Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders

Presented by Bodong Zhang

Schonfeld, Edgar, et al. "Generalized zero-and few-shot learning via aligned variational autoencoders." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019

Abstract

- Proposed a model where a **shared latent space** of image features and class embeddings is learned by **modality-specific aligned variational autoencoders** to achieve **zero-shot and few-shot learning**.
- The key to the approach is that we **align the distributions learned from images and from side-information to construct latent features** that contain the essential multi-modal information associated with unseen classes.
- Results on ImageNet with various zero-shot splits show that our latent features generalize well in large-scale settings.

Introduction

- In Generalized zero-shot learning (GZSL), as visual data of unseen classes is not available at training time, typically **knowledge transfer from seen to unseen classes** is achieved via some form of side information that encode semantic relationship between classes, i.e. class embeddings.
- In this work, we **train VAEs to encode and decode features from different modalities**, e.g. images and class attributes, and use the learned latent features to train a generalized zero-shot learning classifier.
- **Main contributions**:
- (1) Propose the **CADA-VAE model** that learns shared cross-modal latent representations of multiple data modalities using VAEs via distribution alignment and cross alignment objectives.
- (2) Evaluate our model using conventional benchmark datasets, i.e. CUB, SUN, AWA1 and AWA2, on zero-shot and few-shot learning settings.
- (3) The latent features learned by our model improve the state of the art in the truly large-scale ImageNet dataset.

Structure of model

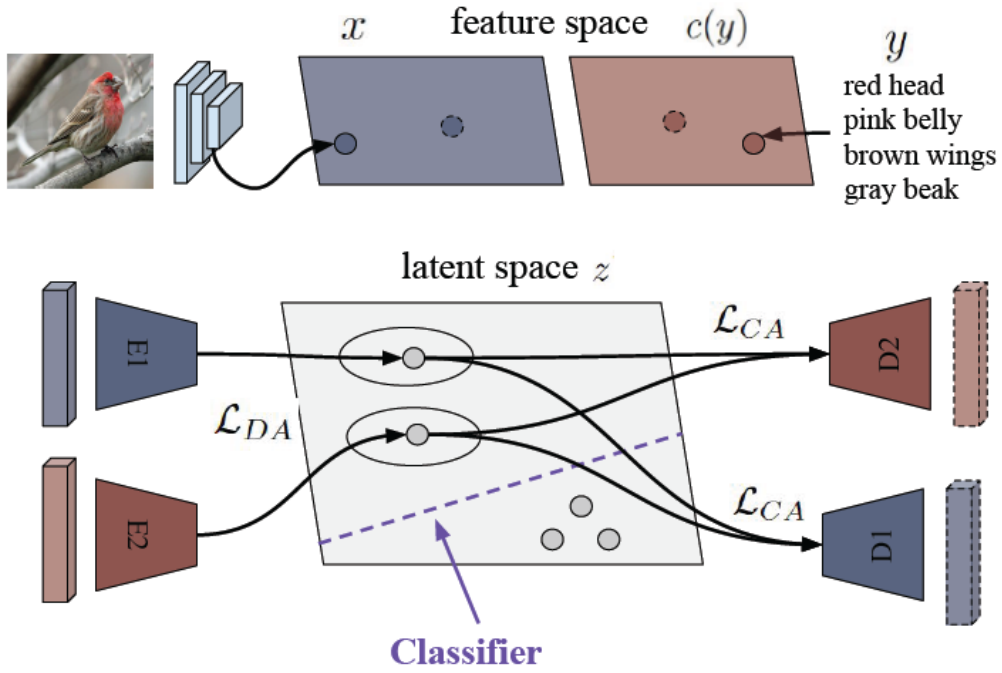


Figure 1: Our CADA−VAE model learns a latent embedding (z) of image features (x) and class embedding ($c(y)$ of labels y) via aligned VAEs optimized with cross-alignment ($\mathcal{L}_{CA}$) and distribution alignment ($\mathcal{L}_{DA}$) objectives, and subsequently trains a classifier on sampled latent features of seen and unseen classes.

- There are two types of input in VAE: image features and class embedding. Image features x could be generated by ResNet. Each class has different attributes c(y). They will be encoded to same latent space z.

Generalized Zero-and Few-Shot Learning

- (1) In zero-shot learning, training and testing classes are disjoint with shared attributes annotated on class level, and the performance of the method is solely judged on its classification accuracy on the novel classes.
- (2) Generalized zero-shot learning model is judged on the harmonic mean of the classification accuracy on seen and unseen classes.
- (3) In few-shot learning, there are k examples provided at training time for the previously unseen classes.

CADA-VAE Model

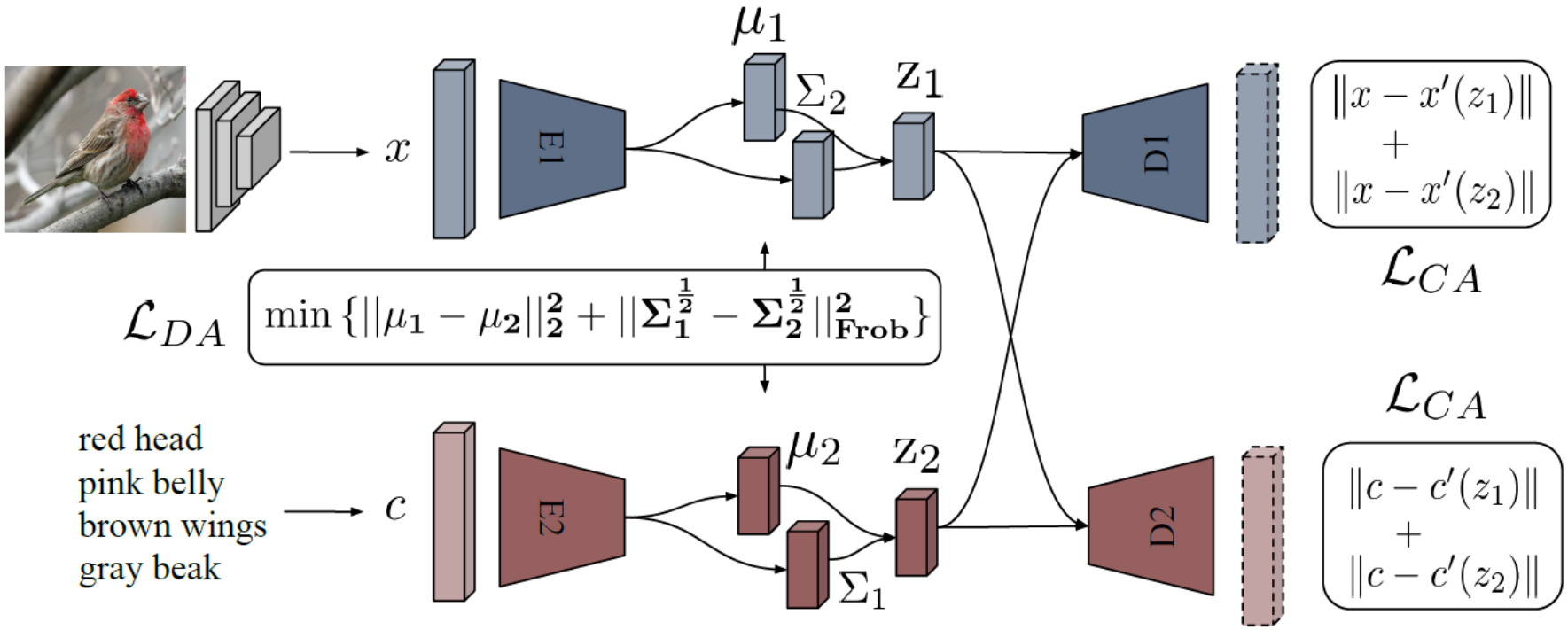


Figure 2: Our Cross- and Distribution Aligned VAE (CADA-VAE). Latent distribution alignment is achieved by minimizing the Wasserstein distance between the latent distributions ($\mathcal{L}_{DA}$). Similarly, the cross-alignment loss ($\mathcal{L}_{CA}$) encourages the latent distributions to align through cross-modal reconstruction.

- The key of the approach is the choice of a VAE latent-space, a reconstruction and cross-reconstruction criterion to preserve class-discriminative information in lower dimensions, as well as explicit distribution alignment to encourage domain-agnostic representations.

Background

Generalized Zero-shot Learning:

- Training set: S={(x, y, c(y))|x ∈ X, y ∈ Y^S, c(y) ∈ C}
- x is image-features, e.g. extracted by CNN. y is class labels. c(y) is class embeddings.
- Auxiliary training set: U={(u, c(u))|u ∈ Y^U, c(y) ∈ C}
- u denotes unseen classes from a set Y^U, which is disjoint from Y^S.
- f_{ZSL}: X → Y^U f_{GZSL}: X → $Y^U \cup Y^S$

Variational Autoencoder (VAE)

- The encoder predicts μ and ∑ such that $q_\emptyset$(z|x) = N(μ, ∑), from which a latent vector z is generated via the reparametrization trick. The encoder tries to reconstruct original x from generated z.
- Loss function: $\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z))$
- The first term is the reconstruction error and the second term is the unpacked Kullback-Leibler divergence between the inference model $q_\emptyset$(z|x), and p_θ(z).

Cross and Distribution Aligned VAE

- The goal of our model is to learn representations within a common space for a combination of M data modalities.
- Our model includes M encoders, one for every modality, to map into this representation space.
- In the case of matching image features with class embeddings, M = 2, $x^{(1)} \in X$ and $x^{(2)} \in C(Y^S)$.
- **VAE-Losses**: $\mathcal{L}_{VAE} = \sum_{i}^{M} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x^{(i)}|z)] - \beta D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z))$
- **Cross-Alignment (CA) Loss**: $\mathcal{L}_{CA} = \sum_{i}^{M} \sum_{j \neq i}^{M} |x^{(j)} - D_j(E_i(x^{(i)}))|$
- Every modality specific decoder is trained on the latent vectors derived from the other modalities. E_i is the encoder of a feature of i-th modality and D_j is the decoder of a feature of the same class but the j-th modality.
- **Distribution-Alignment (DA) Loss**: $\mathcal{L}_{DA} = \sum_{i}^{M} \sum_{j \neq i}^{M} W_{ij} \quad W_{ij} = \left[||\mu_i - \mu_j||_2^2 + Tr(\Sigma_i) + Tr(\Sigma_j) - 2(\Sigma_i^{\frac{1}{2}} \Sigma_i \Sigma_j^{\frac{1}{2}})^{\frac{1}{2}}\right]^{\frac{1}{2}}$
- **Cross- and Distribution Alignment (CADA-VAE) Loss**:

$$\mathcal{L}_{CADA-VAE} = \mathcal{L}_{VAE} + \gamma \mathcal{L}_{CA} + \delta \mathcal{L}_{DA}$$

- γ and δ are the weighting factors of the cross alignment and the distribution alignment loss.

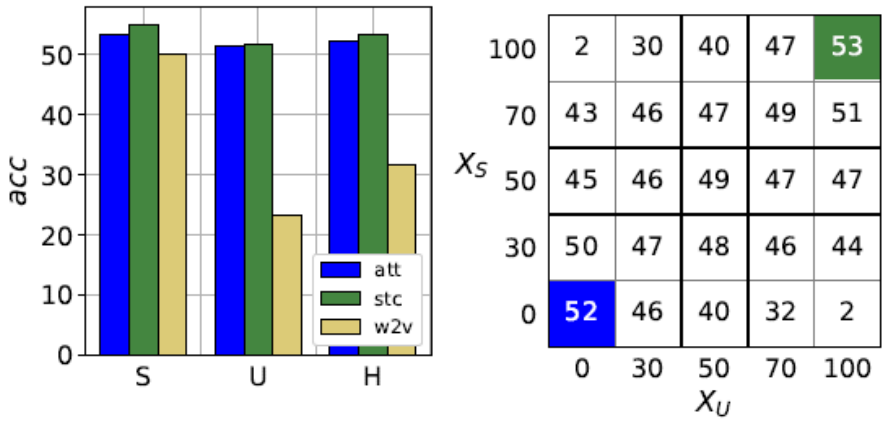
Experiment on ablation study

Model	S	U	H
DA-VAE	48.1	43.8	45.8
CA-VAE	52.6	48.1	50.2
CADA-VAE	**53.5**	**51.6**	**52.4**

Table 1: Ablation study. We compare GZSL accuracy on CUB for different multi-modal alignment objective functions, i.e. DA-VAE (distribution aligned VAE) , CA-VAE (cross-aligned VAE) and CADA-VAE (cross and distribution aligned VAE).

- Cross-alignment objective noticeably improves performance compared to distribution alignment.
- Their combination leads to the highest result on both seen, unseen classes and their harmonic mean.

Experiment on different modalities



X_S \ X_U	0	30	50	70	100
100	2	30	40	47	53
70	43	46	47	49	51
50	45	46	49	47	47
30	50	47	48	46	44
0	52	46	40	32	2

Figure 3: Effect of different class embeddings. (Left) Seen, unseen and harmonic mean accuracy for CUB using different class embeddings as side information. (Right) Using both attributes and sentences as side information, i.e. X_S: the percentage of seen classes with sentences, X_U: the percentage of unseen classes with sentences. Attributes are the class embeddings for the $(100 - X)\%$ of the classes.

- The left graph shows per-class sentence embeddings result in the best performance among all three(attributes, sentence, Word2Vec embeddings).
- X_S% of seen class image features are paired with sentence embeddings while the other (100−X_S)% of seen classes are paired with attributes, same for X_U.
- The high accuracy values in the center of table in the right graph prove that when either sentences or attributes are not available, our model can recover the missing information from the other modality and still learn discriminative representations.

Experiment on dimentionality of latent features

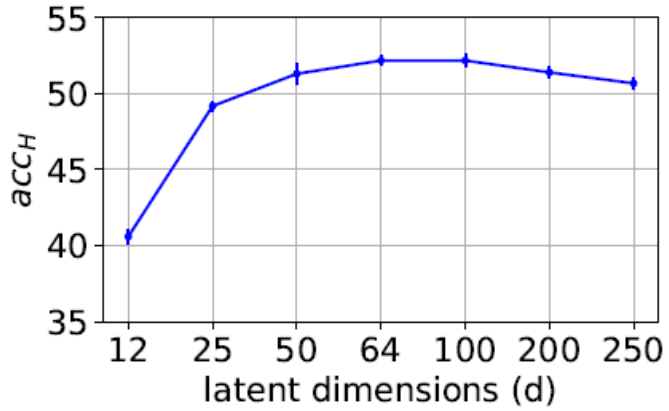


Figure 4: The influence of the dimentionality of the latent features that are generated by CADA−VAE and used to train the GZSL classifier. We measure the harmonic mean accuracy on the CUB dataset

- Accuracy initially increases with increasing dimensionality until it achieves its peak accuracy of 52.4% at d = 64 and flattens until d = 100 after which it declines upon further increase of the latent dimension.
- d=64 is used in other experiments.

Experiment on comparing with other models

		CUB			SUN			AWA1			AWA2		
Model	Feature Size	**S**	**U**	**H**	**S**	**U**	**H**	**S**	**U**	**H**	**S**	**U**	**H**
CMT [27]		49.8	7.2	12.6	21.8	8.1	11.8	87.6	0.9	1.8	90.0	0.5	1.0
SJE [2]		59.2	23.5	33.6	30.5	14.7	19.8	74.6	11.3	19.6	73.9	8.0	14.4
ALE [1]		62.8	23.7	34.4	33.1	21.8	26.3	76.1	16.8	27.5	81.8	14.0	23.9
LATEM [34]	2048	57.3	15.2	24.0	28.8	14.7	19.5	71.7	7.3	13.3	77.3	11.5	20.0
EZSL [24]		63.8	12.6	21.0	27.9	11.0	15.8	75.6	6.6	12.1	77.8	5.9	11.0
SYNC [4]		70.9	11.5	19.8	43.3	7.9	13.4	87.3	8.9	16.2	90.5	10.0	18.0
DeViSE [6]		53.0	23.8	32.8	27.4	16.9	20.9	68.7	13.4	22.4	74.7	17.1	27.8
f-CLSWGAN [36]		57.7	43.7	49.7	36.6	42.6	39.4	61.4	57.9	59.6	68.9	52.1	59.4
CVAE [18]	1024	–	–	34.5	–	–	26.7	–	–	47.2	–	–	51.2
SE [14]		53.3	41.5	46.7	30.5	40.9	34.9	67.8	56.3	61.5	68.1	58.3	62.8
ReViSE [29]	75/100	28.3	37.6	32.3	20.1	24.3	22.0	37.1	46.1	41.1	39.7	46.4	42.8
ours (CADA-VAE)	64	53.5	51.6	**52.4**	35.7	47.2	**40.6**	72.8	57.3	**64.1**	75.0	55.8	**63.9**

Table 2: Comparing CADA-VAE with the state of the art. We report per class accuracy for seen (S) and unseen (S) classes and their harmonic mean (H). All reported numbers for our method are averaged over ten runs.

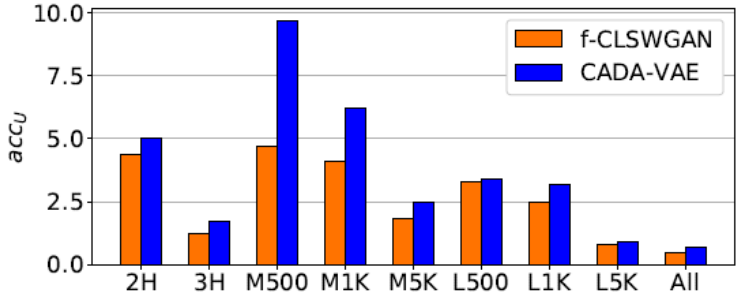


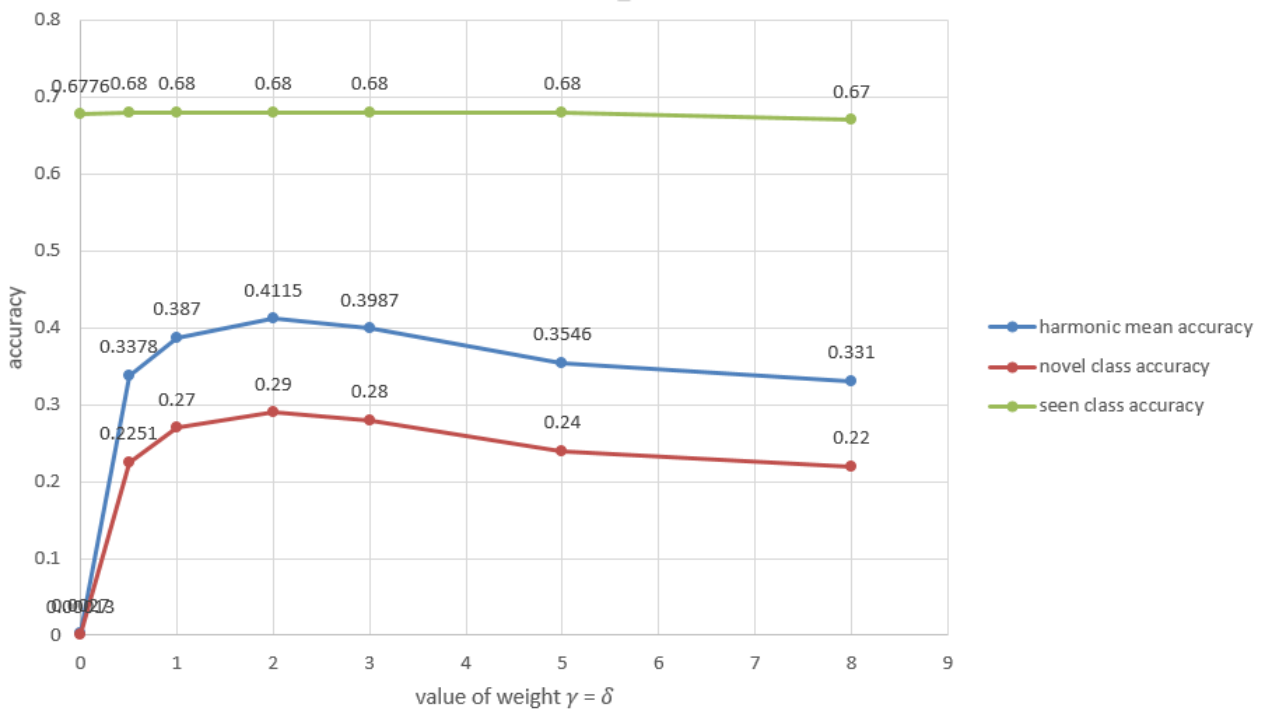
Figure 7: ImageNet results on GZSL. We report the top-1 accuracy for unseen classes. Both f-CLSWGAN and CADA-VAE use a linear softmax classifier.

- CADA-VAE uses less feature sizes and achieves better performance on CUB, SUN, AWA1, AWA2 and ImageNet dataset.

Extended experiments on VAE loss, Cross-Alignment (CA) Loss and Distribution-Alignment (DA) Loss in GZSL on CUB dataset

- In experiments, γ and δ are initialized to 0 and changed with warm-up process. The final values are γ =2.4 and δ =8.1. Now we analyze the effect of each loss.
- We set $\gamma = \delta$, and change its value.
- CA + DA loss plays key role in zero-shot learning.
- VAE loss is helpful in classifying seen cases.

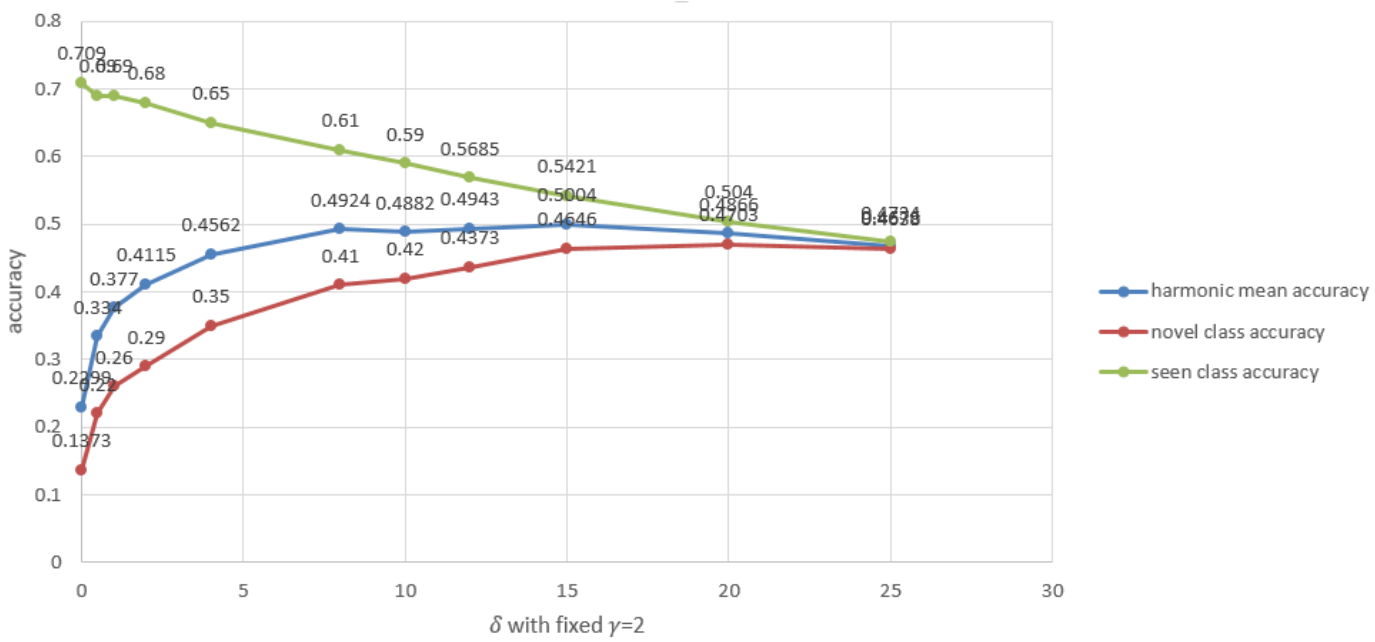
$$\mathcal{L}_{CADA-VAE} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{CA} + \delta\mathcal{L}_{DA}$$



Extended experiments on importance of DA Loss

- Fix γ =2, change δ(weight of DA loss).
- With increase of weight of DA loss, the latent representation z from different modalities becomes closer, which is helpful in learning unseen classes from attributes, seen classes have less advantage over unseen classes, but it will sacrifice the accuracy of seen cases.

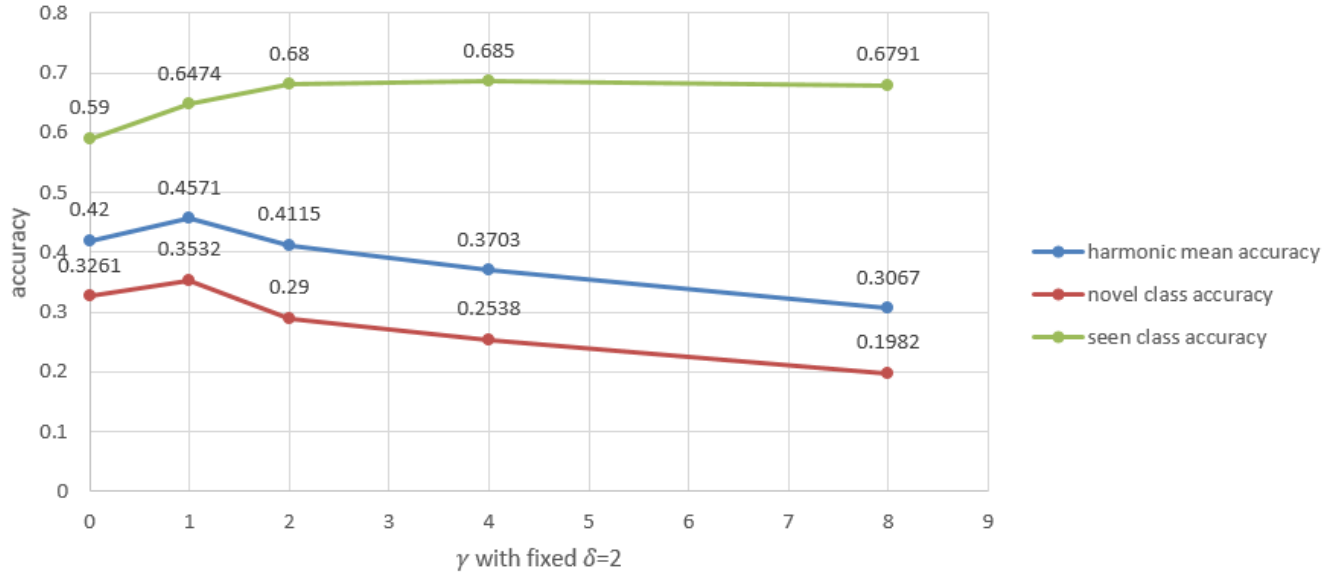
$$\mathcal{L}_{CADA-VAE} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{CA} + \delta\mathcal{L}_{DA}$$



Extended experiments on importance of CA Loss

- Fix δ =2, change γ(weight of CA loss).
- Adding CA loss to an appropriate weight could improve accuracy of both seen and unseen classes.
- In my opinion, during training, the unseen classes only have attributes, but not image features, so unseen classes can't directly impact CA loss and doesn't benefit from CA loss too much.

$$\mathcal{L}_{CADA-VAE} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{CA} + \delta\mathcal{L}_{DA}$$



Conclusion

- Proposed CADA-VAE, a cross-modal embedding framework for generalized zero- and few-shot learning.
- Train a VAE for both visual and semantic modalities.
- This procedure leaves us with encoders that can **encode features from different modalities into one cross-modal embedding space**, in which a linear softmax classifier can be trained.
- Cross-modal embedding model for generalized zero-shot learning achieves better performance than data generating methods, establishing the new state of the art.
- CA loss and DA loss greatly helps improve zero-shot learning results.