Zero-Shot Domain Generalization

Udit Maniyar[1] K J Joseph[1] Aniket Anand Deshmukh[2] Urun Dogan[2]
Vineeth N Balasubramanian[1]
[1]Indian Institute of Technology Hyderabad, [2]Microsoft

Abstract

Standard supervised learning setting assumes that training data and test data come from the same distribution (domain). Domain generalization (DG) methods try to learn a model that when trained on data from multiple domains, would generalize to a new unseen domain. We extend DG to an even more challenging setting, where the label space of the unseen domain could also change. We introduce this problem as Zero-Shot Domain Generalization (to the best of our knowledge, the first such effort), where the model generalizes across new domains and also across new classes in those domains. We propose a simple strategy which effectively exploits semantic information of classes, to adapt existing DG methods to meet the demands of Zero-Shot Domain Generalization. We evaluate the proposed methods on CIFAR-10 [9], CIFAR-100 [9] and F-MNIST [19] datasets and show promising results, establishing a strong baseline to foster interest in this new research direction.

1. Introduction

Generalization is a key desideratum for machine learning models to scale to the dynamic nature of the real world. The standard supervised learning framework assumes that train and test data are from the same distribution (domain). Domain generalization techniques [4, 8, 11, 12, 14] demand to train a model in such a way that it can generalize to a novel domain at inference, by gracefully handling domain shift. However, current domain generalization methods assume the same classes to be present in all domains (including unseen test domains), which is a restriction on the application of such methods. Our work attempts to relax this assumption, and allow novel test domains to have new classes that were not present in any training domain. We introduce this harder problem as *Zero-Shot Domain Generalization*, and to the best of our knowledge, is the first such effort (Fig 1 illustrates the setting). We note that the standard zero-shot learning problem [1, 2, 7, 17, 18] provides a model to generalize to unseen classes, but assumes that datapoints come from a single known domain.

We hypothesize that learning a domain-invariant feature representation, with explicit class information, would help address Zero-Shot Domain Generalization. Encoding class information in the feature space ensures smooth transition of information from the classes that are seen during training on multiple source domains, to the unseen set of classes in the new domain. To this end, we adapt state-of-the-art domain generalization methods - in particular, Feature Critic Network (FC) [14] and Multi-Task Auto Encoder (MTAE) [8] - to ensure that their intermediate feature representations are semantically consistent. Our experimental evaluation on domain generalization variants of CIFAR-10 [9], CIFAR-100 [9] and F-MNIST [19] provide promise and establish baselines for this new research problem.

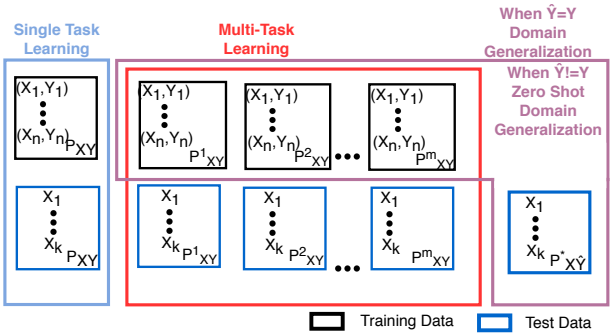

Figure 1. In a single-task learning (blue), both training data-points and test data-points comes from the same distribution P_{XY}. In multi-task learning (red), data from all distributions are available for training ($P^1_{XY} \cdots P^m_{XY}$). In domain generalization (pink), the model is evaluated on data from unseen distribution P^*_{XY}, while in Zero-Shot DG, the model is needs to be performant on $P^*_{X\hat{Y}}$, where $\hat{Y} \neq Y$.

2. Related Work

Domain Generalization: In domain generalization (DG), we are given training data from different domains and the objective is to generalize to a novel domain. [3, 4] proposed a kernel mean embedding based algorithm to get similarity between different domains and transfer the learning. The same kernel mean embedding idea was used to extend this DG framework to a multiclass setting in [6]. Ghifary *et al.* [8] proposed an autoencoder based method called Multi-Task Auto Encoder (MTAE) to learn domain-invariant features. Motiian *et al.* [15] proposed classification and contrastive semantic alignment (CCSA) to learn a domain-invariant embedding by minimizing the sum of classification loss, confusion alignment loss,

and semantic alignment loss. Li *et al*. [11] proposed a method that takes advantage of the robustness of deep learning models to domain shift, and developed a low-rank parameterized CNN model for end-to-end DG learning. Li *et al*. [12] proposed a meta-learning approach to solve for DG by simulating train-test domain shifts during training by synthesizing virtual test domains. Li *et al*. [14] extended this approach to produce a more general feature extractor that can be used with any classifier, by simultaneously learning an auxiliary loss function that trains the feature extractor for improved domain invariance. None of these efforts attempted Zero-Shot DG, setting our work apart.

Zero-Shot Learning (ZSL): The main idea in most zero-shot algorithms is to learn a semantically consistent correlation between seen and unseen classes, using semantic information between classes. They first learn an embedding function which maps visual space to semantic space, to aid in this task. Socher *et al*. [17] minimizes Euclidean distance loss with word vectors in semantic space as objective, and learns a two-layer neural network for classification. Attribute Label Embedding (ALE) [1] and Deep Visual Semantic Embedding (DEVISE) [7] learn a bilinear compatibility function to map image to semantic space. To learn the compatibility function, different objectives functions have been explored: DEVISE [7] uses pairwise ranking objective, [2] takes a dot product between the embedded visual feature and semantic vectors considering three training losses, including a binary cross entropy loss, hinge loss and Euclidean distance loss. Yongqin *et al*. [18] presents a survey of zero-shot learning methods.

Even though both DG and ZSL have been explored by the community in isolation, to the best of our knowledge, no existing work in literature has attempted to solve the Zero-Shot DG problem. This setting can have practical use in applications such as robotics, medical image analysis and general scene understanding. We identify this problem in this work, formalize the same and provide a methodology that can serve as a strong baseline for this problem.

3. Zero-Shot Domain Generalization

Problem Definition:

We define a domain as the joint distribution of feature and label space. Let the training data for the i^{th} domain be: $(X_{ij}, Y_{ij}) \sim P^i_{XY}$ and $P^i_{XY} \sim \mu$. During testing, in DG and Zero-shot DG, features and seen labels are drawn from the same distribution μ as in training.

We assume all (X, Y) pairs are drawn i.i.d. from their respective distributions. In particular, let Y^{tr} and Y^{ts} represent the set of classes in training and test data respectively, such that $Y^{tr} \cap Y^{ts} = \emptyset$. The training data for the i^{th} domain is given by $D_i = \{X_{ij}, Y_{ij}\}_{1 \leq j \leq n_i}$ and $Y_{ij} \in Y^{tr}$, and the test data set be $D^T = \{X^T_j, Y^T_j\}_{1 \leq j \leq n_T}$ and $Y^T_j \in Y^{ts}$ where n_i is the number of images in the i^{th} domain. The main objective of both DG problems is to train a model on all training domains $D = \{D_1, D_2, \ldots D_N\}$ to perform well on D^T.

Proposed Approach: We propose a generic approach to extend the state-of-the-art DG techniques to solve zero-shot domain generalization. The key insight is to bind the intermediate domain-invariant feature representations to a semantic space that is shared across the seen classes of the old domains and the unseen classes of the new domain. Such use of semantic space has been successfully used in recent ZSL methodologies [18].

Existing DG methods [8, 14] can be considered as a composition of a feature extractor function f_θ and classifier function g_ϕ: $(g \circ f)(\boldsymbol{I})$, where $\boldsymbol{I}$ is the input image. In this paper, we only consider DG methods which generate domain invariant features and train a common classifier. We restrict domain invariant features to semantic space, forcing the model to be semantically consistent along with domain invariance. Our proposed method can be extended to any DG method which learns domain-invariant features but does not apply to methods which learn different features for different domains. While training the proposed semantically constrained DG approach, the features generated by f_θ are projected to a semantic space of labels. Images of similar classes are grouped together, and dissimilar classes spaced apart by a semantic alignment loss given in Equation 1. By using the semantic embedding and restricting lower-level features of the model to a semantic space, a shared invariant representation is learned where semantic alignment is accounted for. We make an inherent assumption that classes which appear visually similar should also be semantically similar (as in other ZSL methods). Thus, using semantic space helps us in the visual classification task. We use word embeddings of classes - in particular, simple GloVe embeddings [16] trained on Common Crawl corpus - as the semantic space in this work. One could use more complex embedding functions to study this even further. During inference, we depart from traditional DG methods which train a separate classifier (neural network or SVM) on the domain-invariant features. We instead use nearest neighbour (NN) search in the semantic space. Our method is lightweight and can scale to large numbers of unseen classes using existing efficient NN search methods. We now describe how we infuse semantic information into three DG methods, to solve zero-shot DG.

Semantic AGG (S-AGG): Aggregation (AGG) is a simple baseline, which has strong DG performance [11, 14]. Here, we group data from all domains and train the network on this multi-domain dataset. In line with our generic definition for DG functions, the model is split into two parts: feature extractor f_θ and classifier g_ϕ. f_θ contains a series of convolutional layers followed by fully connected layers which map the image from a higher dimension $\mathbb{R}^x$ to class-

vector dimension $\mathbb{R}^h$. g_ϕ is also a classifier (neural network) which maps from $\mathbb{R}^h$ to number of classes. We now define our semantic loss in this case as follows:

$$\mathcal{L}^{Semantic}(\theta) = ||(f_\theta(X_{ij}) - w[Y_{ij}])||^2 \quad (1)$$

where $w[Y_{ij}]$ denotes the word embedding of the label of image X_{ij}. Semantic AGG (S-AGG) is hence trained to minimize the following loss function: $\min_{\theta,\phi} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathcal{L}^{CE}(g_\phi(f_\theta(X_{ij})), Y_{ij}) + \mathcal{L}^{Semantic}(\theta) + \eta\mathcal{R}(\theta,\phi)$, where $\mathcal{L}^{CE}$ is cross-entropy loss and $\mathcal{R}$ is a regularizer.

Semantic MTAE (S-MTAE): Multi-task Auto Encoder (MTAE) [8] learns domain-invariant features using an autoencoding framework. In MTAE, the encoder acts as a feature extractor. A single encoder (h_Θ) is maintained across domains, which projects each image to a latent representation. Domain-specific decoders ($g_{\Phi_1}, g_{\Phi_2}, \ldots g_{\Phi_N}$) take these representations and regenerate the image back to each domain. This implicitly forces the encoder to learn an unbiased, domain-agnostic projection function. The final classification function is learned using these domain-invariant features with an SVM or a simple MLP. We adapt MTAE for Zero-Shot DG by restricting the latent representation to the semantic embedding of the class labels.

For every pair of domains $(D_i, D_j)_{1\leq i,j\leq N}$, images of domain D_j are generated from images of domain D_i : $(g_{\Phi_j} \circ h_\Theta)(\boldsymbol{I})$ where $\boldsymbol{I}$ is an image of class k in domain D_j that is regenerated (decoded) from images of class k in domain D_i. Our semantic loss restricts the feature space of MTAE to the semantic embedding space of classes, along with standard reconstruction loss. The loss function for S-MTAE while training domain i is hence: $\mathcal{L}(\Theta, \Phi_1, \ldots, \Phi_N, i) = \sum_{j=1}^{N}\mathcal{L}^{MSE}(g_{\Phi_j}(h_\Theta(D_i)), D_j) + \mathcal{L}^{Semantic}(h_\Theta(D_i), w[Y_{D_i}])$, where $w[Y_{D_i}]$ is the word embedding of classes in Y_{D_i}. We minimize the following objective to train across N domains:

$$\min_{\Theta,\Phi_1,\ldots,\Phi_N} \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\Theta, \Phi_1, \ldots, \Phi_N, i) + \eta\mathcal{R}(\Theta, \Phi_1, \ldots, \Phi_N)$$

Semantic FC (S-FC): Feature Critic Networks (FC) [14] provides a meta-learning approach to DG. It learns a domain-invariant feature extractor by meta-learning an auxiliary loss that 'criticizes' the effectiveness of features generated by the feature extractor, when dealing with an unseen domain. The network is trained by simulating training-to-testing domain shift by splitting the source domains into virtual training and testing meta-domains, following standard meta-learning practice. FC can also be viewed as a composition of a feature extractor f_θ and classifier g_ϕ. While training, we split D into $D^{tr}_{meta}, D^{ts}_{meta}$ such that $D^{tr}_{meta} \cap D^{ts}_{meta} = \emptyset$. We train the model on D^{tr}_{meta} but expect it to perform well on D^{ts}_{meta}. The model is trained using the following loss function:

$$\min_{\theta,\phi} \sum_{D_i \in D^{tr}_{meta}} \sum_{X,Y \in D_i} \mathcal{L}^{CE}(g_\phi(h_\theta(X)), Y) + \mathcal{L}^{Semantic}(h_\theta(X), w[Y]) + \mathcal{L}^{Aux}$$

$w[Y]$ refers to word vectors of classes of Y, $\mathcal{L}^{CE}$ is cross-entropy loss, and $\mathcal{L}^{Semantic}$ is the semantic loss defined in Eqn 1. $\mathcal{L}^{Aux}$ is the meta-loss that encourages feature extractor to generate domain-agnostic features. We refer readers to [14] for specifics of the meta-training strategy.

4. Experiments and Results

We evaluate the proposed methods on three different datasets: CIFAR-10 [9], Fashion-MNIST [19], CIFAR-100 [9]. Similar to [5, 8, 13], different domains are obtained by rotating the images by 15° (zero-padded as required after rotation). PACS [11] and Rotated-MNIST[10] are often used in earlier DG work [5, 8, 12, 14, 13]. However, we restrain from using Rotated-MNIST because of the lack of connection between the visual space and semantic space of numbers here, and PACS contains only 7 classes, which is too few to study zero-shot classification. We perform leave-one-out experiments with these generated domains for all datasets. We measure standard DG performance as well as Zero-Shot DG performance on the left-out domain in each experiment. Our experiments are carried out on multiple settings as described below (each setting denotes the zero-shot classes used in the experiment on each dataset): (i) *CIFAR-10:* Setting 1: (cat, truck); Setting 2: (cat, dog); Setting 3: (deer, ship); Setting 4: (car, deer); Setting 5: (airplane, car); Same zero-shot classes were used in [17] (ii) *Fashion-MNIST:* Setting 1: (t-shirt, sandal); Setting 2: (sandal, shirt); Setting 3: (t-shirt, boot); Setting 4: (sandal, Boot); (iii) *CIFAR-100:* Setting 1: (whale, fish, rose, can, orange, lamp, couch, beetle, tiger, skyscraper, mountain, kangaroo, fox, snail, man, snake, squirrel, pine-tree, motorcycle, streetcar); Setting 2: (seal, shark, poppy, bottle, apple, keyboard, table, caterpillar, lion, bridge, forest, camel, raccoon, crab, girl, dinosaur, rabbit, maple, bicycle, tractor).

Implementation Details: All the reported results are averaged across five runs. The partition size of D^{tr} and D^{ts} in Semantic FC is 3:2 i.e 3 domains are chosen as meta-train and 2 as meta-test. We compare the performance across 6 different methods (AGG, S-AGG, FC, S-FC, MTAE, S-MTAE). By AGG, FC and MTAE, we mean the vanilla method without semantic loss. In these cases, even though we don't use semantic loss during training, the ZSDG accuracies are computed using semantic distances (since that is required for zero-shot classification). The vanilla method helps understand the usefulness of adding the semantic loss.

Results: Tables 1, 2 and 3 present DG and ZSDG results on F-MNIST [19], CIFAR-10 [9] and CIFAR-100 [9] datasets respectively. For brevity, we present the average accuracy in

TARGET	Setting 1		Setting 2		Setting 3		Setting 4	
	DG	ZSDG	DG	ZSDG	DG	ZSDG	DG	ZSDG
AGG	67.16	61.51	70.24	51.59	67.21	57.13	60.87	53.36
S-AGG	69.11	57.673	73.19	56.87	68.08	41.88	62.37	52.5
MTAE	18.12	73.105	18.41	70.79	17.50	79.44	17.62	**64.26**
S-MTAE	**72.97**	**92.45**	**77.29**	**89.54**	**72.17**	**89.00**	**65.56**	52.71
FC	66.17	56.61	69.03	52.33	66.14	53.18	59.18	51.41
S-FC	66.53	49.03	69.93	54.24	66.33	61.04	58.83	53.94

Table 1. Domain Generalization (DG) and Zero-Shot Domain Generalization (ZSDG) performance on rotated domains on Fashion-MNIST dataset.

TARGET	Setting 1		Setting 2		Setting 3		Setting 4		Setting 5	
	DG	ZSDG	DG	ZSDG	DG	ZSDG	DG	ZSDG	DG	ZSDG
AGG	51.58	48.94	50.85	49.87	48.79	42.63	49.42	52.40	47.93	51.21
S-AGG	48.55	79.77	48.68	53.58	46.04	**81.75**	46.26	82.59	44.94	65.86
FC	**52.18**	54.82	51.87	50.19	**49.56**	51.69	**49.95**	45.00	**48.37**	52.40
S-FC	51.35	**81.1**	50.98	55.3	48.53	77.37	49.29	81.59	47.79	71.15
MTAE	12.14	54.55	11.74	51.19	12.67	56.55	12.41	52.91	12.66	54.62
S-MTAE	51.92	80.12	**51.94**	**55.35**	49.13	79.94	49.42	**83.2**	47.95	**71.63**

Table 2. Domain Generalization (DG) and Zero-Shot Domain Generalization (ZSDG) performance on rotated domains on CIFAR-10 dataset.

	Setting 1		Setting 2	
	DG	ZSDG	DG	ZSDG
AGG	80.31	5.87	80.47	6.08
S-AGG	74.98	19.99	75.5	20.11
FC	**83.62**	5.5	**83.62**	5.52
S-FC	83.47	**20.17**	**83.62**	**20.7**
MTAE	1.45	5.00	1.29	5.44
S-MTAE	82.03	19.26	82.16	19.24

Table 3. Domain Generalization (DG) and Zero-Shot Domain Generalization (ZSDG) performance on rotated domains on CIFAR-100 dataset.

Target:	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5
zero-shot	93.17	53.55	86.06	88.31	67.54

Table 4. Zero-Shot classification accuracies on CIFAR-10 using [17].

these tables, and present the complete results with standard deviations across our multiple runs in the Supplementary material. We see in the tables that the proposed semantically consistent adaptations of DG methods perform better both on DG and ZSDG. On rotations of F-MNIST, S-MTAE performs better than all other methods with the exception of Setting 4. On rotations of CIFAR-10, S-MTAE and S-FC both perform the best. On rotations of CIFAR-100, S-FC performs better when compared to other methods. From the results, one can hypothesize that for simpler datasets, basic DG methods such as MTAE are sufficient and yield good performance, but when the complexity in the dataset increases (as in CIFAR-100), more complex methods such as FC are required for better performance. For purposes of better understanding, we also present in Table 4 ZSL performance on the considered settings on the CIFAR-10 dataset using the method proposed in [17]. We note that the comparatively lower numbers in Table 2 is because we only use 4000 images from the rotated domains of CIFAR-10, which is lower than the original domain. We expect to get higher numbers when this is increased. More results and ablation studies (different weightings of semantic loss) are in the Supplementary Section due to space constraints.

5. Conclusion

We introduce a novel Zero-shot Domain Generalization problem, where a model is expected to generalize to new classes in an unseen domain. We find that learning semantically consistent domain-invariant features helps address this challenging problem. We adapt current state-of-the-art DG methods [8, 14] to this setting to reveal the efficacy of our proposed approach, as well as provide a baseline for further efforts on this problem.

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *CoRR*, abs/1503.08677, 2015. 1, 2

[2] Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. *CoRR*, abs/1506.00511, 2015. 1, 2

[3] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017. 1

[4] G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NeurIPS*, pages 2178–2186. 2011. 1

[5] Aniket Anand Deshmukh, Yunwen Lei, Srinagesh Sharma, Urun Dogan, James W Cutler, and Clayton Scott. A generalization error bound for multi-class domain generalization. *arXiv preprint arXiv:1905.10392*, 2019. 3

[6] Aniket Anand Deshmukh, Srinagesh Sharma, James W Cutler, and Clayton Scott. Multiclass domain generalization. In *NIPS workshop on Limited Labelled Data*, 2017. 1

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129. 2013. 1, 2

[8] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559, 2015. 1, 2, 3, 4

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3

[10] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. 3

[11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 1, 2, 3

[12] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 1, 2, 3

[13] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018. 3

[14] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pages 3915–3924, 2019. 1, 2, 3, 4

[15] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5715–5725, 2017. 1

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 2

[17] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NeurIPS*, pages 935–943, 2013. 1, 2, 3, 4

[18] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017. 1, 2

[19] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1, 3