International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Zero-Shot Learning to Detect Object Instances from Unknown Image Sources

Chowdhury Shahriar Muzammel, Partha Chakraborty, Md. Noweshed Akram, Khalil Ahammad, Md. Mohibullah

Abstract*: Inspired by the human capability, zero-shot learning research has been approaches to detect object instances from unknown sources. Human brains are capable of making decisions for any unknown object from a given attributes. They can make relation between the unknown and unseen object just by having the description of them. If human brain is given enough attributes, they can assess about the object. Zero-shot learning aims to reach this capability of human brain. First, we consider a machine to detect unknown object with training examples. Zero-shot learning approaches to do this type of object detection where there are no training examples. Through the process, a machine can detect object instances from images without any training examples. In this paper, we develop a dynamic system which will be able to detect object instances from an image that it never seen before. Which means during the testing process the test image will completely unknown from trained images. The system will be able to detect completely unseen objects from some bounded region of given images using zero shot learning approach. We approach to detect object instances from unknown class, because there are lots of growing category in the world and the new categories are always emerging. It is not possible to limit objects in this fast-forwarding world. Again, collecting, annotating and training each category is impossible. So, zero-shot learning will reduce the complexity to detect unknown objects.

***Keywords*: Zero-shot learning, Region Proposal Network, Object recognition.**

I. INTRODUCTION

Human brain is capable of recognize an unknown object, they never seen before just by having a description of the related object. It has been estimated that humans mostly can recognize between 5000 and 30000 categories of object [1] [2]. Humans can learn completely unknown classes from a high-level description without having any previous training data.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Chowdhury Shahriar Muzammel, Department of CSE, Comilla University, Cumilla - 3506, Bangladesh. Email: anik.com@gmail.com
Partha Chakraborty*, Department of CSE, Comilla University, Cumilla - 3506, Bangladesh. Email: partha.chak@cou.ac.bd
Md. Noweshed Akram, Department of CSE, BAIUST, Cumilla Cantonment, Cumilla - 3506, Bangladesh. Email: noweshed@gmail.com
Khalil Ahammad, Department of CSE, Comilla University, Cumilla - 3506, Bangladesh. Email: khalil.cou.cse@gmail.com
Mohammad Mohibullah, Department of CSE, Comilla University, Cumilla - 3506, Bangladesh. Email: mohib.cse.bd@gmail.com

For example, if we ask someone to find a person whom he never seen before and give him some characteristics of that person then it is possible for him to recognize the unknown person from any kind of domain. In such, Zero-shot learning (ZSL) approach is meant to recognize an unknown object without any prior training image.

ZSL can be treated as a special case of unsupervised domain adaptation, where the source domain and the target domain have completely disjoint label spaces [3]. It is basically a two-stage process, which are training stage and inference stage. Knowledge about the attributes is captured in the training phase, and this knowledge is used to categorize instances among a new set of classes in the inference phase [4]. The existing object detection models require training data for all classes which are not optimal because the number of object class is growing and there are huge number of living spaces. So, for this large-scale data we need to train for each object class individually otherwise it couldn't detect outside the box. Again, it is time consuming and arises space complexity for large data-set.

To overcome the problem, we proposed a model where you don't need to train your machine every time with visual image for new object class, instead set the attributes for this new class. The ZSL formula is generally been realized by using the attributes to eliminate the semantic gap between human class descriptions and low-level features [5]. The system will detect the object from test image, compare the attributes using semantic level vectors and recognize the unknown object.

II. RELATED WORK

Zero-shot Learning: Humans can recognize an object by relating known objects, without prior visual experience. Simulating this behavior into an automated machine vision system is called Zero-shot learning (ZSL). ZSL attempts to recognize unseen objects without any prior examples of the unseen data.

Zero-shot Object Recognition: Zero-shot learning recognize unknown object in a single dominant. But, As a part of complex scene, unknown object appears only. Shafin et al [6] showed that, how Zero-shot localize and recognize each instance of new object classes individually from a complex domain and also can recognize without having any prior visual examples of those classes during the training stage.

Semantic Embedding: In natural language processing, Word embedding is the process to represent a set of word modeling and feature learning techniques. This process maps word into vectors of real number.

It also involves a mathematical embedding of one dimension per word from a space to a much lower dimension having continuous vector space. Semantic embedding presents the similarity between the words according to their vector representation.

Shujon et al [7] established a relationship between known and unknown class using semantic label vectors.

III. METHODOLOGY

The flowchart of our proposed methodological steps are given below:

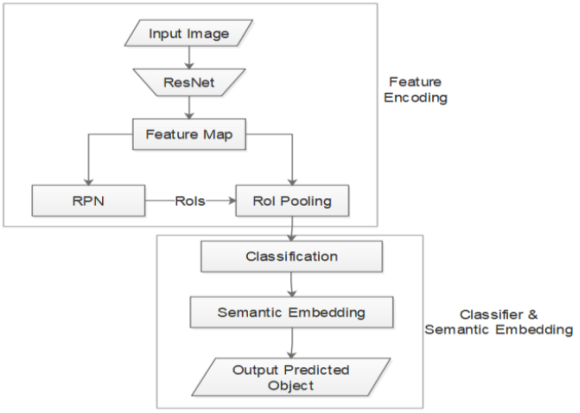

Fig. 1. Flowchart of methodology

We illustrate all the methods in step by step process as follows:

Step 1: Input Image: Collect image from ImageNet ILSVRC-2017.

Step 2: Resnet: A residual neural network (ResNet) is used to obtain intermediate convolution activation.

Step 3: Feature Map: Feature map is used to obtain object bounding box and their features.

Step 4: RPN: Region proposal network or RPN generates object proposal by automatically ranking.

Step 5: ROI Pooling: This layer operates the starting feature map.

Step 6: Classification: Classification is trained to predict each object category and their class.

Step 7: Semantic embedding: Semantic embedding is applied to perform zero-shot object detection.

Step 8: Output image: Finally, we get the output result of unknown object class with their bounding box.it.

IV. EXPERIMENTAL DETAILS

The implementation of ZSL is a two-step process. The first process is training and second inference. We experiment on ImageNet ILSVRC-2017 for our proposed model for dataset, which has 200 object categories. The dataset has 456,567 images and for training purpose the dataset has 478,807 bounding box annotations around object instances. 200 object categories with 20,121 fully annotated images, which totally includes 55,502 object instances is used for validation purpose. For some objects having multiple parents, there has been defined hierarchy. We define a single parent for each category because, we evaluate our approach on meta-class detection and tagging.

First, we train Faster R-CNN for only seen classes using the training set. RPN or Region Proposal Network involves pre-trained model ResNet for shared layers. Then, we applied semantic network instead of Faster R-CNN classification layer. While we train our model, we trained shared layer and RPN is fix because object proposal remains similar from previous step. For each input image, RPN consist of total 'R' RoIs in both negative object proposal and positive too. Any of the known class are positive proposals and negative ones contains background. In RoI pooling layer a feature branch is generated for each RoI which forwarded to the regression branch and classification as well. The overall loss is calculated by the summation of classification and regression loss as follows:

$$L_{cls}(\mathbf{o}_i, y_i) = \lambda L_{mm}(\mathbf{o}_i, y_i) + (1-\lambda) L_{mc}(\mathbf{o}_i, g(y_i)),$$

Classification Loss: Classification loss is calculated for both seen and unseen class.

$$L_{mm}(\mathbf{o}_i, y_i) = \frac{1}{|\mathcal{C}' \setminus y_i|} \sum_{c \in \mathcal{C}' \setminus y_i} \log\Big(1 + \exp(o_c - o_{y_i})\Big),$$

Where, lambda is a parameter which works as a bridge and tries to trade-off between the two losses.

Here, o represents the prediction class. For separating true class prediction response L_m_m is used and for clustering together the member of each meta class L_m_c is used. Regression loss is fine tuned for each seen class. The calculation is based on the ground truth co-ordinates and the co-ordinates. In the semantic space, high level supervision is being added by the meta-class assignment. For this experiment, we have taken into consider both seen and unseen classes. With the same approach, the maximum loss margin considers the set C consisting of both seen and unseen classes. For ZSL, to address domain adaptation this problem setting helps to identify the clustering structure of the semantic embeddings. To solve the problem, we use seen word vectors only as fixed embedding to train the whole framework with max-margin loss,

$$L'_{mm}(\mathbf{o}_i, y_i) = \frac{1}{|\mathcal{S}' \setminus y_i|} \sum_{c \in \mathcal{S}' \setminus y_i} \log\Big(1 + \exp(o_c - o_{y_i})\Big)$$

Prediction: To calculate each output prediction value we normalize each prediction. Between word vectors and image features, it calculates the cosine similarity.

V. SAMPLE INPUT OUTPUT

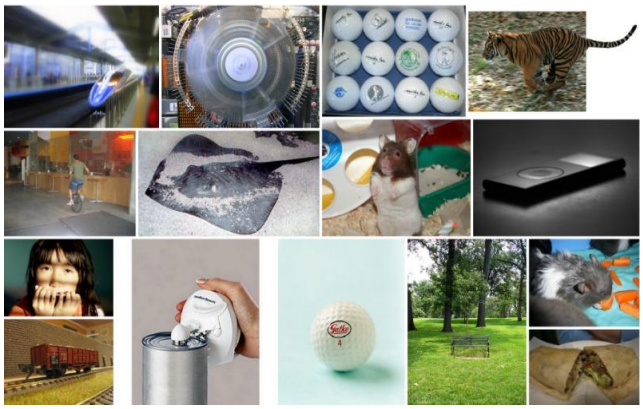

Fig. 2. Selected Input image for Zero-shot learning detection

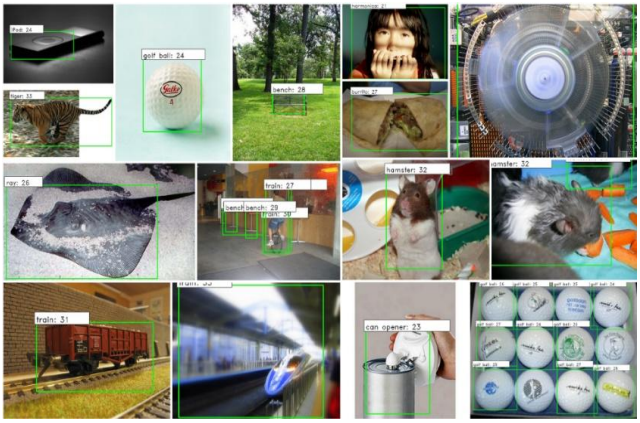

Fig. 3. Output of the selected image with their bounding box

VI. RESULT ANALYSIS

We compare our result along with the baseline according to loss configuration L'_m_m and L_c_l_s. The loss deals with both known and unknown class. L'_m_m denotes the max-margin loss and L'_m_c denotes meta-class clustering loss.

Table- I: Result analysis

	Average	Syringe	Harmonica	Burrito	Can opener
	Similar class not present				
Base	10.85	3.9	0.5	36.3	2.7
L'_{mm}	13.05	8.0	0.2	39.2	4.8
L_{cls}	7.78	1.0	0.1	27.8	2.2

Average precision of individual unknown classes using ResNet & w2v loss calculation L'_m_m and L_c_l_s when similar classes aren't present.

Average precision of individual unknown classes using ResNet & w2v loss calculation L'_m_m and L_c_l_s when similar classes are present.

As we can see, across all zero-shot tasks performance get improved from baseline to ours (with L'_m_m). The reason behind the improvement is, during the training the baseline method did not consider word vectors. As a result, about the semantic embeddings of classes, overall detection could not get enough supervision. For this situation, our desired L'_m_m loss formulation considers word vector for better outcome.

Table- II: Result analysis

	Average	Bench	Electric Fan	iPod	Hamster	Tiger	Ray	Train	Golf Ball
Similar class not present									
Base	22.56	1.0	0.6	22.0	40.9	75.3	0.3	28.4	12.0
L'_{mm}	28.74	0.0	7.1	23.3	50.5	75.3	0.0	44.8	28.9
L_{cls}	31.71	4.1	5.3	26.7	47.3	71.3	21.5	51.1	26.2

VII. CONCLUSION

In supervised learning we detect objects from trained image only. But, zero-shot learning detect object instances from a completely unknown image. In our work, we have developed a system by using zero-shot learning, where we are able to detect objects from completely unknown class. We used ILSVRC-2017 dataset for both training and testing of this paper. We showed that, zero-shot learning is capable of detect object instances from unknown examples while other supervised learning can't where we got better result from baseline results.

Zero-shot learning further warrants limitation of training the attributes. We expect to detect object which has visually and semantically similar to a corresponding class. While doing this, we see that the system detect object which is semantically and visually similar but of different object. To get better results to avoid this circumstance we will experiment by using word vectors in future.

REFERENCES

1. Wai Lam Hoo and Chee Seng Chan. Zero-Shot Object Recognition System Based on Topic Model. IEEE Transactions on Human-Machine Systems, Vol. 45, No. 4, August 2015.
2. Yanwei Fu, Timothy M. Hospedales, Tao Xiang and Shaogang Gong. Transductive Multi-view Zero-Shot Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015
3. Meng Ye and Yuhong Guo. Zero-Shot Classification with Discriminative Semantic Representation Learning. IEEE Conference on Computer Vision and Pattern Recognition, 2017
4. Xun Xu, Timothy Hospedales and Shaogang Gong. Semantic Embedding Space for Zero-Shot Action Recognition. IEEE Conference, 2015.
5. Ioannis Alexiou, Tao Xiang and Shaogang Gong. Exploring Synonyms as Context in Zero-Shot Action Recognition. IEEE Conference, 2016.
6. Shafin Rahman, Salman Khan and Fatih Porikli. Zero-Shot Object Detection: Learning to Simultaneously Recognize and Localize Novel Concepts. Asian Conference on Computer Vision, 2018.
7. Shujon Naha and Yang Wang. Zero-Shot Object Recognition Using Semantic Label Vectors. IEEE 12th Conference on Computer and Robot Vision, 2015.

AUTHORS PROFILE

Chowdhury Shahriar Muzammel, a faculty member of Comilla University, Bangladesh, Faculty of Engineering, is currently working as a lecturer in the Department of Computer Science and Engineering, Comilla University. He was awarded M.Sc. (Engineering) degree and B.Sc. (Engineering) degree from Comilla University in Computer Science and Engineering. Signal processing, Bangla Natural Language Processing, Image Processing, and Artificial Neural Network are his current research interests. In various international journals, he presented his research articles.

Partha Chakraborty is a member of the faculty at the Department of Computer Science and Engineering, Comilla University, Bangladesh is currently working as an assistant professor. He received M.Sc and B.Sc. in the field of computer science and engineering from the Jahangirnagar University, a renowned Public University of Bangladesh. He also has an interest in computer vision, robotics, image processing and artificial intelligence for his work. To various international journals, he presented his research articles. He is actively involved in the fields of education and learning.

Md. Noweshed Akram completed Bachelor of Science (B.Sc.) in Computer Science & Engineering from Bangladesh Army International University of Science & Technology (BAIUST), Cumilla. His research area includes Machine Learning, Artificial Intelligence, and Natural Language Processing. He have completed his undergraduate research work on Zero-shot Learning where he worked to detect object instances from unknown image sources from their attributes. Currently, He is working as a software engineer where he research and develop new technologies and API's to integrate them with existing software in order to enhance user experience.

Khalil Ahammad is a faculty member of the Faculty of Engineering, Comilla University, Kotbari, Cumilla-3506, Bangladesh. He is currently working as a lecturer in the department of Computer Science and Engineering. He has completed his Bachelor of Science (Engg.) and Master of Science (Engg.) degree from the department of Computer Science and Engineering of Comilla University. His current research interest includes Machine Learning, Knowledge Engineering, Artificial Intelligence, Pattern Recognition and Bengali NLP. He is an active participant of various national conferences and he has already published his several research articles in several international journals. He is actively engaged with different research communities both in national and international arena.

Md. Mohibullah, a faculty member of the Faculty of Engineering, Comilla University, Bangladesh is currently working as a Lecturer at the department of Computer Science and Engineering since 27 February 2019. He secured first position (Thesis Group) in Master of Science (Engineering) degree and obtained second position in Bachelor of Science (Engineering) degree in Computer Science and Engineering (CSE) from Comilla University, a state-run university in the south-east region of Bangladesh. His current research interest includes Machine Learning, Data Mining, Artificial Intelligence and Big Data.