A Zero-shot Learning Method with a Multi-modal Knowledge Graph

Yuhong Zhang, Haitao Shu, Chenyang Bu and Xuegang Hu
Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei, China
School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
zhangyh@hfut.edu.cn, shuhaitao@mail.hfut.edu.cn, {chenyangbu, jsjxhuxg}@hfut.edu.cn

***Abstract*—Zero-shot learning aims to recognize unseen-classes using some seen-class samples as training set. It is challenging owing to that the feature representations of unseen-class samples are unavailable. Existing methods transfer the mapping from seen-classes to unseen-classes with the correlation as a bridge, in which, the semantic representations are used to discriminate the classes. However, the unavailability of visual representations for unseen-classes and the insufficient discrimination of semantic representations make the zero-shot learning challenging. Therefore, the visual representations are learned as complements to semantic representations to construct a multi-modal knowledge graph (KG), and a zero-shot learning method based on multi-modal KG is proposed in this paper. Specially, a semantic KG is introduced to capture the correlation of classes, and with the correlation, the visual feature representations of all classes are learned. Then, the discriminative visual representations and the semantic representations are used together to construct a multi-modal KG. With the multi-modal KG, the classifier for seen-classes is transferred to unseen classes. Extensive experimental results show the effectiveness of our method.**

***Index Terms*—Zero-shot Learning, Knowledge Graph, Multi-modal Representation.**

I. Introduction

Transfer learning has been widely used in CV, NLP, and other fields [1]–[3]. Existing methods assume that source and target domains share the same class space. However, this assumption may not be hold in applications. There may be samples with new classes in target domain that do not appear in source domain, which is called zero-shot learning [4], [5]. Zero-shot learning relaxes the assumption of both domains sharing the same classes and is a very challenging and meaningful research problem.

The zero-shot learning methods try to train a classifier for unseen-class samples based on seen-class samples, in which the representations of unseen-classes need to be learned by the correlation between seen- and unseen-classes. Specially, the mapping for seen-classes is trained, and then with the correlation between seen and unseen classes as a bridge, the mapping is transferred to learn the representations for unseen-classes. Existing methods can be roughly divided into two categories. One category of methods represent classes in view of semantic features with implicit correlation. These semantic features of classes include attributes, word vectors or sentence descriptions [4], [6]–[9]. The other category of methods use the correlation based on graphs, in which, the correlation is represented explicitly. Compared with semantic representations, graphs can represent not only semantic distance but also the relationship between semantics, which are more comprehensive prior knowledge. A subgraph is constructed based on WordNet [10] in [11], [12] to represent the correlation between classes, and represent the graph nodes with word vectors, and then Graph Convolutional Network (GCN) [13] is used to extend the classifier for seen-classes to unseen-class samples with correlations as bridge.

It can be seen that the representations for different classes are important for zero-shot learning. Most methods use semantic representations owing to the unavailability of visual representations, they still have some limitations. In some applications, the semantic representations may be less discriminative. For example, the semantic similarity between zebras and birds is low, so they can be distinguished by semantic embedding. While the similarity between polar bears and pandas is is higher, it is not be easy to distinguish them using semantic embedding only. However,the visual representations for unseen-class samples are unavailable in zero-shot learning. If all classes, especially the unseen-classes can be represented by visual features, polar bears and pandas will be distinguished more easily. Thus, to learn the visual representations and treated it as complements to the semantic representations will benefit to zero-shotting learning.

In this paper, we propose a zero-shot learning method based on a multi-modal KG. the visual representations for all classes are learned, and the visual representations and the word embedding are combined to construct a multi-modal KG, in which, the nodes of classes are represented in a multi-modal manner and the edges are treated to model the correlation of classes explicitly. That also means that the classes are represented with both semantics and visual features. With multi-modal KG as a bridge, more related information will be transferred from seen-class nodes to unseen-class nodes to achieve zero-shot learning. Our contributions are as following.

- To address the unavailability of visual features for unseen-classes, correlation of classes in KG is used to learn the discriminative visual features for unseen-classes. It can avoid the instability and collapse caused by GAN.
- The visual representations are combined with semantic representations to construct a multi-modal KG. The

This work is supported by the National Natural Science Foundation of China under grant 61976077 and the Natural Science Foundation of Anhui Province under grant 2208085MF170.

2375-0197/22/$31.00 ©2022 IEEE
DOI 10.1109/ICTAI56018.2022.00064

multi-modal KG is more discriminative to distinguish the classes, and also benefits to learn more informative representations for unseen-class samples. To our best knowledge, it is the first time to introduce multi-modal KG into zero-shotting.

II. Related Work

Zero-shot learning methods can be divided into two types, implicit correlation based, and explicit correlation based.

A. Implicit correlation methods for zero-shot learning

Most methods with implicit correlation use generative adversarial network (GAN) [14], [15] to construct samples for unseen-classes, and then use these samples and seen-classes to jointly train the classifier. The challenging problem is the lack of unseen-class samples, which makes some methods unusable. The emergence of GAN can just solve this problem, using GAN to synthesize samples of unseen-classes. To improve the quality of synthetic samples, Oral-GMN [16] proposes a new gradient matching loss, which is used to measure the difference between real samples and gradient vectors obtained by synthetic samples. Compared with the previous methods of generating sample images, FGN [17] directly generates feature representations for unseen class samples. However, the feature learning with GAN tends to be instable and collapse easily. Furthermore, the time cost increases with size of training increasing.

B. Explicit correlation methods for zero-shot learning

Recently, KGs are gradually introduced into zero-shot learning, which contains both rich semantic and the correlation of classes. Methods with explicit correlation introduce KG and use the edges as correlation between classes. GCNZ [11] represents the nodes of KG with word vectors, integrating the semantic and correlation into a KG, and then GCN is used to expand the classifier of seen-classes to unseen-classes to classify for unseen classes. To address the redundant Laplace smoothing in GCNZ, a GCN network with only two layers is used in [12], and proposes a DGP method to aggregate a wider range of information.However, the above methods represent the nodes only in view of semantic representations, ignoring that the visual representations are also informative for classes.

III. Our Method

Given a training set D_s with seen-classes C_s and a class set for testing C_u, zero-shot learning aims to build a classifier for testing set, where $D_s = \{x_i, c_i\}^{n_s}$ is defined as training data with n_s samples, and $x_i \in R^d$ and $C_s = \{1 \ldots\ldots c_s\}$ corresponds to the class of i-th sample. Define $D_u = \{x_j\}_j^{n_u}$ as test set with n_u samples, and the class space of D_u is denoted as $C_u = \{c_s + 1 \ldots\ldots c_s + c_u\}$, and the class of each x_j is unknown. In zero-shot learning, C_s and C_u do not intersect. The word vectors can be used to represent the semantics of classes, denoted as $\omega_{s+u} = \{\omega_i\}_i^{c_s+c_u}$. The correlation between classes is represented by KG $G = \{V, R\}$, where $c_s \in V$ and $c_u \in V$, R represents the correlation between classes. Our task is to train a classifier for unseen-class samples in the test set.

The framework of our method is shown in Fig 1, which includes three components. 1) Learning the visual representations for seen-classes: For the training set D_s, the convolutional neural network (CNN) is used to extract the visual representations for seen-classes, denoted as F_{C_s}. It provides the supervised information for classification of unseen-classes. 2) Optimizing the visual representations of seen- and unseen-classes: The training set D_s and its corresponding word vectors ω_s are used to establish the mapping from word vectors to visual representations with the generative adversarial network (GDAN) [18], and then the ω_{s+u} can be mapped to generate the visual representations of all classes F_{s+u}. In order to enrich the visual representations and improve the discriminability for classes, the correlations from KG are introduced to optimize the generated visual representations $\tilde{F}_{s+u}$. 3) Training the classifier with a multi-modal graph: Both the visual representations and semantic vectors of the classes learned above are combined to represent KG in multi-modal way. With the multi-modal representation of classes and the correlation in KG, the classifier for unseen-class samples is trained based on GCN.

A. Learning the visual representations for seen-classes

To avoid the complicated parameter calculation process, we design a simple step to learn the visual representations for seen-classes. Since the samples with the same classes have similar distributions, we directly use the pre-trained ResNet-50 to extract visual vectors for the samples with the same classes in the training set and adopt the average vectors of these visual vectors as the visual representations of seen-classes. The specific operation is shown as follows.

$$F_c = \{MEAN(v_i | \forall_i C_i^s = c)\} \tag{1}$$

where F_c represents the average visual representations of the corresponding class, v_i and C_i^s are the visual vectors and class of the i-th sample.

B. Optimizing visual representations for all classes

The visual representations for unseen-classes are unavailable. In this subsection, we learn the initial representations for unseen-class and then optimize them.

1) The initial visual representations: We obtain the initial visual representations for unseen-classes with the help of seen-classes. Based on seen-classes, the mapping for seen-classes is constructed from word vectors to visual features, and then this mapping is transferred to learn visual features for unseen-classes from their word vectors.

Here, the word vectors are treated as the initial representations for all classes. It should be noted that our goal is not to generate unseen images, but to generate visual representations for unseen-class samples. GDAN is chosen in this paper, which is a generative dual adversarial network with multiple components. With the training set D_s and the word vectors ω_s for their classes, GDAN is used to train the mapping from ω_s

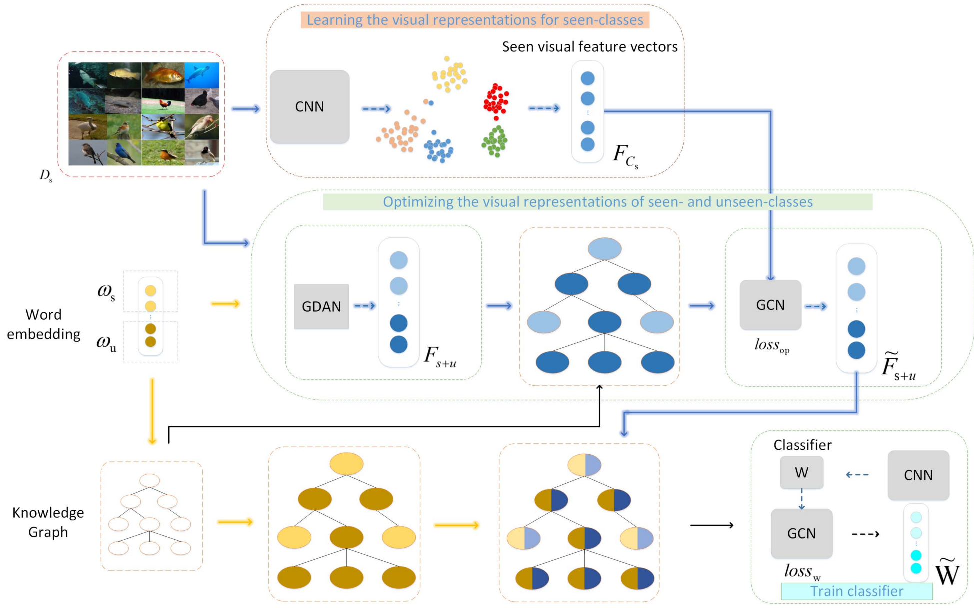

Fig. 1. The framework of our method. The yellow and blue lines and shapes denote the semantic and visual representations respectively.

TABLE I

TOP-K ACCURACY OF ALL METHODS ON IMAGENET DATASET. RESULTS DENOTED BY *, †, AND ‡ ARE FROM [19], [20], AND [11], AND RESULTS OF SGCN AND DGP ARE FROM THEIR PUBLIC CODES [12].

	Hit@k(%)														
	2-hops					3-hops					All				
	1	2	5	10	20	1	2	5	10	20	1	2	5	10	20
$ConSE^*$	8.3	12.9	21.8	30.9	41.7	2.6	4.1	7.3	11.1	16.4	1.3	2.1	3.8	5.8	8.7
$SYNC^*$	10.5	17.7	28.6	40.1	52	2.9	4.9	9.2	14.2	20.9	1.4	2.4	4.5	7.1	10.9
$EXEM^\dagger$	12.5	19.5	32.3	43.7	55.2	3.6	5.9	10.7	16.1	23.1	1.8	2.9	5.3	8.2	12.2
$GCNZ^\ddagger$	19.8	33.3	53.2	65.4	74.6	4.1	7.5	14.2	20.2	27.7	1.8	3.3	6.3	9.1	12.7
SGCN	25.51	39.4	58.88	70.73	80.03	5.79	10.3	18.39	26.41	35.94	2.71	4.74	8.87	13.15	18.78
DGP	26.16	39.9	59.31	71.14	80.29	6.13	10.37	18.78	27.06	36.83	2.87	4.91	9.08	13.54	19.32
our	25.94	40.16	59.26	71.02	80.29	5.85	10.18	18.58	26.65	36.08	2.73	4.8	8.96	13.28	18.92
our-dense	**26.25**	**40.3**	**59.33**	**71.37**	**80.67**	**6.22**	**10.6**	**19.07**	**27.4**	**37.18**	**2.93**	**5.04**	**9.27**	**13.77**	**19.62**

to D_s. Specially, CVAE in GDAN is used to generate visual features, and the generator can be regarded as the mapping between word vectors and visual representations. Based on the generator, word vectors can be used to obtain the visual feature representation F_{s+u}.

2) Optimizing the visual features for all classes: Owing to the unstability of GDAN, visual representations are not discriminative and informative strongly. Therefore, we introduce the explicit correlation from WordNet to optimize visual representations.

Based on the subgraph containing both implicit correlation of word vectors and explicit correlation from graph, the visual features are optimized by GCN. The generated visual representations F_{s+u} obtained by GDAN are used as the initial node representations in the graph. Using GCN, the representations are optimized by the following propagation.

$$\tilde{F}_{s+u} = \sigma(D^{-1}A\sigma(D^{-1}AF_{s+u}\Theta)\Theta) \tag{2}$$

where $\tilde{F}_{s+u}$ represents the result of two-layer network activation, and Θ represents the trainable weight matrix. Taking F_{s+u} as the input matrix of the first layer, $\sigma(\cdot)$ represents a nonlinear activation function, which is a Leaky ReLU in our experiments. $A \in R^{C\times C}$ is a symmetric matrix that represents the correlation of classes from graph, and C represents the number of classes in our graph. $D \in R^{C\times C}$ is a degree matrix that normalizes the rows in A to ensure that the scale of the feature representation is not modified by A.

With F_{C_s} as the supervised data for the training process, and the optimized visual representations of the seen-classes are learned with the following losses.

$$Loss_{op} = \frac{1}{c_s}\sum_{i}^{c_s} L_{mse}(\tilde{F}_i, F_i) \tag{3}$$

Thus, the F_{C_s} of the seen-classes are extended to the unseen-classes in the training and the visual representations F_{C_u} of the unseen-classes are gotten, the visual representation for seen-class is optimized as well as.

TABLE II
TOP-1 ACCURACY RESULTS FOR UNSEEN-CLASSES ON AWA2 AND aPY. THE RESULTS OF THE CONSE, DEVISE AND SYNC ARE FROM [21]. FOR AWA2, THE RESULTS OF SE-GZSL, GCNZ, AND DGP ARE FROM [12], AND THE RESULTS OF SGCN ARE GENERATED BY OUR REPRODUCTION OF [12]. ON aPY, THE RESULTS OF EXPLICIT METHODS ARE FROM THEIR PUBLIC CODES.

	Implicit correlation methods					Explicit correlation methods			
	ConSE [22]	Devise [6]	SYNC [19]	SE-GZSL [23]	Gaussian-Ort [21]	GCNZ [11]	SGCN [12]	DGP [12]	Our
AWA2	44.5	59.7	46.6	69.2	70.5	70.7	76.64	77.3	**79.56**
aPY	26.9	39.8	23.9	-	**45.3**	17.91	19.03	19.08	26.87

C. The training for classifier with multi-modal KG

After the previous step of training, we have obtained informative visual representations for all classes, which are combined with the word vectors to obtain the multi-modal representations for KG.

$$P = \omega_{s+u} \bigoplus \tilde{F}_{s+u} \tag{4}$$

where $\bigoplus$ means the concanction operation.

In order to get a classifier for unseen-class samples, we extract the last layer weights of the ResNet-50 model, which has been pre-trained on the ImageNet2012 dataset. With the multi-modal KG and the extracted classifier weights for seen-class (denoted with W_i), GCN is used to train the last layer weights for both seen- and unseen-classes. Thus, the classifier weights for seen-classes will be extended to unseen-classes. In this extending process, the model is trained to predict the classifier weights for the seen-class by optimizing the $Loss_w$.

$$Loss_w = \frac{1}{c_s} \sum_{i}^{c_s} L_{mse}(\tilde{W}_i, W_i) \tag{5}$$

IV. EXPERIMENTS

A. Experimental details

We conduct experiments on three benchmark datasets to evaluate the effectiveness of our method, including **ImageNet** [24],**AWA2** [25] and **Attribute Pascal and Yahoo (aPY)** [26].

In the learning for initial visual representations, the word vectors are 300-dimensional vectors trained by Glove. We choose Adam [27] as our optimizer with momentum set to (0.9, 0.999). In order to reduce the training cost of GAN, we set the epoch to 1. As for GCN used in both the optimizing representation for visual representations and the training for classifier with multi-modal KG, their parameters are set identically. To avoid the over-fitting, dropout [28] is further used with a dropout rate of 0.5 for each layer. The Adam optimizer is also used here, trained for 3000 epochs, with a learning rate of 0.001 and a weight decay of 0.0005. We use Leaky ReLUs with a negative slope of 0.2. We replace the original weights in ResNet-50 with the predicted weights, and use SGD to fine-tune the features for 20 epochs with a learning rate of 0.0001 and a momentum of 0.9 to accommodate the new weights. In our experiments, considering the huge number of samples and large number of unseen classes on ImageNet, we use a fine-tuned feature extractor. However, the fine-tuned feature extractor is not used on the small datasets AWA2 and aPY. Training and testing are performed on a GTX 3060 GPU.

TABLE III
ABLATION RESULTS ON AWA2.

Model	Semantic	Visual			ACC(%)
	word	random	initF	op	
SGCN	√				76.64
Our-op-initF			√	√	22.26
Our-W-randF	√	√			62.63
Our-W-initF	√		√		66.76
Our-W-op-randF	√	√		√	78.01
Our	√		√	√	**79.56**

B. Overall results

In the experiments, both implicit and explicit correlation baselines are compared with our method. The implicit correlation baselines include ConSE [22], EXEM [20], SYNC [19], Devise [6], SE-GZSL [23], Gaussian-Ort [21], while explicit correlation baselines include GCNZ [11], SGCN [12], DGP [12]. The results of all methods on three datasets are shown in Table I, Table II. 1) Explicit correlations modeled by graph are more effective than all implicit correlations modeled by word vectors. 2) Multi-modal graph represented with both semantic and visual features is more effective than represented only with semantics. Compared with GCNZ SGCN and DGP, our method achieves the best results on three dataset. Specially, our method achieves the best results on three challenges of different difficulty on the Imagenet dataset, meanwhile, our method outperforms 2.34% and 7.79% on the other two datasets. This is because that GCNZ SGCN and DGP use the single-modal KG with semantic representations, while our method use multi-modal KG with visual and semantic representations. In addition, DGP uses dense graph while SGCN uses non-dense graph. Thus our-dense is designed to compare with DGP for a fair comparison, in which, the dense-graph is constructed. 3) Compared with ImageNet and AWA2, the methods achieve lower results on the challenging aPY. However, our method improves significantly on the aPY dataset compared to other methods using graphs. The reason may be that some images in aPY dataset include multiple objects. This multiple-object images makes the classification difficult. On this occasion, our method combines the visual features with word vectors, which will provide more discriminative information for unseen-classes and then improve the classification for unseen-class samples.

C. Ablation analysis

The Ablation is shown in Table III. The performance of all variants decrease, which shows each component of our method is effective for zero-shot learning. In this subsection, we will explore the effect of each contribution further.

The effectiveness of multi-modal representations SGCN only uses the semantic representations with word vectors, while our-op-initF only uses the visual representations. Compared with our method (our-w-op-initF), the performance of both SGCN and our-op-initF decrease. It shows that the multi-modal representations for classes will be more discriminative and be beneficial for zero-shot. In addition, our-op-initF performs worse obviously, it shows that the semantic representations are more discriminative than visual representations in most situations. The reason may lie in that the visual representations are learned from the seen-classes with the correlations, while the semantic representations are obtained from pre-trained Glove model.

The effectiveness of optimized visual representations We design two ways to learn the initial visual representations and an optimization step in the ablation. Compared with the random and optimized random, the init-features and optimized init-features perform better, and it illustrate that the initial visual features learned with GDAN is more effective than random initialization. Moreover, compared with random and init-features, optimized random and optimized init-features achieve higher performance, it illustrates that the optimization strategy for visual representations with the KG is more effective. It can be induced that the optimization for visual representations can capture more informative visual features with the correlation of classes from KG.

V. Conclusions

In this paper, we propose a multi-modal KG for zero-shotting. A KG is introduced to capture the correlation of classes explicitly, and the semantic and visual features are combined together to represent the classes in a multi-modal manner. The experiments show that our proposed method outperforms state-of-the-art zero-shot learning methods. In near future, we will focus on the visual representations learning in more complex occasions. When maintaining the discrimination of classes, we will focus on the favorable visual features learning from complex scenes. In addition, we will also try to find a way to obtain more appropriate correlations between classes, and use a more efficient way to integrate the correlations of different aspects.

References

[1] G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu, "Text classification without negative examples revisit," *IEEE Transactions on Knowledge and Data Engineering*, 2006.

[2] H. Al-Mubaid and S. Umair, "A new text categorization technique using distributional clustering and learning logic," *IEEE Transactions on Knowledge and Data Engineering*, 2006.

[3] K. Sarinnapakorn and M. Kubat, "Combining subclassifiers in text categorization: A dst-based solution and a case study," *IEEE Transactions on Knowledge and Data Engineering*, 2007.

[4] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," *computer vision and pattern recognition*, 2013.

[5] Lampert, Nickisch, and Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," *computer vision and pattern recognition*, 2009.

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[7] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[8] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," *computer vision and pattern recognition*, 2016.

[9] B. Romera-Paredes and P. H. S. Torr, "An embarrassingly simple approach to zero-shot learning," *international conference on machine learning*, 2015.

[10] G. A. Miller, "Wordnet: a lexical database for english," *Communications of The ACM*, 1995.

[11] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," *computer vision and pattern recognition*, 2018.

[12] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," *computer vision and pattern recognition*, 2019.

[13] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv: Learning*, 2016.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *neural information processing systems*, 2014.

[15] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv: Learning*, 2015.

[16] S. M. Bulent and C. R. Gokberk, "Gradient matching generative networks for zero-shot learning," *IEEE Conference Proceedings*, 2019.

[17] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," *computer vision and pattern recognition*, 2017.

[18] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," *computer vision and pattern recognition*, 2018.

[19] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," *computer vision and pattern recognition*, 2016.

[20] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," *arXiv: Computer Vision and Pattern Recognition*, 2016.

[21] H. Zhang and P. Koniusz, "Zero-shot kernel learning," *computer vision and pattern recognition*, 2018.

[22] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *ICLR*, 2014.

[23] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," *computer vision and pattern recognition*, 2017.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *computer vision and pattern recognition*, 2009.

[25] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learningła comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[26] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," *computer vision and pattern recognition*, 2009.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv: Learning*, 2014.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, 2014.