Aligarh Muslim University

Zakir Husain College of Engineering and Technology

Colloquium Lab Report

Zero Shot Learning and Its Application

April 29, 2019

Authors

Mansi Agarwal
mansi.29ag@gmail.com
16PEB055
GI2058

Tezuesh Varshney
tezuesh.varshney@gmail.com
16PEB123
GH4766

Contents

1 Introduction

1.1 Motivation

The motivation of this assessment was to explore the state-of-the-art neural network frameworks that reaches human level cognition at tasks such as Image Classification and Language Processing. The group members are both interested in learning about how machines would learn when insufficient or no examples are available. The project deals with the resource-bound reasoning methods that would help generalizing the neural network to diverse yet dense domain.

Our motivation was our interest in the implications of uncertainty and limited computational resources on the design of state-of-the-art techniques for the real world problems that are not feasible or desirable. AI is a field full of nuances and is used to solve some of the hardest problems in the outside world, and we therefore could not attempt exhaustive research about the subject, but were motivated to get started and to study theoretical issues and their approach towards development of effective algorithms and its application.

Finally, the greatest motivation was to learn new and interesting things about how machines can teach itself to adapt different domains, as well as get experience with a medium-term seminar presentation, and therefore also develop our interpersonal skills.

1.2 Objective

Objective is divided into two parts both focusing on to study the approaches how machines can teach itself when sufficient example training set are not available.

- Getting familiarized with the concept of Zero-Shot Learning framework and learning what, why's and how's of the framework.
- Discuss applications, ranging from language processing to image classification, where Zero Shot Learning can be implied
- To discuss the current on-goings and future work in the domain of Zero Shot Learning.

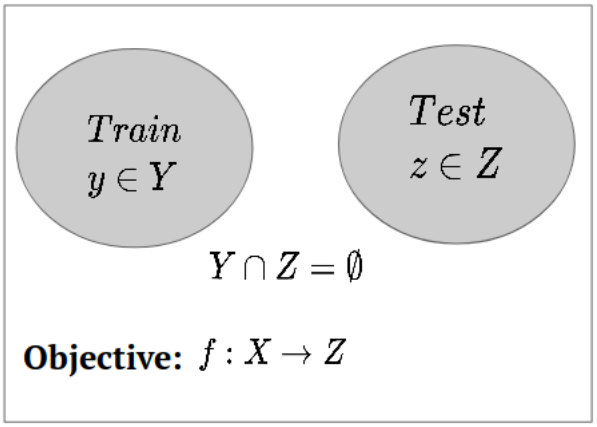
2 Literature Review

2.1 Overview

The concept of Zero Shot learning (ZSL) is inspired by humans ability to recognize an unseen class when are provided with description of the class. This ability of humans is because of their existing knowledge base which provides high-level description of a new or unseen class.

In case of machines, Zero Shot machine learning technique is used to construct recognition models for unseen target class that are not labelled during training time.

Definition 1 (Zero-Shot Learning). Given labeled training instances D^{tr} belonging to the seen classes Y, zero-shot learning aims to learn a classifier $f^u(\Delta) : X \rightarrow Z$ that can classify testing instances X^{te} (i.e., to predict Z^{te}) belonging to the unseen classes Z.



In case of machines, Zero Shot machine learning technique is used to construct recognition models for unseen target class that are not labelled during training time. It utilises the class attributes as aside information and transfers information from source classes with labelled samples. ZSL is done in two stages:

- **Training** : Where the knowledge about the attributes is captured
- **Inference** : The knowledge is then used to categories instances among a new set of classes.

Recently, there has been a surge in interest in automatic recognition of attributes, due to the availability of data containing meta information. Zero-shot learning approaches are designed to learn intermediate semantic layer, their attributes, and apply them at inference time to predict a new class of data.

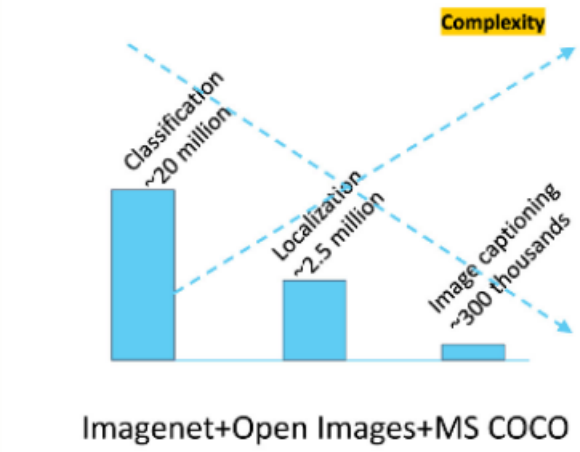
2.2 Why Zero Shot Learning?

The crux of zero-shot learning also relies on the existence of a labelled training set of seen classes and unseen class. Both seen and unseen classes are related in a high dimensional

vector space, called semantic space, where the knowledge from seen classes can be transferred to unseen classes.

- *Uneven distribution of data*: This is because the distribution of no. of images is not uniform for all classes. For example: In Imagenet dataset the common objects such as cars have a large number of images but there are only few images for uncommon objects such as insects. It is vital to use zero shot learning for these problems.

- *Target classes change over time*: An example is recognizing images of products belonging to a certain style and brand. As products of new styles and new brands appear frequently, for some new products, it is difficult to find corresponding labeled instances.



- *Annotation and Complexity*: In some particular tasks, it is expensive to obtain labeled instances. In some learning tasks related with classification, the instance labeling process is expensive and time consuming. Thus, the number of classes covered by existing datasets is limited, and many classes have no labeled instances.

Since its inception, ZSL has become a fast-developing field in machine learning, with a wide range of applications in computer vision, natural language processing, and ubiquitous computing.

2.3 Framework

From the *Definition 1*, we can see that the general idea of zero-shot learning is to transfer the knowledge contained in the training instances to the task of testing instance classification. The label spaces covered by the training and the testing instances are disjoint. Thus, zero-shot learning is a subfield of *transfer learning*

The general idea of zero-shot learning is to transfer the knowledge contained in the training instances to the task of testing instance classification.

**Here we will be discussing about ZSL framework of Image classification which can be used to impose on other application.*

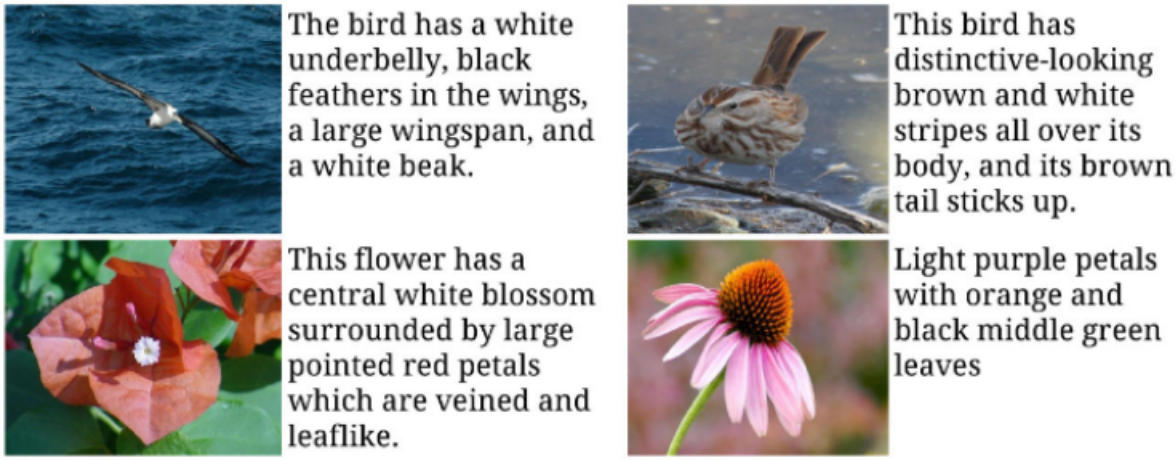
Requirements

- Side Information
- Model

2.3.1 Side Information

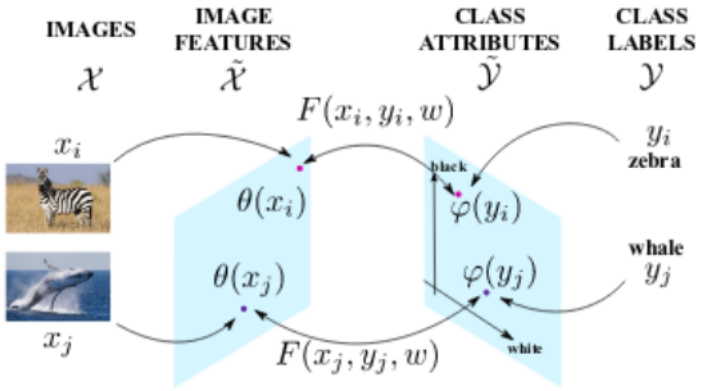
Side Information, extra information, domain knowledge are all similar terms. Side information is data that is neither in input nor in the output but include useful information to learning the input. Given the case of image classification side information can be in the form of:

- *Attributes*: Attributes are the visual qualities of objects , such as 'red', 'striped', or 'spotted' i.e. they are the key features of the image. For example: Zebra has black and white strips, a land animal and has a tail.
- *Sources*: Wikipedia and wordnet provides high-level description of the object and we can extract the information from this text and represent them as a vector using word embedding techniques such as word2vec, GLove and Hierarichal similarity measures.
- *Visual Abstraction*: Side information can also be provided by human as detailed visual description.



2.3.2 Model

Once we have the side information, we process this side information in a vector form which can be fed into Zero Shot Learning framework. We extract the features from side information using Neural Network and represent the information in $n-$dimensional vector space. the auxiliary information involved by zero-shot learning methods is usually some semantic information. It forms a space that contains both the seen and the unseen classes. As this space contains semantic information, it is often referred to as the semantic space. Being similar to the feature space, the semantic space is also usually a real number space. In the semantic space, each class has a corresponding vector representation, which is referred to as the class prototype of this class.



Simultaneously we try to extract features from the images using another Neural Network which are represented in other $m-$dimensional vector space. Once we have the features of Side information and Image we try to find the compatibility between Image Features, Side Information and Class Labels using a third neural network.

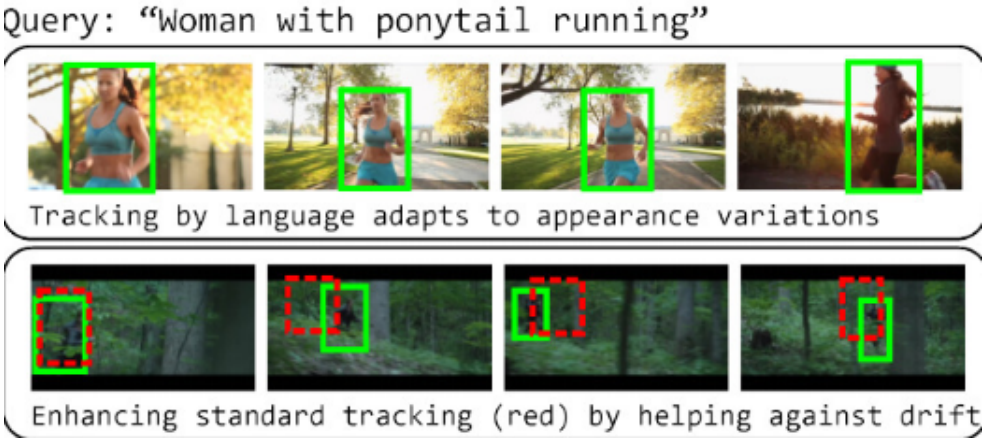
Benchmark Models Existing ZSL models can be grouped on the basis of compatibility:

- **Linear Compatibility**: ALE, DEVISE, SJE, ESZSL, SAE
- **Non-Linear Compatibility**: LATEM, CMT
- **Two-Stage Inference**: DAP, CONSE
- **Hybrid Model**: SYNC

3 Applications

3.1 Localization in Video

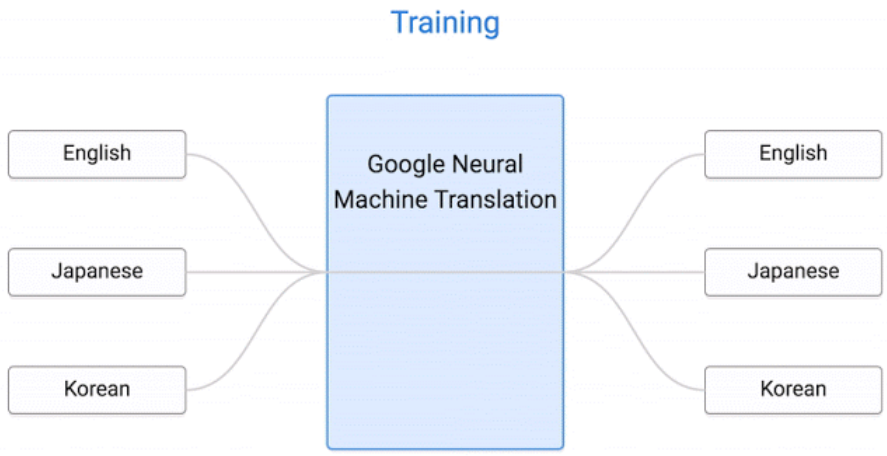
Tracking is a long standing challenge in computer vision. The common approach is to specify a target by means of a bounding box around the object and to track this target as it moves throughout the video. One of the zero shot approach to object tracking in video is to localize an object in an image by means of a natural language query only, returning a bounding box. Such tracking by natural language specification allows for novel type of human-machine interaction in tracking. In several real-life applications, such as robotics or autonomous driving, defining the target by a description is more natural,e.g., "track the red car in the middle lane". It also allows for novel applications of tracking, because there is no first frame requirement when the target is defined by text so one can start tracking in the middle of the video whereas in traditional deep tracking, you need to have a first frame where the object appears and then we have to draw boxes to start tracking. It also allows multiple video-multiple target tracking at the same time. Moreover, it allows for robust kind of tracking as the model adapts for appearance variation as there can be many objects with little appearance variation which can be tracked by the same natural language statement.



3.2 Neural Machine Translation

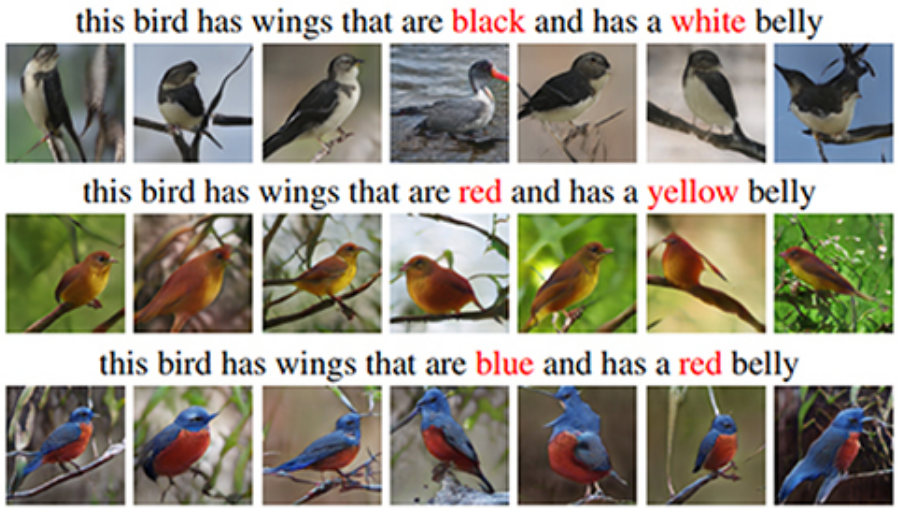
Translation refers to the process of translating words or text from one language into another. Translation has enabled effective communication between people around the world. Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. Neural Machine Translation (NMT) is an end-to-end learning approach for automated translation,with the potential to overcome many of the weaknesses of conventional phrase-based translation systems. Neural Machine Translation is promising approach with the potential of addressing many shortcomings of traditional machine translation systems. The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text. Its architecture typically consists of two recurrent neural networks (RNNs), one (encoder network) to consume the input text sequence and one (decoder network) to generate translated output text. But such a system requires a new model for each language pair, and scaling up to all the 103 supported languages (Languages in Google Translate) is a significant challenge. To address this challenge, Google extending its previous GNMT system,

allowing for a single system to translate between multiple languages. The proposed architecture requires no change in the base GNMT system, but instead uses an additional "token" at the beginning of the input sentence to specify the required target.



3.3 Image Generation using GANs

Generative Adversarial Networks abbreviated as GAN is a machine learning system, where two neural networks battles each other to create image that are superficially authentic to humans. Given the demonstrated capability of Generative Adversarial Networks(GANs) to generate images, these networks can be used to generate images from text. GANs can be leveraged to imagine unseen categories from text descriptions and hence recognize novel classes with no examples being seen. Specifically, the model takes as input noisy text description about an unseen class (eg wikipedia articles) and generate synthesized visual features for this class. With added pseudo data, zero-shot learning is naturally converted to a traditional classification problem. From the given text description of the GAN hallucinates the psuedo visual features for corresponding image class



3.4 Others

Besides the above areas, zero-shot learning has also been used for applications in other areas.

- In the area of ubiquitous computing, it is used for human activity recognition from sensor data.
- In the area of computational biology, it is used for molecular compound analysis, neural decoding from fMRI images, and ECoG.
- In the area of security and privacy, it is used for new transmitter recognition
- Other problems like knowledge representation learning, classifying unseen tags in StackExchange, unseen class article classification
- Generating unseen emoji labels for an image have also witnessed zero-shot learning-based solutions.

4 Conclusion

In this article, we provide an introduction of zero-shot learning. We first gave an rundown of the need for zero-shot learning by analyzing several other learning paradigms and listing several typical application scenarios in which zero-shot learning is needed. Then, we give an overview of zero-shot learning and introduce the different learning settings. We then explain the general zero shot learning framework and Model Architecture. Then we finally discuss various application areas where zero shot learning has performed well.

4.1 Good, Bad and Ugly Dimension

- **Good**: An important direction that has gained interest.
- **Bad**: No unified evaluation protocol exists.
- **Ugly**: Test Classes overlap with large Datasets.

5 References

Mensink; E. Gavves; Z. Akata; G.M. Snoek, **'Zero-Shot Learning for Computer Vision'**, CVPR 2017, Honolulu

Dr. Timothy Hospedales, **'Zero-Shot Learning'** , Yandex School of Data Analysis

Wei Wang, Vincent W. Zheng, Han Yu and Chunyan Miao, **'A Survey of Zero Shot Learning: Settings, Methods, and Applications'**

Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, Arnold W.M. Smeulders QUVA Lab : **'Tracking by Natural Language Specification'**

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean **'Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation'**