Title: CHANNEL-WISE MIX-FUSION DEEP NEURAL NETWORKS FOR ZERO-SHOT LEARNING

Paper ID:4852

Authors: Guowei Wang[1,2] , Naiyang Guan[1]*, Hanjia Ye[3] , Xiaodong Yi[1] , Hang Cheng[1] , Junjie Zhu[1]

Affiliates: 1Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing
2College of Intelligence and Computing, Tianjin University, Tianjin
3State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing
*Corresponding Author: Naiyang Guan (nyguan@sina.com)

Motivation

The key in ZSL lies in the learning of ***visual and semantic cross-domain mappings***. We consider introducing more attributes-related visual information for the model to enhance this mapping and constructing the relationship between the objects and their background information.

Problem Formulation

In ZSL, seen classes $S \equiv \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and unseen classes $\mathrm{U} \equiv \{(x_j^u, y_j^u)\}_{j=1}^{N_u}$ are strictly disjoint. We have S and auxiliary attributes for training, and our goal is to recognize unseen class U correctly.

Our Model

Our model takes the original scale image as input and generates a delicate scale image through the cropping module. The backbone CNN extract features from both scale images, and two cooperative attention-based modules are applied on two CNN, respectively. We then project the features to the attribute space as well as the latent space. All parameters are jointly optimized.

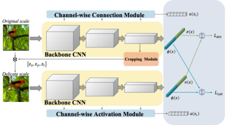


Fig.1 $a(x_i)$ denotes the integrated attribute attention. $\sigma(x)$ and $\phi(x)$ denote the latent features and semantic prediction features, respectively.

Our Results

Table I. The Mean Class Accuracy (%) of CMFZ.

Methods	CUB		AwA2	
	SS	PS	SS	PS
ALE [7]	53.2	54.9	65.3	59.9
SJE [19]	55.3	53.9	62.0	65.6
SYNC [20]	54.1	53.6	72.7	54.0
LDF [4]	67.1	67.5	83.3	65.5
LFGAA [3]	67.7	67.7	84.3	68.1
SGMA [5]	**70.5**	71.0	83.5	**68.8**
CMFZ†	67.7	71.4	84.4	64.7
CMFZ‡	68.6	**72.7**	**84.7**	65.3
CMFZ	**70.0**	**73.7**	**85.9**	**68.4**

The results showed that our model achieved the best performance on CUB PS and AwA2 SS.

The ablation analysis showed that both CCM and CAM could boost the performance.

Learning and Inference

Learning:

For visual-semantic projection, we use softmax loss, i.e.,

$$L_{att} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_j)}{\sum_{Y_s}\exp(s_j)},$$

$$s_j = \theta(x)^T W \varphi(y_j), y_j \in Y_s.$$

For visual-latent projection, we use triplet loss, i.e.,

$$L_{lat} = \frac{1}{N}\sum_{i=1}^{N} \max\left\{0, \|\sigma(x_i) - \sigma(x_j)\|^2 - \|\sigma(x_i) - \sigma(x_k)\|^2 + mrg\right\}.$$

The cropping loss, i.e.,

$$L_{Mse} = ([t_x, t_y, t_l] - [z_x, z_y, t_l])^2.$$

The overall loss function, i.e.,

$$L = \sum_n L_{att}^n + \alpha L_{lat}^n + \beta L_{Mse}^{n-1}.$$

Inference:

For visual-semantic projection, we have, i.e.,

$$y_{att}{}_j^c = arg \max_{c \in Y^u} \left(s\left(\phi(x_j^u), \varphi(c)\right) \right),$$

For visual-latent projection, we choose the predicted labels as follows, i.e.,

$$\sigma_s = \frac{1}{N}\sum_{i=1}^{N} \sigma(x_i),$$

$$\beta_s^u = arg \min_{s \in Y^s} \left\| \varphi(u) - \sum \beta_s^u \varphi(s) \right\|_2^2 + \lambda \|\beta_s^u\|_2^2,$$

$$\sigma_u = \sum_{s \in Y^s} \beta_s^u \sigma_s,$$

$$y_{lat}{}_j^c = arg \max_{c \in Y^u} \left(s\left(\sigma(x_j^u), \sigma_u\right) \right).$$

$$y_j^c = arg \max_{c \in Y^u} \left(y_{att}{}_j^c, y_{lat}{}_j^c \right).$$

Title: CHANNEL-WISE MIX-FUSION DEEP NEURAL NETWORKS FOR ZERO-SHOT LEARNING

Paper ID:4852

Authors: *Guowei Wang*[1,2] , *Naiyang Guan*[1]∗ , *Hanjia Ye*[3] , *Xiaodong Yi*[1] , *Hang Cheng*[1] , *Junjie Zhu*[1]

Affiliates: [1]Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing

[2]College of Intelligence and Computing, Tianjin University, Tianjin

[3]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing

*Corresponding Author: Naiyang Guan (nyguan@sina.com)

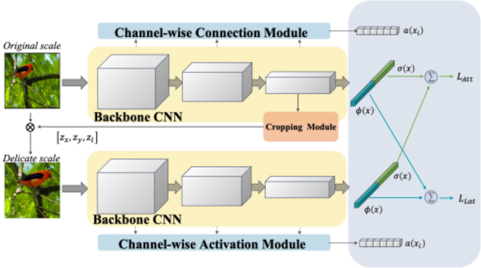
Problem Formulation

- In ZSL, seen classes $S \equiv \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and unseen classes $\mathrm{U} \equiv \{\left(x_j^u, y_j^u\right)\}_{j=1}^{N_u}$ are strictly disjoint. We have S and auxiliary attributes for training, and our goal is to recognize unseen class U correctly.

Motivation

- The key in ZSL lies in the learning of ***visual and semantic cross-domain mappings***. We consider introducing more attributes-related visual information for the model to enhance this mapping and constructing the relationship between the objects and their background information.

Our Model



- Our model takes the original scale image as input and generates a delicate scale image through the cropping module. The backbone CNN extract features from both scale images, and two cooperative attention-based modules are applied on two CNN, respectively. We then project the features to the attribute space as well as the latent space. All parameters are jointly optimized.

Learning and Inference

Learning:

For visual-semantic projection, we use softmax loss, i.e.,

$$L_{att} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_j)}{\sum_{Y_s} \exp(s_j)},$$

$$s_j = \theta(x)^T W \varphi(y_j), y_j \in Y_s.$$

For visual-latent projection, we use triplet loss, i.e.,

$$L_{lat} = \frac{1}{N}\sum_{i=1}^{N} \max\left\{0, \left\|\sigma(x_i) - \sigma(x_j)\right\|^2 - \|\sigma(x_i) - \sigma(x_k)\|^2 + mrg\right\}.$$

The cropping loss, i.e.,

$$L_{Mse} = ([t_x, t_y, t_l] - [z_x, z_y, t_l])^2.$$

The overall loss function, i.e.,

$$L = \sum_{n} L_{att}^{n} + \alpha L_{lat}^{n} + \beta L_{Mse}^{n-1}.$$

Inference:

For visual-semantic projection, we have, i.e.,

$$yatt_j^c = arg \max_{c \in Y^u} \left(s\left(\phi\left(x_j^u\right), \varphi(c)\right)\right),$$

For visual-latent projection, we choose the predicted labels as follows, i.e.,

$$\sigma_s = \frac{1}{N}\sum_{i=1}^{N} \sigma(x_i),$$

$$\beta_s^u = arg \min_{s \in Y^s} \left\| \varphi(u) - \sum \beta_s^u \varphi(s) \right\|_2^2 + \lambda \|\beta_s^u\|_2^2,$$

$$\sigma_u = \sum_{s \in Y^s} \beta_s^u \sigma_s,$$

$$ylat_j^c = arg \max_{c \in Y^u} \left(s\left(\sigma\left(x_j^u\right), \sigma_u\right)\right).$$

$$y_j^c = arg \max_{c \in Y^u} \left(yatt_j^c, ylat_j^c \right).$$

Results

Methods	CUB		AwA2	
	SS	PS	SS	PS
ALE [7]	53.2	54.9	65.3	59.9
SJE [19]	55.3	53.9	62.0	65.6
SYNC [20]	54.1	53.6	72.7	54.0
LDF [4]	67.1	67.5	83.3	65.5
LFGAA [3]	67.7	67.7	84.3	68.1
SGMA [5]	**70.5**	71.0	83.5	**68.8**
CMFZ†	67.7	71.4	84.4	64.7
CMFZ‡	68.6	**72.7**	**84.7**	65.3
CMFZ	**70.0**	**73.7**	**85.9**	**68.4**

The results showed that our model achieved the best performance on CUB PS and AwA2 SS.

The ablation analysis showed that both CCM and CAM could boost the performance.

Thank you for listening!
If you have any questions, please feel free to contact our corresponding author (nyguan@sina.com).