2023 4th International Conference for Emerging Technology (INCET)
Belgaum, India. May 26-28, 2023

Enhanced Zero shot Learning using Deep Neural Network ResNet50

Ansari Shaista Khanam
EXTC Department
Rajiv Gandhi Institute of Technology
Mumbai, India
shaista.khan@vcet.edu.in

Dr.Poonam . N.Sonar
EXTC Department
Rajiv Gandhi Institute of Technology
Mumbai, India
poonam.sonar@mctrgit.ac.in

2023 4th International Conference for Emerging Technology (INCET) | 979-8-3503-3575-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/INCET57972.2023.10170025

***Abstract*— Computer vision has advanced with lot of development in visual recognition systems, which poses restrictions to expand for huge numbers of image classes. This restriction is due to huge image classes with unlabeled images which cannot be used in training the machine learning algorithm. As Traditional machine learning method of classification are based on the classification of categories which are available at the time of training. Technique of Zero-shot learning (ZSL) recognizes categories of test sets which is not appearing while training the model. The enhanced ZSL technique proposed is based on deep visual semantic embedding method. In this method Visual and semantic features are used for the classification of unknown categories. The extraction of visual features is accomplished with convolutional neural network (CNN) and ResNet 50. FastText is used to convert labels of classes into word embedding vectors. Visual and word embedding features are mapped. The model is predicting the Top 5 labels for an unknown image category (zero-shot class). The experiments are performed on standard datasets SUN and AWA2. The proposed technique of enhanced ZSL with ResNet 50 gives better accuracy and reduced model loss then CNN model.**

Keywords*— word embedding, CNN, ResNet50 Zero shot learning, visual semantic embedding, word embedding, CNN, ResNet50

I. INTRODUCTION

Conventional Classification approaches have accomplished extensive achievement in many areas [1]. Though, there are some constraints for these methods under this learning model. Sufficient labeled training examples are required for each class. The trained model can only categorize the examples of categories covered by the training data and cannot recognize unseen types. Nevertheless, it is not always possible to have labeled categories in practical scenarios. The lack of labeled data is due to large data sets which are expensive to annotate and due to rare data classes.

Zero-shot learning (ZSL) overcomes this problem by identifying unseen categories of the test set absent in the training set. ZSL is based on learning some transitional features from training data, and learned features are used to identify unknown categories during testing. To learn intermediate features, each class uses semantic features in a set of attributes, word embedding of class labels or sentences using natural language processing. Attributes are characteristics of categories that describe the features of the classes. Labels or text are used to describe the class semantically. This semantic information concatenated with Visual elements recognize unseen classes (zero-shot classes).

This paper proposed a method to identify unknown categories by extracting visual features with deep neural networks and semantic features (word embedding) using word2vec of training data. The model gives the top 5 predictions of the most suitable labels for unknown categories.

II. LITERATURE REVIEW

The main objective of ZSL is to discriminate unknown categories with the association between known and unknown categories. ZSL employs certain kinds of semantic information in the form of attributes and word vectors. Using semantic output code, ZSL was first proposed in [1] to predict the classes excluded in the training set. ZSL conquers the problem of classification of unlabeled data. In 2009 Lampert [2] et al. proposed attribute-based classification for unseen classes. Attributes are manually made features for groups of classes. Attributes like feather type, body structure, animal habitat etc., are used as auxiliary information for unseen categories that are not available in training. Attribute-based zero-shot image classification was introduced in [3]. This method uses direct attribute prediction (DAP) and indirect attribute prediction (IAP), which are probabilistic classifiers. Attribute label embedding [ALE] suggested in [4] works better than DAP. This mechanism introduces a compatibility function between the image and the label based on which model is evaluated. This method can work with any side information encoded in a vector. Frome [5] has developed deep visual semantic embedding (DeViSE), which extracts visual features with a convolutional neural network and semantic features using a skip-gram language model. The trained model is checked for its prediction using the Hinge loss function. Cross-model Transfer (CMT) [6] is a ZSL model based on a mixture of seen and unseen classes. This method is not based on attributes; it maps Visual features with semantic word vectors of comparable class. Zero-Shot Learning by the Convex Combination of Semantic Embeddings (ConSE) method [7] maps the image to its label embedding using the convex method. The model can be used without additional training.

Zero-shot classification by understanding the compatibility of Input embedding and output embedding is given in [8]. Input Embeddings used Fisher Vectors (FV) and Deep CNN Features (CNN). The author discovers combinations of five types of output embeddings: supervised

attributes, unsupervised Word2Vec, Glove, a bag of words (BoW) and

WordNet-derived similarity embeddings. Different embeddings are either concatenated or combined to improve the fine-grained classification. Ziming Zhang (SSE) in [9] proposed the classification of unknown categories by mixing each source/target data with the seen class. This method revealed significant improvement in accuracy on the benchmark dataset. The Deep Weighted Attribute Prediction method [10] uses a deep neural network for feature extraction and class prediction. And it does not require hand-crafted specific features. Models perform better with more accurate attribute prediction. SRC (Sparse representation coefficient) uses weighted attributes for prediction, improving classification accuracy.

The methods mentioned in [1], [2], [3], [4], [5],[6],[7],[8] [9] and [10] are embedding-based methods for zero-shot learning.

A novel method of ZSLis using the generative adversarial network in [11]. This method produces visual features using conditional GAN. Conditional GAN makes use of the semantic feature as a dependent parameter. Rui Gao has proposed an approach [12] combining conditional generative adversarial networks. Conditional variational autoencoder conditioned on semantic attributes. This model Works well even for Generalized ZSL settings. The framework of [13] is using a generative model which is trained using Generative adversarial network minimizing the domain shift problem of seen and unseen class distribution. This method is simpler than GAN/VAE, which outperforms the existing method. [14] is a Zero-Shot Learning method for incomplete semantic generation Which is Considered the most practical case leading to better ZSL performance. The hybrid Feature model [15] uses a conditional autoencoder conditioned on semantic space. This method uses two autoencoders; one encoder uses visual and semantic information, and another autoencoder is provided with only visual information.

The proposed method is an embedding-based method influenced by the [5] and [6], based on visual and semantic feature extraction. Visual features are extracted using a deep neural network of CNN and ResNet 50. Unknown categories are classified after mapping image features with a word embedding vector (semantic features).

III. ZERO SHOT LEARNING

ZSL is set of seen categories S= $\{C_i^s \mid i = 1,2, \ldots . . N_s\}$ where each C_i^s is seen category and unseen categories U= $\{C_i^u \mid i = 1,2, \ldots . . N_U\}$ where each C_i^u is an unseen category [16]. Whereas seen categories are completely disjoint of unseen categories i: e S∩U. ZSL can recognize categories which model has not seen while training. ZSL method is based on visual feature from image and semantic feature from the respective class label. The mapping associations are learnt between the visual feature space and semantic feature space to form the embedding space. Relationship between the image features and class labels using semantic is learnt. In testing stage learned mapping is used `to predict class labels of unknown classes. Based on unseen image features semantically similar labels are predicted. ZSL classifier learns by using (1)

$$C = U(S(.)) \quad (1)$$

$$S: X_d - S_v$$

$$U: S_v - Y_u$$

C maps the d dimensional input feature space (unseen) X_d to a set of Y labels from semantic space. C is a structure of two functions U and S. S learns from d dimensional seen classes feature space X_d to seen semantic vectorsS_v. Once learning of s is over, U learns the mapping of semantic vector S_v to unseen class labelsY_u. The mapping of Visual semantic is learnt using visual feature extraction, semantic representation.

A. *Visual features Extraction*

Extracting visual features is very important in computer vision. Effective methods should be used for extracting visual features such as colour, texture, shape etc. Deep learning models have shown great improvement over traditional methods of feature extraction. Currently, VGG, Google Net and RestNet etc. models are used.

B. *Semantic Representation*

ZSL is a method of recognizing unseen data based on knowledge gained from seen data. ZSL is trained with labelled seen data and tested on unseen classes which are not available at the time of training. The traditional method finds it difficult to transfer knowledge which is learned in training for the identification of unseen classes. To overcome this problem some secondary information is used to relate seen classes with unseen classes. This secondary information is called semantic feature space. The semantic features work as a link between seen and unseen categories which are attribute-based, or word vector based.

Attribute space. List of human comprehensible characteristics illustrating various properties of the classes are defined as attributes. Attributes are words or sentences describing one property of the classes. For example, animals' attributes are their properties such as their body color, their habitat, and their visual properties etc. Like zebras can be described by stripes, forest animals etc are attributes of zebras. These attributes are called semantic features which are common in some animals and hence can be used to identify unseen classes.

Word Vector Based. Natural language processing techniques are using word embedding vectors which are generated by Word2vec. Word2Vec is one of the most widely used methods to understand word embeddings using shallow neural networks. Most commonly used two basic architectures are continuous bag-of-words (CBOW) and skip-gram which are used for producing word2vec vectors [17].

In the CBOW method, the context of the sentence predicts the word in the middle of the sentence. CBOW is quite faster to train than the skip-gram and it gives better accuracy for the frequent words. In the skip-gram method,

Context is predicted based on the input word. This method represents rare words well and gives good accuracy with a small amount of data.

C. Visual Semantic Mapping

Visual features are mapped with semantic features. After developing visual semantic mapping in visual space, semantic space or intermediate space unseen classes are predicted. There are three ways of Visual semantic mapping based on it ZSL methods are classified.

Forward Mapping. Forward mapping is a type of mapping in which visual features are associated with semantic features, and the identification of unknown classes are achieved in the semantic space. [5] is using forward mapping where classification is performed in semantic space.

Common Mapping. Image features and semantic features are associated to common space. Identification of unseen class is accomplished in common space.[9] is an example of common mapping.

Reverse mapping. Reverse mapping is a type of mapping in which semantic features are mapped to visual space. Identification of unseen class is accomplished in visual space. [11], [12], [13], [14] are the cases of Reverse mapping.

IV. Proposed Method

Deep Visual semantic embedding is a method for Image classification of unseen data (zero shot classes). This method merges two models, the visual model, and the language model. This method is using forward mapping approach of visual semantic mapping where visual features are mapped to semantic space, Classification is performed in semantic space. The experiment results are tested on standard datasets SUN [19]and AWA2[20]. The SUN [19] dataset comprises 14,340 images of scenes images such as rivers, hills, churches, beaches etc. It consists of a split of 645 seen categories for training and 72 unseen classes for testing is used for zero-shot learning. The AwA-2 dataset [20] comprises 50 categories of 37322 images of animals. In this experiment for the AWA2 dataset, 40 categories are utilized for training and 10 categories are utilized for testing. The proposed method is consisting of a Visual model, language model and Deep visual semantic model.

A. Visual model

Visual model is used to extract visual features using deep Convolutional neural network. Various models are available in literature such as Alex Net, VGG19, Google Net, RestNet etc. In this method visual features of images are obtained using CNN model and ResNet 50. ResNet 50 model is pretrained on ImageNet dataset. Visual models are used to get input embeddings using CNN and ResNet 50.

Convolutional Neural Network (CNN). It is the very extensively utilized deep neural network. The basic operation used in convolution is a linear operation of matrices. CNN is always composed of the first layer as a convolutional layer which is used to extract various features of an image based on filters. The dimensionality of the network is lowered by the pooling layer.

The fully connected layer is used after pooling layer which connects all neurons of the prior layer. Various applications of CNN exist in computer vision for image classification, object detection etc. CNN also finds wide application in natural language processing [18].

The proposed CNN model for ZSL is shown in Figure.1 which is consisting of two convolutional layers followed by a Max pooling layer, two dense layers and a Relu activation function in the convolution layer. Drop out layer is used to avoid overfitting of model. CNN visual model gives a feature vector of 300 dimensional.

ResNet50. ResNet 50 is an architecture of Deep Convolutional neural networks which are used to extract low, mid, and high-level features from the images. A deep convolutional network is stacked with layers to get better accuracy, but it also leads to challenges of vanishing gradient and exploding gradient. This model converges slowly or starts oscillating. One of the solutions to reduce the challenge of vanishing gradient and exploding gradient is to use skip connection which is used in the Resnet 50 model.

ResNet 50 consists of several convolutional layers, a pooling layer, and a fully connected layer fusing to a total of 50 layers Models consist of several convolutional layers, max pooling etc to extract deep features of an image. The special feature of ResNet 50 architecture is Skip Connection. Skip connection skips some of the layers in the ResNet 50 architecture and provides the output of one layer as the input to the next layers. Skip connections overcome the problem of vanishing gradient to a great extent. The Proposed ResNet 50 for the ZSL model is shown in figure 2 and is pre-trained on the ImageNet dataset which is followed by two dense layers and drop out layer. The visual model with RestNet 50 also gives a feature vector of 300 dimensions.

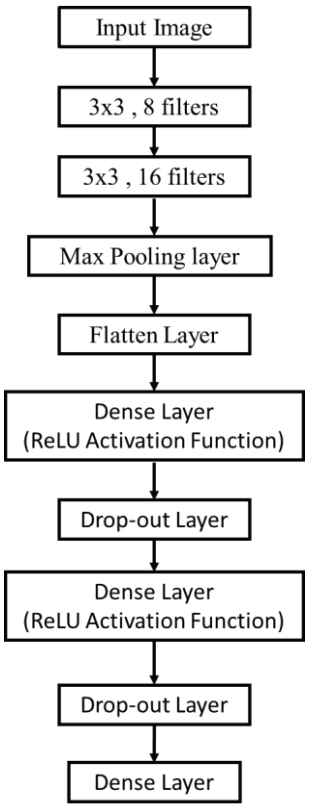


Fig. 1. Proposed CNN Model for ZSL

B. Language model

The language model is used to convert words into word vectors. The Wor2vec method is used for the language model. The word2vec algorithm uses a deep neural network model to understand word similarity based on a large

amount of text. Once a model is trained it can predict semantically similar words. This model represents each label by fixed length vectors called embedding vectors.

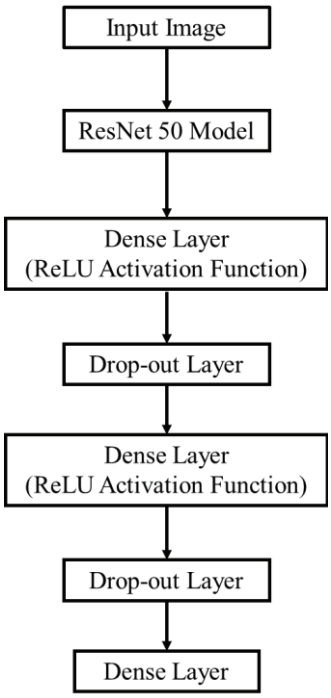


Fig. 2. Proposed ResNet50 Model for ZSL

As synonyms tend to occur in similar contexts, this leads the model to learn similar embedding vectors for semantically related words. Fig. 3 shows a visualization of the 300-dimensional semantic space with word vectors for some classes. The mapping from 300 to 2 dimensions is done with t-SNE plot. It shows semantically similar categories are near in the plot. For example, monkey and chimpanzee are semantically similar, so it is near in the plot.

The language model is implemented using Fast text an open-source library which is used to convert words into embeddings based on text data. Fast Text is created by Facebook's AI Research (FAIR) lab. Labels of Classes are converted into embedding vectors of 300 Dimensional. The model can generate feature vectors of 50, 100, 200 and 300 dimensions. A more dimension word vector gives more information hence for this experiment 300 dimension is selected.

C. *Deep visual semantic model*

Deep visual semantic model [1] combines visual model with language model. Deep characteristics are extracted using deep convolutional neural networks like ResNet 50. ZSL Training model is shown in Figure 3. (a) is taking two inputs which are input image and its label. Input image is given to visual model for feature extraction. Labels of the images are given to the language model for extracting its word embedding also called semantic features. Visual features and semantic features are embedded, and the model is trained. Model is ready for prediction of zero shot classes.

Zero shot image model first extract visual feature, predict its embedding vector as shown figure3. (b). Based on the Predicted embedding vector, the model searches for the top 5 nearest labels in the embedding space using cosine similarity measure.

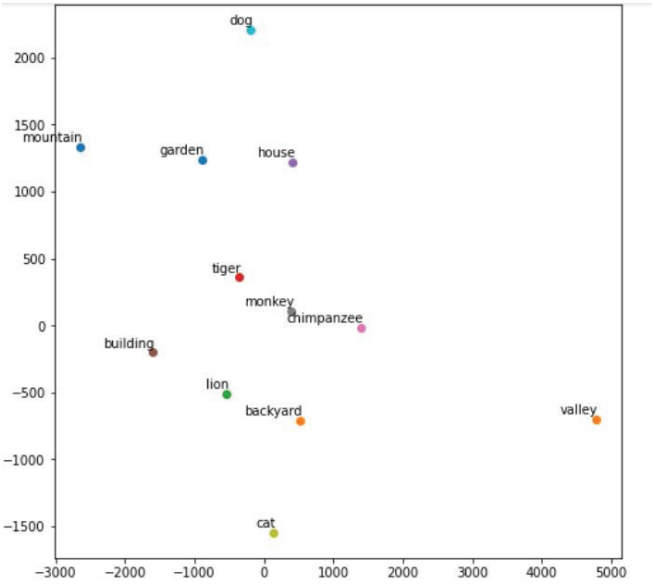
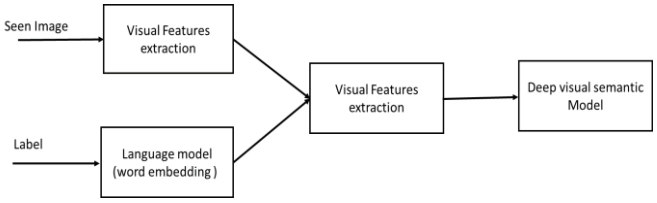


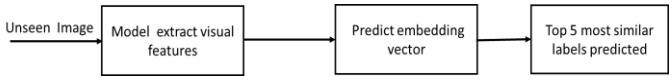
Fig. 3. tsne plot of wor2vec model

Visual model is trained with convolutional neural networks and ResNet 50. Embedding vectors of 300 dimensions are calculated using a Fast Text. 300 dimensions are selected which give more features of embedding vectors.

Visual features and language models are mapped with 300-dimensional representation. The model is trained for 20 epochs with early stopping. Early stopping is used to avoid overfitting in training of model. During testing, for test image from zero shot class, model first computes visual feature vector using visual model, then search for nearest labels in embedding vector space using cosine similarity measure. Top5 nearest embedding vectors are displayed as result.



(a) Training Model



(b) Prediction

Fig. 4. Proposed deep visual semantic model for ZSL

V. RESULT

The enhanced ZSL using the deep visual semantic model is developed by building a visual model using CNN and ResNet 50 and the language model is developed using FastText. The implemented model can classify unknown image categories. The proposed model is tested on the standard datasets SUN [19] and AWA2 [20]. Once a model is tested for unknown categories (zero-shot classes), the performance of the designed Model is evaluated by calculating per class average accuracy given by the following equation

$$\text{Per class Average Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_{correct\ class}^{class\_i}}{Y_{Total}^{class\_i}}\right] \quad (2)$$

Per class average accuracy is calculated by the average of correct classes identified divided by the total number of classes in a dataset. For the proposed Enhanced ZSL method per class average accuracy and model loss is calculated. Results are shown in Table I. Experimental results shows that Model loss is less for the ResNet 50 model as compared to the CNN model. It is evident from the results that per class average accuracy is better for ResNet 50 model as compared to the CNN model.

TABLE I. RESULT TABLE FOR PER CLASS AVERAGE ACCURACY AND MODEL LOSS

Model		**Per class average Accuracy and model loss**		
	Dataset	***Methods***	***Accuracy***	***Model loss***
CNN	SUN	Proposed Enhanced ZSL	13.09	-0.65
	AWA2	Proposed Enhanced ZSL	11.25	-0.74
ResNet 50	SUN	IAP [15]	19.4	-
		ConSE [15]	38.8	-
		Proposed Enhanced ZSL	**39.5**	-0.70
	AWA2	IAP [15]	35.9	-
		ConSE [15]	44.5	-
		CMT [15]	37.9	-
		Proposed Enhanced ZSL	**44.786**	-0.9

Top 5 prediction Result for unknown categories for some sample cases of SUN and AWA2 datasets are shown in figure 5 (a), (b) and figure 6(a), (b). Actual labels of the images along with Top 5 predictions of labels are shown.

Top 5 predictions of the images are based on cosine similarity measure. In prediction accurate categories are predicted whereas in some of the cases very similar labels are predicted. The Result of Proposed approach of Enhanced ZSL shows that ResNet 50 can predict actual category labels more precisely as compared to CNN model. In sample case of Chimpanzee, ResNet 50 gives prediction of chimpanzee at top 4 place whereas CNN gives at top 5 place hence, ResNet 50 gives improved result in terms of Top n prediction then CNN.

(a) Actual label: lawn

ResNet 50:
Top 5 Prediction for an image: garden, back-garden, bankside, polytunnel, rockery
CNN:
Top 5 Prediction for an image: back-garden, cark, house, bankside, garden

(b) Actual label: Valley

ResNet 50:
Top 5 Prediction for an image: mountain, **valley**, hill, hillside, cliff-side
CNN:
Top 5 Prediction for an image: bankside, cliff-side, cark, landside, cross-island

Fig. 5. Result for SUN dataset

(a) Actual label: chimpanzee

ResNet 50:
Top 5 Prediction for an image: gorilla, gorillas, chimp, **chimpanzee**, orangutan
CNN:
Top 5 Prediction for an image: gorilla, gorillas, chimp, monkey, **chimpanzee**

(b) Actual label: pig

ResNet 50:
Top 5 Prediction for an image: cow, cows, goat, **pig**, heifer
CNN:
Top 5 Prediction for an image: horse, giraffe, elephant, cow, deer

Fig. 6. Result for AWA2 dataset

Experimental results of proposed visual semantic embedding system with ResNet 50 model for SUN dataset [19] gives improved accuracy as compared with IAP and CONSE method [15]. For AWA2 dataset [20] proposed method with ResNet 50 model gives better accuracy as compared to IAP, ConSE and CMT method [15].

VI. CONCLUSION

ZSL is a technique used to classify the unknown categories which are not the part of training set. The enhanced ZSL proposed is consisting of visual feature extraction, semantic feature illustration in a form of embedding vector, visual semantic mapping, and classification of unknown categories. The enhanced ZSL using deep neural network is implemented using CNN, ResNet 50 and fastText model on SUN dataset and AWA2 dataset. Model is predicting the embedding vectors for unknown image category and searches for top5 similar labels using cosine similarity measure. The experiment results of Proposed Enhanced ZSL shows that, for SUN dataset and AWA2 dataset ResNet50 model gives accuracy of 39.5%, 44.786% which is better as

compared with CNN. It is also evident from result that model loss is less for ResNet 50 as compared to CNN.

Proposed Model is extracting many features and calculating plenty of parameters of an images so increasing the complexity of the model in prediction. Results of the proposed model can be improvised in future by selecting significant features. Uncorrelated and significant feature selecting techniques such as autoencoder can enhance accuracy of the proposed visual semantic model for ZSL. These techniques can reduce time and space complexity of the model.

REFERENCES

[1] Palatucci M, Pomerleau D, Hinton G, Mitchell T (2009) "Zero-shot learning with semantic output codes", Adv Neural Inf Proces Syst 1:1410–1418

[2] LAMPERT C, Nickisch H, HARMELING S (2009) "Learning to detect unseen object classes by between-class attribute transfer", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 20–25 June, pp951–958

[3] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling."Attribute-based classification for zero-shot visual object categorization", IEEE T. Pattern Analysis and Machine Intelligence, 36(3):453–465, 2014.

[4] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. "Label-embedding for image classification". IEEE TPAMI, 38(7):1425–1438, 2015.

[5] 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 Advances in neural information processing systems, pages 2121–2129, 2013.

[6] Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A. Zero-shot learning through cross-modal transfer. Adv. Neural Inf. Process. Syst. 2013, 26, 1–10.

[7] Norouzi M, Mikolov T, Bengio S, Singer Y, Shlens J, Frome A,Corrado G, Dean J (2013) Zero-shot learning by convex combination of semantic embeddings. arXiv 2013, arXiv:1312.5650

[8] Akata, Z. et al. (2015) "Evaluation of output embeddings for fine-grained image classification," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Preprint]. Available at: https://doi.org/10.1109/cvpr.2015.7298911.

[9] Ziming Zhang and Venkatesh Saligrama. "Zero-shot learning via semantic similarity embedding," In Proceedings of the IEEE international conference on computer vision, pages 4166–4174, 2015

[10] Xuesong Wang, Chen Chen, Yuhu Cheng , Xun Chen and Yu Liu "Zero-Shot Learning Based on Deep Weighted Attribute Prediction" IEEE Transactions on Systems, Man, and Cybernetics: Systems (Volume: 50, Issue: 8, Aug. 2020)

[11] Yongqin Xian, Tobias Lorenz, Bernt Schiele, Zeynep Akata "Feature Generating Networks for Zero-Shot Learning", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5542-5551

[12] Gao, R., Hou, X., Qin, J., Liu, L., Zhu, F., Zhang, Z. (2019). A Joint Generative Model for Zero-Shot Learning. In: Leal-Taixé, L., Roth, S. (eds) Computer Vision – ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science (), vol 11132. Springer, Cham. https://doi.org/10.1007/978-3-030-11018-5_50

[13] Varun Khare, Divyat Mahajan, Homanga Bharadhwaj, Vinay Verma, Piyush Rai, "Generative Framework for ZSLwith Adversarial Domain Adaptation", in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV).

[14] Xiaojie Zhao, Yuming Shen, Shidong Wang, Haofeng Zhang "Boosting Generative Zero-Shot Learning by Synthesizing Diverse Features with Attribute Augmentation",in Computer Vision and Pattern Recognition.

[15] Al Machot, F.; Ullah, M.; Ullah, H. HFM: A Hybrid Feature Model Based on Conditional Auto Encoders for Zero-Shot Learning. J. Imaging 2022, 8, 171. https://doi.org/10.3390/jimaging8060171

[16] Wang, W. et al. (2019) "A survey of Zero-shot learning," ACM Transactions on Intelligent Systems and Technology, 10(2), pp. 1–37. Available at: https://doi.org/10.1145/3293318.

[17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, 2017.

[18] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[19] Sun Attribute Database: Discovering, annotating, and recognizing scene attributes (no date) SUN Attribute Dataset. Available at: https://cs.brown.edu/~gmpatter/sunattributes.html (Accessed: December 20, 2022).

[20] Available at: https://cvml.ist.ac.at/AwA2/ (Accessed: December 20, 2022).