Semantic Autoencoder for Zero-Shot Learning
Authors: Elyor Kodirov, Tao Xiang, Shaogang Gong

Introduction

With the research in computer vision moving towards large-scale datasets, even in challenges on datasets like ImageNet, usually only 1K classes are used for training purposes, of a total of 21,814 classes, owing to the imbalance in classes. In order to tackle the problem of classification of unseen data, zero-shot learning is being explored upon. The authors in this research work explore the encoder-decoder paradigm in autoencoders to implement the concept of zero-shot learning. The encoder part of the autoencoder is used for training on the dataset to learn a projection of the data into the semantic space, while the decoder part exerts an additional constraint that the projection learnt is able to reconstruct the original visual features. The authors go on to prove the importance of this constraint to be able to generalize on the data and predict labels for unseen data.

Zero-shot learning constitutes developing a projection from the visual space to the semantic space. The encoder in the semantic autoencoder is useful for learning the projection function from a visual feature space to a semantic feature space using word2vec word vector techniques, and the decoder part provides a constraint to tackle the nearest neighbour search problem which gets biased towards a few labels in the high dimensional space. Thus, the semantic autoencoder works like an undercomplete autoencoder.

Experimentation

The authors use six datasets for experimentation, four of which are small-scale datasets (AwA, CUB, aP&Y, SUN), and two are large-scale datasets (ImNet-1, ImNet-2). For the former, attributes are used as the semantic space, while for the latter, word2vec word vectors are trained from a corpus of 4.6M Wikipedia documents. Furthermore, the authors use AlexNet features for ImNet-1 and GoogleNet features for all other datasets. The semantic autoencoder (SAE) model uses only one free parameter, whose value is set by class-wise cross-validation using the training data. Multi-way classification accuracy is used as an evaluation metric for the small-scale datasets and hit5 classification accuracy is used for the large-scale datasets.

The results of the SAE model are compared with 14 existing ZSL models for the small-scale datasets, and with 7 existing ZSL models for the large-scale datasets; all of which have been published a maximum of two years prior to this work, and all of whom are state-of-the-art. The classification accuracies of each of these datasets has been recorded in Table 1 in the paper. It can be observed that SAE model performs better than all the other models for all datasets except AwA, for which the state-of-the-art model $SynC^{struct}$ performs better. The authors also experiment on the generalized zero-shot learning setting, wherein, the test set contains both seen and unseen classes. For this series of experiments, *Area Under Seen-Unseen Accuracy Curve (AUSUC)* is used, which measures how well the model distinguishes between seen and unseen classes.

Among the zero-shot learning models introduced prior to the semantic autoencoder, the more efficient ones include SSE, ESZSL and AMP. The authors compare the performance of these models with SAE on the AwA dataset, as recorded in Table 5. It was observed that SAE is atleast 10 times faster than all the models, of which, ESZSL is the closest in terms of computational cost.

In the next series of experiments, the authors extend the SAE model to the supervised clustering problem. They create two synthetic datasets from the Oxford Flowers-17 dataset, through a clustering technique, wherein, in one of the datasets, there are three clusters each with 1000 samples, and in the other dataset, there are three clusters with 1000, 2000 and 4000 samples respectively. The datasets are created such that the clusters contain subclusters which are closer to subclusters from different clusters, than from it's own when measured in Euclidean distance. In the feature extraction step in these experiments, the SIFT and color features are extracted from each 8 X 8 patch centered at each pixel thus resulting in a 135D feature vector for each pixel. The SAE model is compared with six other state-of-the-art supervised clustering models, as recorded in Table 6 and Table 7 for the datasets with same cluster sizes and different cluster sizes respectively. While SAE outperforms all other models, the training time is second to only the MLCA model.

Strengths & Weaknesses

One of the major strengths of this paper is the computationally superior linear projection function, which is more generalizable than the other ZSL models. The SAE is extremely useful for learning a low-dimensional semantic representation of input data that can be used for data reconstruction. The SAE model also outperforms the other state-of-the-art supervised clustering models with a superior training time, taking more training time only than the MLCA model. One of the weakness of this paper is that the SAE model reconstructs from the attribute space to the visual space by learning a linear compatibility between each of them. This has to be explored for more image classification problems which have non-linear decision boundaries.

Further Work

In [1], the authors explore conditional variational autoencoders which model non-linear decision boundaries better than the semantic autoencoders proposed in this work. Furthermore, in [2], the authors propose novel generalized zero-shot learning algorithms which have shown to also outperform SAE considerably.

References

[1] A. Mishra, M. Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. arXiv preprint arXiv:1709.00663, 2017

[2] Kumar Verma, Vinay, et al. "Generalized zero-shot learning via synthesized examples." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.