
Ours	**70.5**	**71.0**	**83.5**	**68.8**	**65.9**

Generalized Zero-shot learning results on CUB, AWA.

Method	CUB			AwA		
	$\mathcal{A}_\mathcal{U}$	$\mathcal{A}_\mathcal{S}$	$\mathcal{H}$	$\mathcal{A}_\mathcal{U}$	$\mathcal{A}_\mathcal{S}$	$\mathcal{H}$
DEM	19.6	57.9	29.2	32.8	84.7	47.3
GAZSL	31.7	61.3	41.8	29.6	84.2	43.8
LDF	26.4	**81.6**	39.9	9.8	**87.4**	17.6
Ours	**36.7**	71.3	**48.5**	**37.6**	87.1	**52.5**

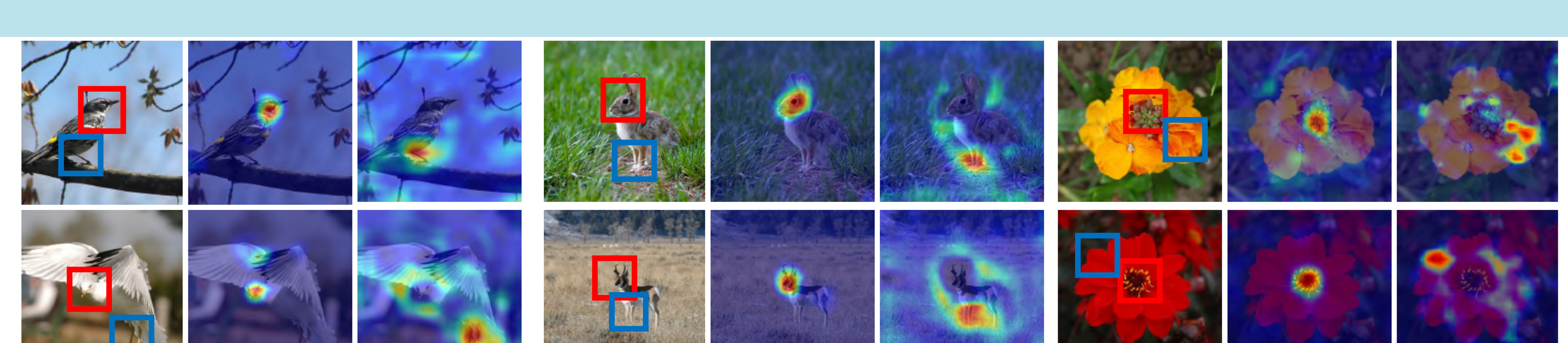
Ablation Study: the performance of variants of our model on zero-shot learning with PS setting.

Method	CUB	AWA	FLO	Avg
Baseline	60.2	61.5	57.7	59.8
Parts	55.4	51.2	49.8	52.1
Baseline+Parts	**67.4**	**64.3**	**63.9**	**65.2**
Baseline+Random Parts	56.3	59.8	56.4	57.5
Embedding Softmax	60.9	62.4	57.2	60.2
Class-Center Triplet	62.1	64.6	61.1	62.6
Combined	**63.5**	**65.7**	**61.8**	**63.7**

Part Detection Results

Method	Head	Tail	Average
SPDA-CNN	90.9	67.2	79.1
Ours	74.9	48.1	61.5
Ours w/o MA	65.7	29.4	47.6
Random	25.6	26.0	25.8

Quantitative results measured by average precision(%).



Qualitative results. Each result consists of three images, where the detected parts are marked with blue and red bounding boxes in the first image, and the rest two images are the corresponding generated attention maps.

Reference

- Yan Li, Junge Zhang, Jianguo Zhang, Kaiqi Huang. Discriminative Learning of Latent Features for Zero-Shot Recognition, *CVPR* 2018.
- Feng Wang, Xiang Xiang, Jian Cheng, Alan L. Yuille. NormFace: L2 Hypersphere Embedding for Face Verification, *ACM-MM* 2017.
- Heliang Zheng, Jianlong Fu, Tao Mei, Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition, *ICCV* 2017.