Zero-Shot Image Classification Guided by Natural Language Descriptions of Classes: A Meta-Learning Approach

R. Lily Hu
Salesforce Research
lhu@salesforce.com

Caiming Xiong
Salesforce Research
cxiong@salesforce.com

Richard Socher
Salesforce Research
rsocher@salesforce.com

Abstract

We propose a model that learns to perform zero-shot image classification from natural language descriptions by using a meta-learner that is trained to produce a correction to the output of a previously trained learner. The model consists of two modules: a task module (learner) that supplies an initial prediction, and a correction module (meta-learner) that updates the initial prediction. The correction module is trained in an episodic approach whereby many different task modules are trained on various subsets of the total training data, with the rest being used as unseen data for the correction module. The correction module takes as input a representation of the task module's training data so that the predicted correction is a function of the task module's training data. The correction module is trained to update the task module's prediction to be closer to the target value. This approach leads to state-of-the-art performance for zero-shot classification on natural language class descriptions on the CUB and NAB datasets.

1 Introduction

The ability to solve a task without receiving training examples – zero-shot learning – is desirable. We as humans can learn new tasks from descriptions of the tasks, as we learn from reading encyclopedia entries, manuals, handbooks, textbooks, etc. We propose a model that learns a correction on predictions in the zero-shot setting, based on the training data set used to generate the initial prediction. Hence, our model is called Correction Networks. The intuition for our model is that a zero-shot query sample that is different from samples in the training data will require a different correction than a zero-shot query sample that is similar to samples in the training data.

Correction Networks update the predictions based on the training data. This updated prediction is trained to be closer to the target value than the original prediction. Correction Networks consist of two modules: a *task module* that supplies an initial prediction, and a *correction module* that provides a correction to the initial prediction. The task module is the learner and the correction module is the meta-learner. The final prediction is the task module's initial prediction combined with the correction module's correction. This method is illustrated in Figure 1. The prediction of the meta-learner is used to modify the output of the learner.

Our approach is for zero-shot learning while previous meta-learning approaches focus on few shot learning and require a few samples. Another novelty is that the correction module, the meta-learner, takes as input the dataset used to train the task module. Correction Networks are independent of the representation of the task module. Existing models that provide predictions can be treated as task modules. One novelty is that the correction module, the meta-learner, also takes as input the dataset used to train the task module. The main contribution of this paper is a zero-shot learning model that corrects zero-shot predictions based on training data used to generate the initial prediction.

Preprint. Work in progress.

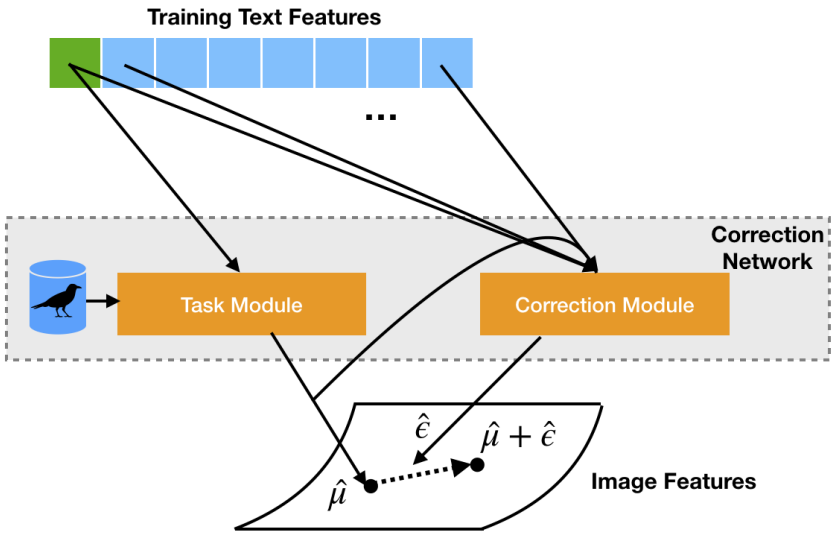

Figure 1: The task module produces an initial prediction. The correction module provides a correction such that when combined with the initial prediction, produces a better prediction.

2 Related Works

For zero-shot learning for images, the majority of state-of-the-art methods are embedding-based methods. This often involves learning a mapping from the visual space to the semantic space of class labels or vice versa. Alternatively, the embedding function between the visual and semantic spaces is jointly learned through a latent space. Given a defined attribute ontology, each class name can be converted to an attribute vector against which image features are compared (1). Text-based embeddings into which to project class names can also be used (2) (3). Semantic representations for zero-shot classes have been created from text documents of the classes eg. a Wikipedia article for each class (4) (5) (6). Zero-shot recognition has also been framed as a conventional supervised classification problem by hallucinating samples for unseen classes (4).

3 Correction Networks

Let $\mathcal{D}_S$ denote our training data and $\mathcal{D}_U$ our testing data. $\mathcal{D}_S$ is subdivided into disjoint sets $\mathcal{D}_S^s$ and $\mathcal{D}_S^u$. For classification, the classes in $\mathcal{D}_S^s$ are disjoint from the classes in $\mathcal{D}_S^u$. Correction Networks M consists of two modules: a task module M_T and a correction module M_C. The task module M_T is so-called because it is task specific and related to the application. The task module is trained on $\mathcal{D}_S^s$. The output of M_T is an estimate $\hat{\mu}$ of a target μ. The predictions of the task module on its training data $\mathcal{D}_S^s$ is $\hat{\mu}_S^s$. Training the task module proceeds by minimizing the distance between $\hat{\mu}_S^s$ and the ground truth μ_S^s. The loss is:

$$L_{M_T} = \mathbb{E}[d(M_T(T_S^s), \mu_S^s)] + \alpha||w_{M_{T,text}}||^2 \tag{1}$$

where T is the class text description, μ is the empirical mean of samples that belong to the class, and d is a distance function. We use the L2 norm as the distance function.

The task module M_T is not trained on $\mathcal{D}_S^u$ nor $\mathcal{D}_U$. The task module's predictions on $\mathcal{D}_S^u$ are $\hat{\mu}_S^u$. Likewise, the task module's predictions on $\mathcal{D}_U$ are $\hat{\mu}_U$. The correction module M_C computes a correction $\hat{\epsilon}_S^u$ that is applied to the prediction $\hat{\mu}_S^u$ of the task module M_T, where $\hat{\epsilon}_S^u$ is calculated based on the data used to train M_T, such that the corrected prediction $(\hat{\mu}_S^u + \hat{\epsilon}_S^u)$ is closer than $\hat{\mu}_S^u$ to the ground truth μ_S^u. Training the correction module proceeds by minimizing the distance between μ_S^u and $(\hat{\mu}_S^u + \hat{\epsilon}_S^u)$. We use the L2 norm. The training data for the task module D_S^s is input into the correction module by representing the training data D_S^s as an un-ordered collection of data by using a pooling function. The objective function of the correction module is to minimize:

$$L_{M_C} = \mathbb{E}[d(M_C(T_S^u), \mu_S^u - M_T(T_S^u))] \tag{2}$$

Table 1: Zero-shot learning classification results accuracy @ 1 on the CUB-200-2011 dataset and the NAB dataset using class descriptions from Wikipedia on the Super-Category-Shared (SCS) and Super-Category-Exclusive (SCE) zero-shot splits

METHOD	CUB		NAB	
	SCS	SCE	SCS	SCE
MCZSL (8)	34.7	-	-	-
WAC-Linear (9)	27.0	5.0	-	-
WAC-Kernel (6)	33.5	7.7	11.4	6.0
ESZSL (10)	28.5	7.4	24.3	6.3
SJE (11)	29.9	-	-	-
ZSLNS (5)	29.1	7.3	24.5	6.8
$\text{SynC}_{\text{fast}}$ (12)	28.0	8.6	18.4	3.8
SynC_{OVO} (12)	12.5	5.9	-	-
ZSLPP (13)	37.2	9.7	30.3	8.1
GAZSL (4)	43.7	10.3	35.6	8.6
Correction Networks	**45.8**	10.0	**37.0**	**9.5**

Table 2: Generalized Zero-shot learning classification area under Seen-Unseen Curve on CUB and NAB datasets

METHOD	CUB		NAB	
	SCS	SCE	SCS	SCE
WAC-Linear (9)	23.9	4.9	23.5	-
WAC-Kernel (6)	22.5	5.4	0.7	2.3
$\text{SynC}_{\text{Fast}}$ (12)	13.1	4.0	2.7	0.8
ESZSL (10)	18.5	4.5	9.2	2.9
ZSLNS (5)	14.7	4.4	9.3	2.3
SynC_{OvO} (12)	1.7	1.0	0.1	-
ZSLPP (13)	30.4	6.1	12.6	3.5
GAZSL (4)	35.4	8.7	20.4	5.8
CorrectionNet	**41.9**	**9.0**	**25.4**	**7.6**

We adopt the meta-learning sampling strategy for training as in (7). Training data for Correction Networks are formed by randomly selecting a subset $\mathcal{D}_S^s$ from the training data $\mathcal{D}_S$. Then, the task module M_T, is trained on $\mathcal{D}_S^s$. The remaining tasks that the task module M_T does not train on are treated as $\mathcal{D}_S^u$ for M_T. To use Correction Networks for evaluation, the task module M_T outputs $\hat{\mu}_U$ and the correction network supplies $\hat{\epsilon}_U$. The output of the Correction Networks is $\bar{\mu}_U = \hat{\mu}_U + \hat{\epsilon}_U$.

4 Experiments

We demonstrate Correction Networks on Caltech UCSD Birds 2011 (CUB) (14) and North America Birds (NAB) (15) with class data from Wikipedia. The top-1 accuracy of our method and eight state-of-the-art algorithms for the CUB and NAB datasets for both the SCS split and the SCE split are tabulated in Table 1. The eight comparison models are MCZSL (8), ZSLNS (5), SJE (11), WAC (6), SynC (12), ZSLPP (13), and GAZSL (4). The performance numbers are copied from (4). We also report generalized zero-shot learning performance using the area under the seen-unseen curve (16). This is tabulated in Table 2. Our model performs favorably against the other models.

5 Conclusion

We propose a zero-shot learning model that consists of a task module and a correction module. The training data is partitioned into a set of data used to train the task module and a disjoint set of data

used to train the correction module. Our model performs favorably against the state-of-the-art on image classification guided by natural language descriptions of novel image classes.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 819–826, IEEE, 2013.

[2] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1717–1724, IEEE, 2014.

[3] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *European Conference on Computer Vision*, pp. 584–599, Springer, 2014.

[4] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] R. Qiao, L. Liu, C. Shen, and A. van den Hengel, "Less is more: zero-shot learning from online textual documents with noise suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2249–2257, 2016.

[6] M. Elhoseiny, A. Elgammal, and B. Saleh, "Write a classifier: Predicting visual classifiers from unstructured text," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2539–2553, 2017.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

[8] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multi-cue zero-shot learning with strong supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 59–68, 2016.

[9] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2584–2591, IEEE, 2013.

[10] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International Conference on Machine Learning*, pp. 2152–2161, 2015.

[11] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 2927–2936, IEEE, 2015.

[12] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336, 2016.

[13] H. Z. Mohamed Elhoseiny, Yizhe Zhu and A. Elgammal, "Link the head to the" beak": Zero shot learning from noisy text description at part precision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[15] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

[16] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *European Conference on Computer Vision*, pp. 52–68, Springer, 2016.