西安电子科技大学
XIDIAN UNIVERSITY

Co-Representation Network for Generalized Zero-Shot Learning

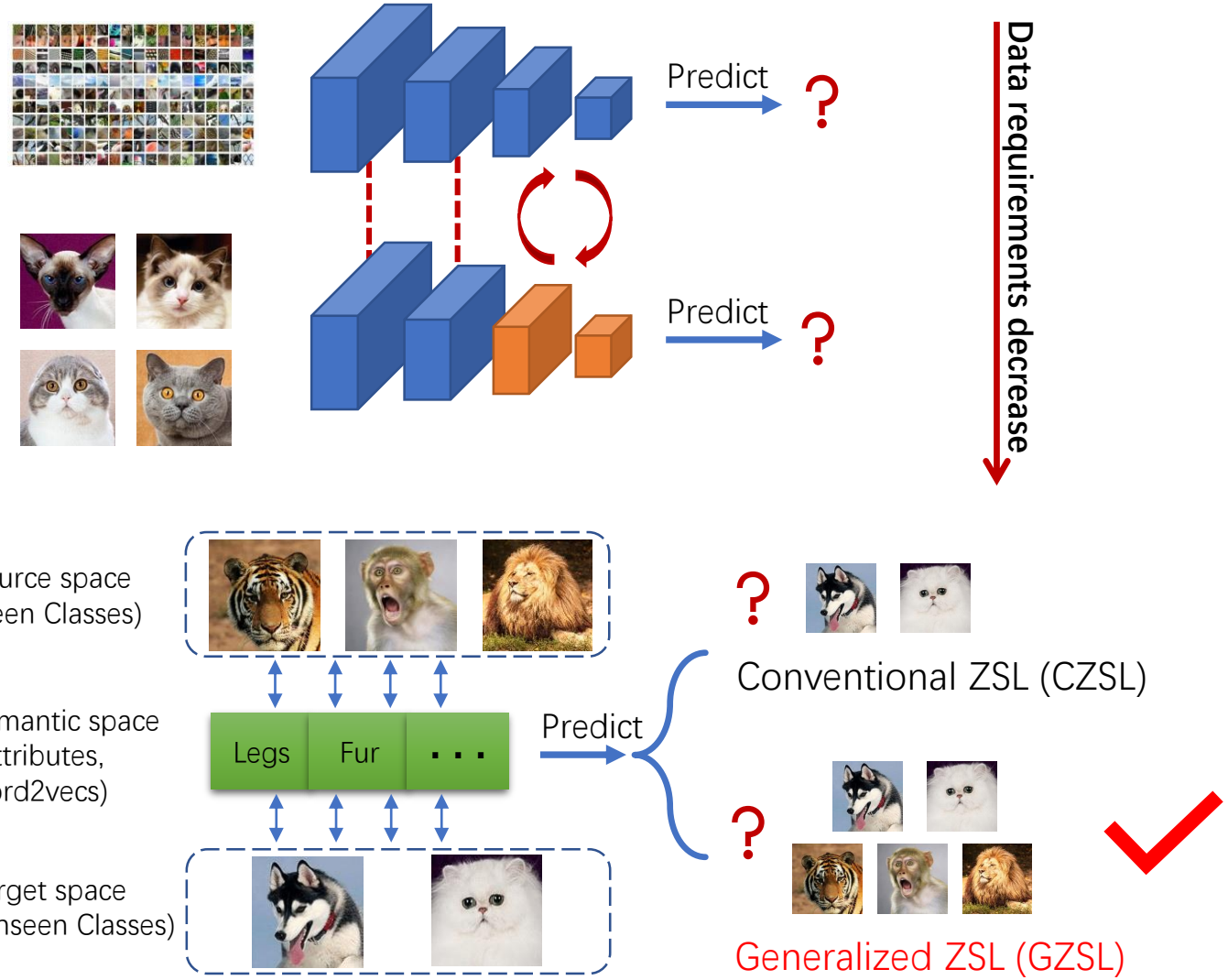
Fei Zhang, Guangming Shi

XIDIAN UNIVERSITY

ICML 2019

Introduction

- Classic Deep CNN
- Transfer Learning
  - Few-Shot Learning
  - One-Shot Learning
  - Zero-Shot Learning (ZSL)



Bias Problem

Existing Embedding Models for GZSL

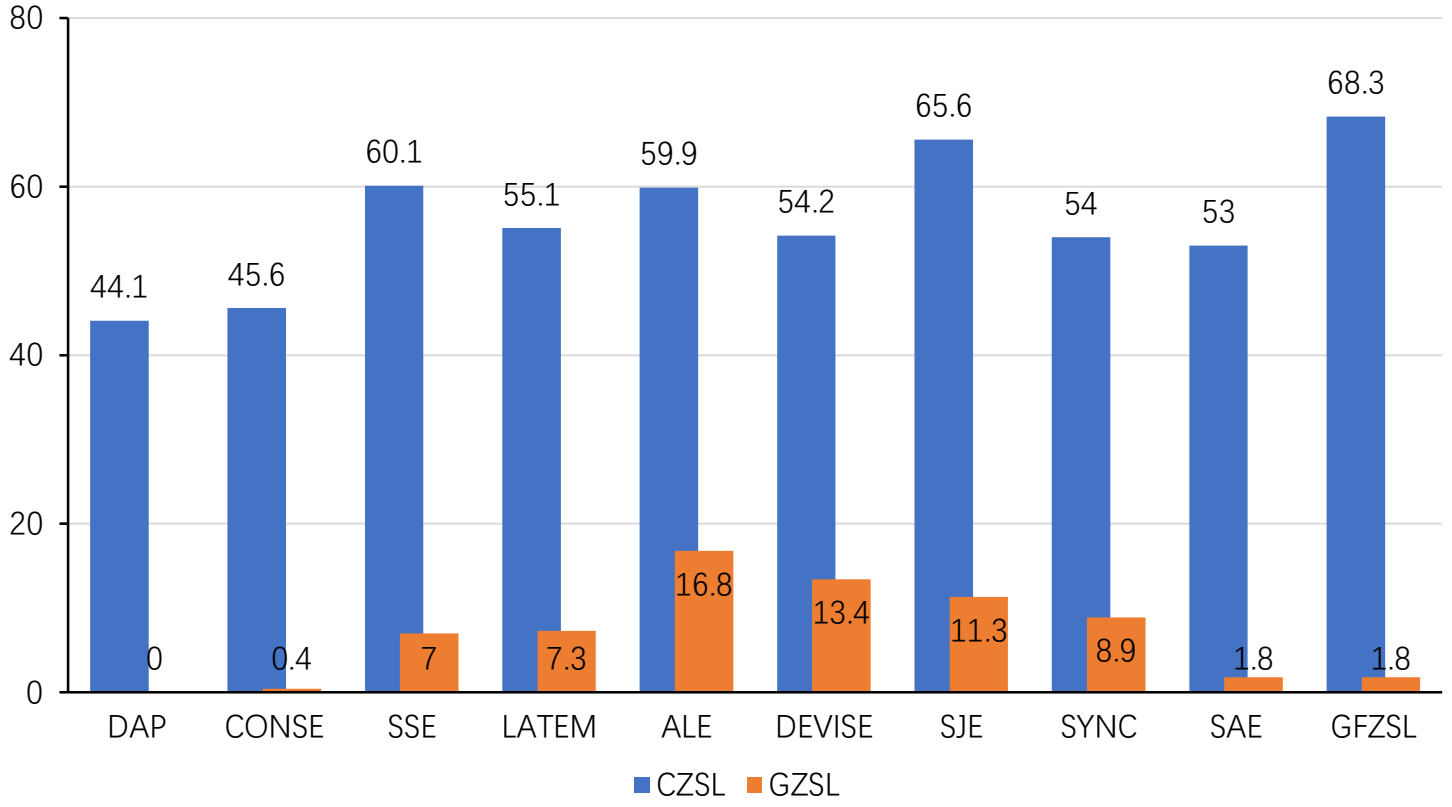
- Visual Space
  to Semantic Space
- Visual & Semantic Space
  to a Latent Space
- Semantic Space
  to Visual Space

Bias Problem

Unseen samples are easily classified into similar seen classes.

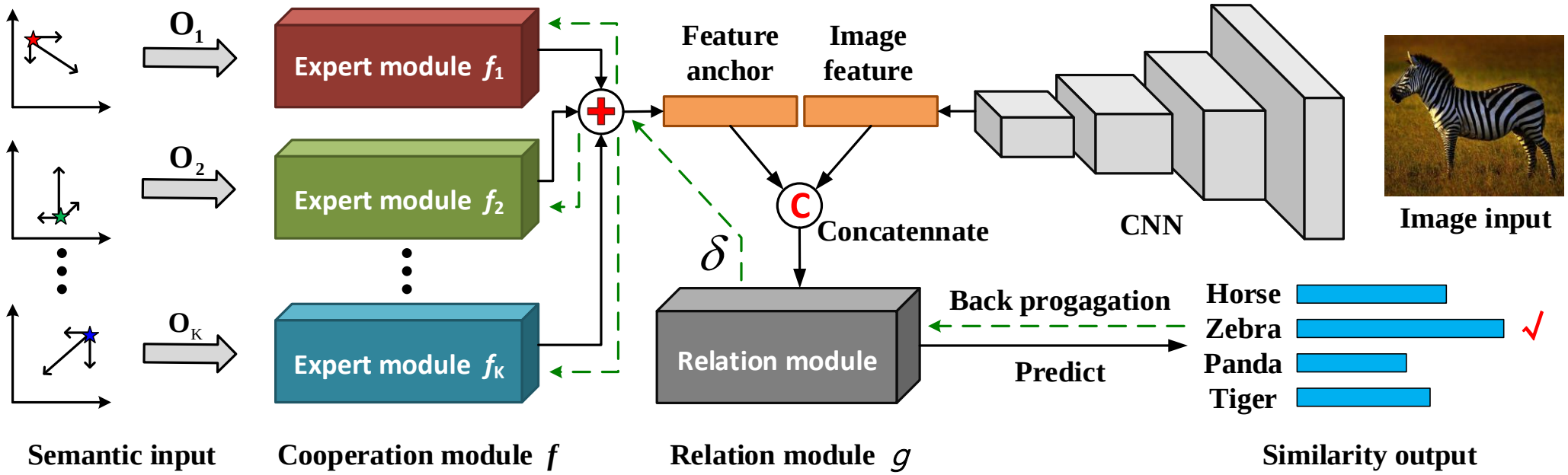
e.g. Zebra → Horse

Average per-class top-1 accuracy in % on unseen classes of various models following CZSL settings and GZSL settings



Yongqin, Xian , et al. "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly." *IEEE TPAMI* 2017

Our Model



- **Co-Representation Network (CRnet)**

1. A cooperation module for visual feature representation (our main contribution).
2. A pre-trained CNN (Resnet-101) for feature extraction.
3. A relation module for similarity output, i.e. the classification.
   (Sung, Flood , et al. "Learning to Compare: Relation Network for Few-Shot Learning." CVPR 2018.)

Algorithm

➢ **Initialization Algorithm**

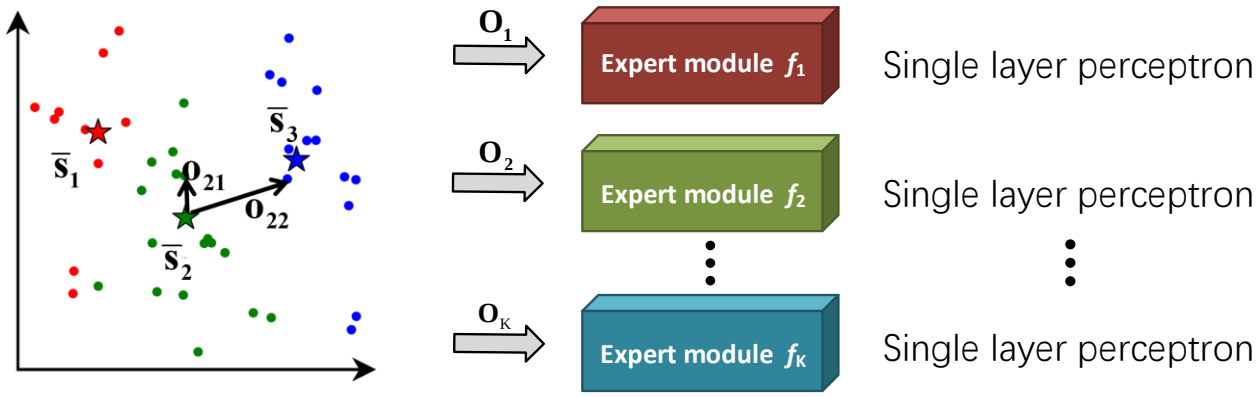
Perform ***K-means Clustering*** on the semantic space.

Semantic vectors: $\mathbf{s}_m^s$

Clustering center: $\bar{\mathbf{s}}_k$

Expert module k:

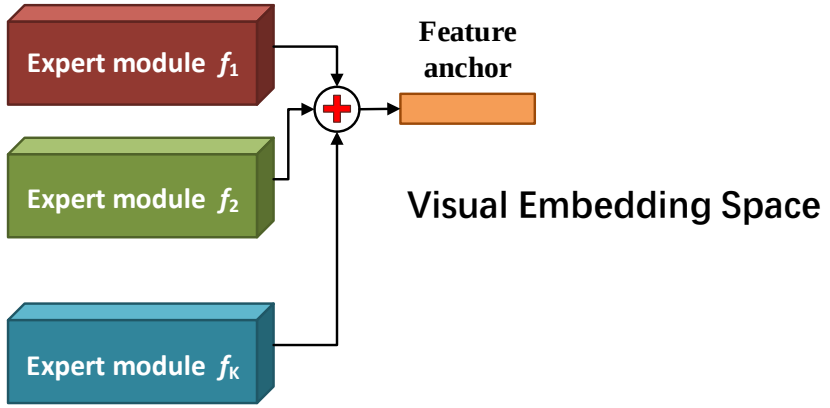
$$f_k(\mathbf{s}_m^s; \bar{\mathbf{s}}_k) = relu\left(\mathbf{W}_k(\mathbf{s}_m^s - \bar{\mathbf{s}}_k) + \mathbf{b}_k\right)$$



➢ **Cooperation Module**

Sum the outputs of expert modules.

$$f(\mathbf{s}_m^s) = \sum_{k=1}^{K} relu\left(\mathbf{W}_k(\mathbf{s}_m^s - \bar{\mathbf{s}}_k) + \mathbf{b}_k\right)$$



Algorithm

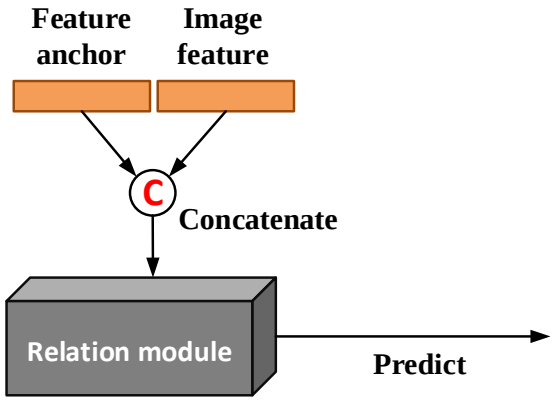
➢ **Relation Module**

Concatenate feature anchor $\hat{\mathbf{v}}_m^s$ (output of cooperation module) and image feature $\mathbf{v}_i^s$ as the input.

Tow-layer perceptron with Sigmoid.

Ground-truth:

$$l(\mathbf{v}_i^s, \hat{\mathbf{v}}_m^s) = \begin{cases} 1, & y_m^s = y_i \\ 0, & y_m^s \neq y_i \end{cases}$$



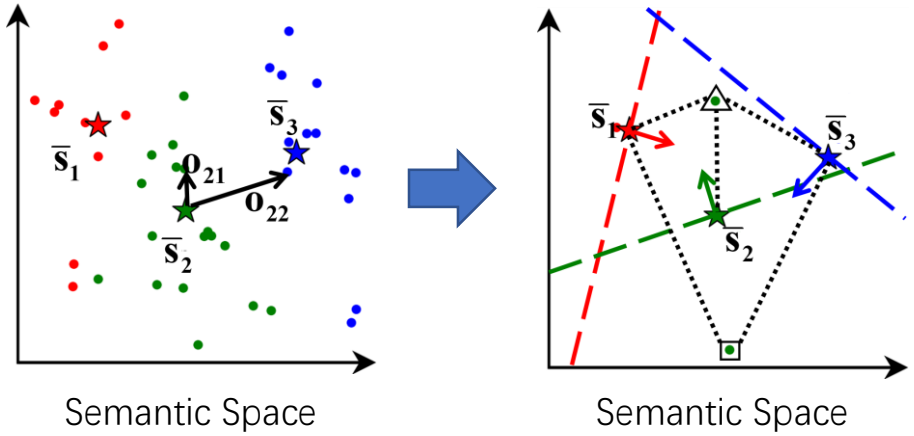
➢ **Training**

Objective function:

$$\arg\min_{\mathbf{w}_f, \mathbf{w}_g} \sum_m \sum_i (g(\mathbf{v}_i^s, \hat{\mathbf{v}}_m^s) - l(\mathbf{v}_i^s, \hat{\mathbf{v}}_m^s))^2 + \alpha ||\mathbf{w}_f||_2^2 + \beta ||\mathbf{w}_g||_2^2$$

End-to-end manner.

- When the model converges, cooperation module divides the semantic space into several parts.
- Semantic vectors located in different parts are projected by several different expert modules.



Benchmark Results

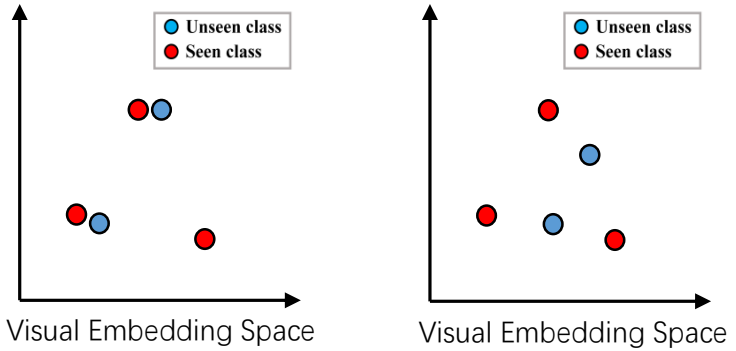
Table 1. Classification accuracies on various datasets following the GZSL setting. As/Au: Average per-class top-1 accuracy in % on seen/unseen classes. H: Harmonic mean accuracy. Cloumn 3-12 are the results of classic CZSL methods on GZSL repruduced by Xian et al. (2017), which show strong bias on As. Cloumn 13-18 are the official results of recent GZSL methods.

	AwA1			AwA2			CUB			SUN			aPY		
Method	**As**	**Au**	**H**	**As**	**Au**	**H**	**As**	**Au**	**H**	**As**	**Au**	**H**	**As**	**Au**	**H**
CONSE (Norouzi et al., 2013)	88.6	0.4	0.8	90.6	0.5	1.0	72.2	1.6	3.1	39.9	6.8	11.6	91.2	0.0	0.0
CMT* (Socher et al., 2013)	86.9	8.4	15.3	89.0	8.7	15.9	60.1	4.7	8.7	28.0	8.7	13.3	74.2	10.9	19.0
SSE (Zhang & Saligrama, 2015)	80.5	7.0	12.9	82.5	8.1	14.8	46.9	8.5	14.4	36.4	2.1	4.0	78.9	0.2	0.4
SJE (Akata et al., 2015)	74.6	11.3	19.6	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8	55.7	3.7	6.9
ESZSL (Romera-Paredes & Torr, 2015)	75.6	6.6	12.1	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8	70.1	2.4	4.6
SYNC (Changpinyo et al., 2016)	87.3	8.9	16.2	90.5	10.0	18.0	70.9	11.5	19.8	43.3	7.9	13.4	66.3	7.4	13.3
SAE (Kodirov et al., 2017)	77.1	1.8	3.5	82.2	1.1	2.2	54.0	7.8	13.6	18.0	8.8	11.8	80.9	0.4	0.9
LATEM (Xian et al., 2016)	71.7	7.3	13.3	77.3	11.5	20.0	57.3	15.2	24.0	28.8	14.7	19.5	73.0	0.1	0.2
ALE (Akata et al., 2016)	16.8	76.1	27.5	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3	73.7	4.6	8.7
DEVISE (Frome et al., 2013)	68.7	13.4	22.4	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9	76.9	4.9	9.2
DEM (Zhang et al., 2017)	84.7	32.8	47.3	86.4	30.5	45.1	57.9	19.6	29.2	34.3	20.5	25.6	11.1	75.1	19.4
RN (Sung et al., 2018)	91.3	31.4	46.7	93.4	30.0	45.3	61.1	38.1	47.0	-	-	-	-	-	-
DCN (Liu et al., 2018)	84.2	25.5	39.1	-	-	-	60.7	28.4	38.7	37.0	25.5	30.2	75.0	14.2	23.9
CVAE-ZSL (Mishra et al., 2018)	-	-	47.2	-	-	51.2	-	-	34.5	-	-	26.7	-	-	-
SE-GZSL (Verma et al., 2018)	67.8	56.3	61.5	68.1	58.3	62.8	53.3	41.5	46.7	30.5	40.9	34.9	-	-	-
f-CLSWGAN (Xian et al., 2018)	61.4	57.9	59.6	-	-	-	57.7	43.7	49.7	36.6	42.6	39.4	-	-	-
CRnet (Ours)	74.7	**58.1**	**65.4**	78.8	52.6	**63.1**	56.8	**45.5**	**50.5**	36.5	34.1	35.3	68.4	**32.4**	**44.0**

Analysis

➢ Bias Problem

Unseen anchors distribute too close to seen anchors in the embedding space used for classification.



Serious bias problem **Slight bias problem**

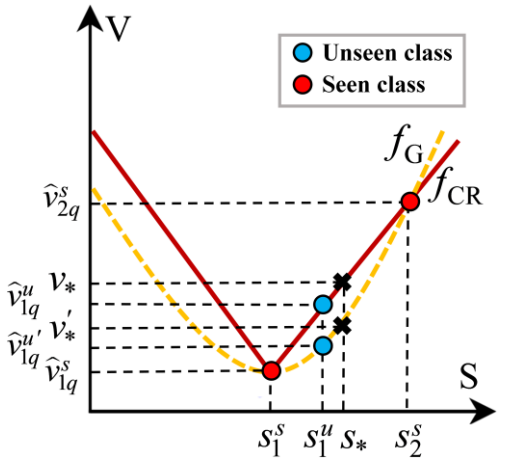
➢ Local Relative Distance (LRD)

We propose the LRD as a metric for bias problem.

$$e(\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2) = \left(\sum_{q=1}^{Q} \frac{(\hat{v}_{1q} - \hat{v}_{2q})^2}{\sigma_q^2} \right)^{\frac{1}{2}} \qquad \mathrm{LRD}(\hat{\mathbf{v}}_j^u) = \frac{e(\hat{\mathbf{v}}_j^u, \hat{\mathbf{v}}_{c(j)}^s)}{e(\hat{\mathbf{v}}_{c(j)}^s, \hat{\mathbf{v}}_{c(j)}^{s'})}$$

Larger LRD means a more uniform embedding space, i.e. slighter bias problem.

1-d semantic space to 1-d visual embedding space:



$$p\left(\mathrm{LRD}(\hat{v}_{1q}^u) > \mathrm{LRD}(\hat{v}_{1q}^{u'})\right) > \frac{s_* - s_1^s}{s_2^s - s_1^s} = 0.5$$

- **High local linearity results in larger LRD.**
- **Cooperation module actually learns a piecewise linear function of K+1 pieces with high local linearity**

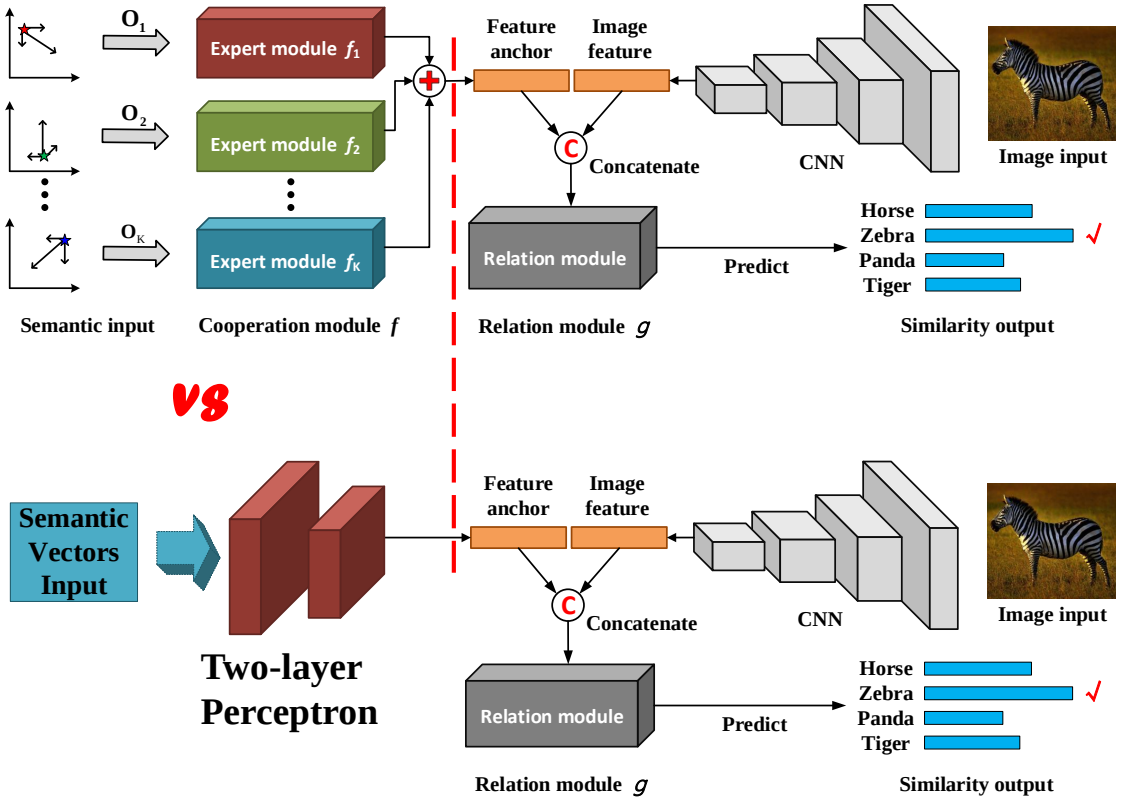
f_G: General fitting curve; f_{CR}: Fitting curve of CRnet
S: semantic space; V: visual embedding space.

Contrast Experiments

- **Relation Network (RN)**

  A two-layer perceptron instead of cooperation module is used.

  (Sung, Flood , et al. "Learning to Compare: Relation Network for Few-Shot Learning." CVPR 2018.)



CRnet

RN

Contrast Experiments

➢ **Results**

Compared with RN, CRnet achieves:

- **More Sparse and Discriminative Features**

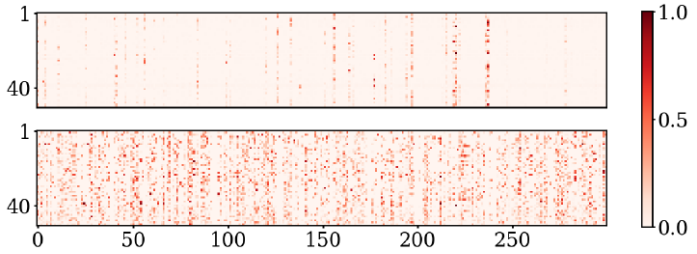


Figure 4. Visualization of the feature anchors of CRnet (above) and RN (below) well-trained on AwA2. The first 300 dimensions' normalization results of the feature anchors of 40 unseen classes and 10 seen classes are presented.

- **More Uniform Embedding Space (Larger LRD)**

Table 3. Average LRD of all unseen class anchors for RN and CRnet on various datasets.

	AwA1	AwA2	CUB
RN	0.711	0.756	0.831
CRnet	0.835	0.956	0.843

- **Slighter Bias Problem**

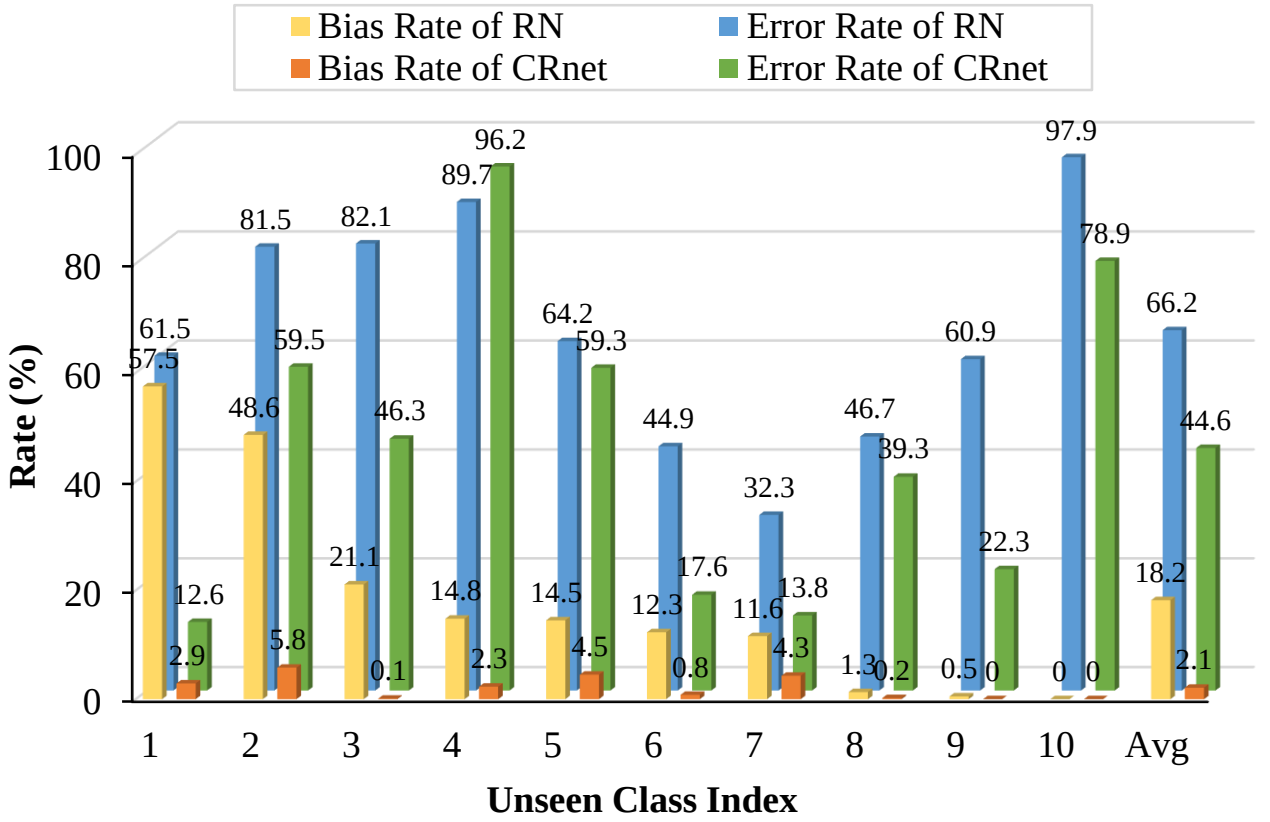


Figure. Bar chart of per-class Bias Rate and per-class Error Rate of RN and CRnet on AwA2. Bias Rate: The rate in % of misclassification into the closest seen class; Error Rate: Per-class classification Error Rate in %.

Summarize

➢ **Co-representation network**

- Decomposition method for projecting semantic space to visual embedding space.
- Cooperation module for representation and learnable relation module for classification.

✓ Training in an end-to-end manner.

✓ Slighter bias problem leads to a good performance on GZSL.

Other advantages:

✓ Simple structure with high expandability.

✓ No need for semantic information of unseen classes during training (compared with generative models)

Email:
f.zhang@stu.xidian.edu.cn