The Analyzed Approach of Zero-Shot Learning

Raghav Agrawal
Computer Science and Engineering
Indian Institute of Information Technology Nagpur
bt19cse044@iiitn.ac.in

Neha Kalbande
Computer Science and Engineering
Indian Institute of Information Technology Nagpur
bt19cse047@iiitn.ac.in

Hemanshu Chaudhary
Computer Science and Engineering
Indian Institute of Information Technology Nagpur
bt19cse056@iiitn.ac.in

Abstract- Zero-shot learning (ZSL) aims at understanding unseen categories with no training examples from class-level descriptions. Due to the importance of zero-shot learning i.e. classifying images where there is a lack of labeled training data, the number of proposed approaches has recently increased steadily. The objective is to enable models to perform under practical settings where data annotation could be infeasible, could be expensive or new classes get added over time. So, Illustratively, you have a few that are known as base classes where you have many examples. The goal is to use them to learn a classifier in such a way that if you have other classes, for which you have very few or zero samples, then the classifier can still work and be able to give you a good prediction.

Keywords— Generalized Zero-shot Learning

I. PROBLEM WITH DL MODELS

Deep learning models are known to be heavily reliant on large amounts of labeled data during training. This is one of the reasons they work so well but also that's one of their major limitations. If you had only a few samples to learn from, deep learning models may not be that effective in giving you a good model and a good prediction performance.To a certain extent, regularization methods are used to avoid overfitting in low data regimes but they don't really give you good performance on low data regimes. So, what do we do ? Solution is to be able to train models explicitly that are capable of rapidly generalizing to new tasks with only a few samples or perhaps no sample at all.

A. Problem Setting

Let x be an image or a feature and y denotes the class label for that data point, then Training Data

$$S = \{(x,y,a(y)) \mid x \in Xs, y \in Ys, a(y) \in A\}$$

And Test Set,

$$U = \{(x,y,a(y)) \mid x \in Xs, y \in Ys, a(y) \in A\}$$

Where x,y denotes the training data, a(y) which are some attributes related to each class y and both x and y belong to seen classes. Set of attributes, a(y) belongs to all classes.

II. INTRODUCTION

Zero-shot learning aims to recognize objects whose instances may not have been seen during training [1]. Recently, zero-shot learning (ZSL), which aims to recognize unseen-class samples given a set of labeled seen-class samples and the semantic features of both the seen and unseen classes, has attracted increasing attention in the fields of machine learning and computer vision. The key to ZSL is to learn an appropriate mapping between visual and semantic features, which could be adapted to the unseen-class domain. Zero shot learning can be classified as,

A. Conventional Zero-Shot

The goal is to learn a classifier f such that the image or feature x to be recognised at test time comes only from the unseen or few/zero shot classes, not from the base classes or seen classes. That is what is known as conventional zero shot.

B. Generalized Zero-Shot

This is more challenging and practical setting where the goal is to still learn a classifier going from x to both seen and unseen classes or base and zero/few shot classes where the image or feature x to be recognised at test time may belong to either seen or unseen , or base or zero/few shot. What makes it more

challenging is in the real world, we cannot predict whether your image at test time will come only from an unseen class or only from a few shot classes. It could come from any class in your universe of classes.

Lets understand more with the help of an example. Consider the animal category “oregon”. As usual, we haven’t heard about this category or seen any visual examples in the past, we can still learn a good visual classifier based on the following description: “zebra-striped four legged animal with a brown torso and a deer-like face”. Apart from this our traditional machine learning algorithms can only recognize the categories they are trained with. To add a new category we require collecting hundreds of training examples and then retraining the classifiers. This is hectic, time consuming and obviously not feasible in every case. To tackle this problem, zero-shot learning is often used. Transferring knowledge obtained from familiar classes to describe the unfamiliar classes is the key to deal with this problem. This can be done by using implicit knowledge representations, i.e. semantic embeddings. In this approach, one learns a vector representation of different categories using text data and then learns a mapping between the vector representation and visual classifier directly.

Fig. 1. Oregon Deer Zebra class

II. NEURAL NETWORK AND LOSS FUNCTION

Siamese networks are a special type of neural network architecture. In this network instead of model learning to classify, the neural network learns to differentiate between two inputs. It learns the similarity between them. So, how do we do that ?

When we encode and put the data into the classification model into the cnn solution neural net then applying pooling and strides and at the end coded with fully connected and finally putting this into softmax regressor for getting the highest probability results out. The objective here is to encode the image by doing convolution and get feature maps. Now, in siamese neural network we are going to re propose the Convolution neural net to be used a mechanism that would encode the data from an image into a certain vectors and then if these vectors are from the same person then we will hypothesize or assume that final encoding layer should be similar to one another that means if we were to compare the distance between the final layer and the coded version, this should be smaller. Consider f(x1), the feature vector of image 1 and f(x2), the feature vector of image 2. So, if these two people come from the same set then the euclidean distance between these two vectors is small.

In order to train this type of neural network, we should have a loss function[2,3] for our assumed hypothesis. Usually we use the Triplet Loss function with the siamese neural network. In this a reference input (called anchor) is compared to a matching input (called positive) and a non-matching input (called negative) sample. Our aim is to minimize the distance from the anchor to the positive sample and maximize the distance from the anchor to the negative sample. The loss function is defined as,

$$L(a, p, n) = \max(0, D(a, p) — D(a, n) + \text{margin})$$

Where D(a,p) represents the euclidean distance between anchor and positive sample, similarly with D(a,n).

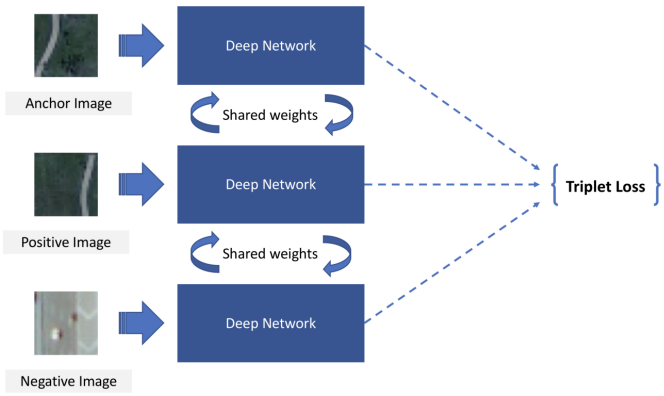

Fig. 2. Triplet Loss Architecture

III. RELATED WORKS

Early Zero-Shot Learning(ZSL) approaches; a key idea to facilitate zero-shot learning is finding a common semantic representation that both seen and

unseen classes can share. Attributes and text descriptions are shown to be effective shared semantic representations that allow transferring knowledge from seen to unseen classes.

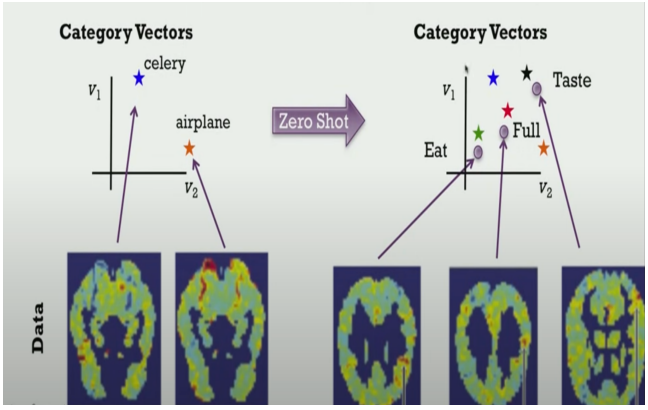

Fig. 3. Mind Reading Pipeline

This Mind Reading idea, scanning people's brain by FMRI actually motivated the whole field and reduced the dependency of EEG Tests.

Every data point is some kind of brain scan data structure and you have category vector for every word of the person might be expressing and we can again train the right sort of regressions and then apply it to some new word and so you input some new brain scans and you can try to generalize, add the ability to read all sorts of words out of people's mind.

IV. CHALLENGES

Although very attractive, one-shot learning is still in the developing stage and requires lots of research to achieve a bigger milestone. Each Siamese neural network trained on triplet loss is only useful for the one task it has been trained on. Can you think why? A neural network trained for zero-shot or one shot learning for a particular recognition task can't be used for some other task, such as telling whether two pictures contain the same cat or the same dog.

This kind of neural network is also reactive to other variations. For example, the accuracy might fall considerably if the person in one of the images is wearing sunglasses, a mask or a scarf and the person in the other image is not.

REFERENCES

[1]. C. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," in TPAMI, 2013.

[2]. Elad Hoffer, Nir Ailoni, Tsuchiya & Nakagawa, Seiji. (2020). Deep metric learning using Triplet network. 2. 153-160. 10.1016/j.sbspro.2010.01.029.

[3]. Xingping Dong , and Jianbing Shn, (2018). Triplet Loss in Siamese Network for Object Tracking.

[4]. Towards Data Science, Geeks for geeks article, wikipedia