Knowledge Distillation using Teacher-Student Network

Kalyan Ghosh
Dept of Electrical &Computer Engineering
North Carolina State University
Raleigh, USA
kghosh@ncsu.edu

Shrey Anand
Dept of Computer Science
North Carolina State University
Raleigh, USA
sanand3@ncsu.edu

Akshay Kalkunte Suresh
Dept of Electrical &Computer Engineering
North Carolina State University
Raleigh, USA
akalkun@ncsu.edu

***Abstract*—We propose to train a deep network as the teacher network on a larger dataset composed of 2 modalities and also a shallow student network composed of a single modality. We hypothesize that, using a teacher-student to distill knowledge, the accuracy of the shallow student network on an unseen test set should be comparable to the accuracy of the deep network and should also have a significantly lower inference time.**

***Index Terms*—Teacher-Student Network, Knowledge Distillation.**

I. Motivation

Complex deep learning networks require enormous amounts of data to train and deploying them on small embedded devices is unfeasible. A promising solution for the problem is *Knowledge Distillation* or more specifically Teacher Student Networks [1].

In Teacher Student Networks, a deep and complex teacher network is trained on a large dataset, powerful enough to learn complex features. Then, we train a much smaller and shallow student network propagating the information from the teacher network to achieve a comparable performance.

For the task of classifying a celestial observation as star, quasar or a galaxy, we aim to develop a small student network that can give comparable performance as compared to a pretrained teacher network.

II. Dataset

The dataset that we are using for our task is the Sloan Digital Sky Survey which offers public data of space observations. The dataset consists of 10,000 observations of space taken by the SDSS. Every observation is described by 17 feature columns which is a combination of data coming from a photometric source and a spectral source and 1 class column which identifies it to be either a star, galaxy or quasar.

Spectral features are found to be more accurate than photometric features for the classification task, but photometric techniques can be applied to objects a couple of magnitudes (a factor of 5) fainter.

III. Methodology

Our model is composed of a teacher and a student network. The teacher network is a Deep Neural Network that is trained on a set of 17 features which is a combination of data acquired from photometric and spectral observations. The student network is a shallow network trained only on data from photometric sources. The dataset is divided into teacher-training, student-training and testing sets. The teacher network is trained on the teacher network and soft-probabilities are obtained on the student-training dataset. The student network is then trained on the student-training set with only the photometric data with the soft labels obtained from performing a forward-pass of the teacher network. We also plan to experiment with other techniques of transferring information from the teacher to the student network like, simultaneously training the teacher and student networks and backpropagating a loss that combines the loss of the teacher and the student network. Another technique would be learn an ensemble of models for the teacher network, which would be more robust than using a single DNN model and transferring information from the soft-labels of this ensemble model to the student network.

If time allows, in another experiment, we aim to explore the use of teacher student network in the field of NLP. Kyrios et al. develop a complex LSTM encoder decoder network to extract a latent feature space representing context of sentences [2]. The dataset for the teacher network is composed of features from 2 modalites, uni-gram and bi-gram vector representations. The inference time and the memory requirements of the model is relatively high and we hypothesize that it can be reduced through a simpler LSTM network.

IV. Evaluation

For our evaluation phase, we will use the performance of the Teacher as our baseline reference. Then,we will perform a quantitative and qualitative comparative analysis of the performance of the shallow Student Network. Accuracy of classification and the Inference time taken to classify one observation will be our prime metrics.

REFERENCES

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.