Autonomous Learning: from Large-Scale Data without Annotation

Sachit Kumar, sk4661, Yangchen Huang, yh3223
Topic
Autonomous Learning
Date
02/14/2020

Abstract:

Semi-supervised learning is a branch of machine learning that leverages unlabeled data since labeling data is expensive. It is recently gaining a lot of research attention especially for classification problems. Therefore, to gain a solid understanding and practical exposure to semi-supervised learning, we decided to implement a generalized semi supervised learning algorithm to solve classification problems. For this, we will choose an appropriate model architecture and perform some novelty in in label guessing algorithms, regularizations, etc. We will also implement and test our model. Our goal will be to obtain a worthy accuracy-privacy trade off and an improved performance overall.

1. Background (Review of Related Literature):

Semi-supervised learning (SSL) is a halfway between supervised and unsupervised learning [1]. The training set for an SSL algorithm is a combination of labeled and unlabeled examples. In practice, the process of SSL is a two-step approach as follows: (1) Given the distribution of a small set of labeled data, SSL will automatically 'guess' the labels for unlabeled examples, and mix them up to become a new much larger training set. (2) Train the model using the new training set, and make predictions on the test set.

Popular semi-supervised learning fall under the following classes: (1) Generative Mixture Models (2) Low-Density Separation Models (3) Graph-Based Models (4) Metric Based Models.

Generative Mixture Models (GMMs) assume a generative model $p(x, y) = p(x)p(y|x)$. In order to estimate $p(y|x)$, one can use information on $p(x)$, which additional unlabeled data would help a lot. In other words, GMM can be considered as classification with additional information on the marginal density, or a 'soft cluster'. Expectation-Maximization (EM) algorithm is a common approach. Nigam el al. (2000) apply the EM algorithm in a text classification problem and show that it achieves better performance than models trained only from labeled data [2]. Fujino el al. (2005) extend GMMs by including a bias correction term and discriminative training using the max-entropy principle [3]. Self-training and Co-training are two popular techniques used often in GMMs. Rosenberg et al. (2005) apply self-training to object detection in images [4]. Jones (2005) usded co-training and co-EM to do text information extraction [5]. Recently, with the booming of deep learning, Generative Adversarial Networks are employed in semi-supervised learning and reached better performance. [6, 7]

Low-Density Separation Models (LDSMs) aim to directly implement the low-density separation assumption which holds for SSL. This smoothness assumption claims that if two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2. Thus, most LDSMs, with various architectures, utilize algorithms to push the decision boundary away from unlabeled data points. Transudative Support Vector Machines (TSVM) was introduced by Vapnik (1998) and was proved to be beneficial for semi-supervised learning [8]. Lawrence and Jordan (2005) employ Gaussian Process parallel of TSVM [9]. Szummer and Jaakkola (2002) design a information regularization framework to control $p(y|x)$ by $p(x)$, guaranteeing that labels would not change too much in regions where $p(x)$ is high [10]. Zhu (2005) uses entropy minimization to tune the hyperparameter [11]. David et al. (2019) proposed MixMatch, unifying current dominant approaches in LDSMs and achieves very good performance [12].

Graph-Based Models are built on the manifold assumption: The high-dimensional data lie roughly on a low-dimensional manifold [13]. Zhu (2005) presents a series of novel semi-supervised learning approaches arising from a graph representation, where labeled and unlabeled instances are represented as vertices, and edges encode the similarity between instances [11]. Liu (2019) propose a generic framework to create much more training data through label propagation from the few labeled examples to a vast collection of unannotated examples [14].

Metric-Based Model is a model selection method to detect hypothesis inconsistency with unlabeled data. using unlabeled data [15]. It's a general method which can be applied to almost any learning algorithms. However, it only selects among hypotheses and does not generate new hypothesis based on unlabeled data. Madani et al. (2005) uses this method for model selection and active learning [16].

2. Introduction to the Project:

Recently, deep neural networks have been gaining traction. This method tends to be very effective and provides a great performance as well. However, although there is an abundance of data, very few of them are actually labeled - as needed by deep learning models. Plus, it seems very tedious and expensive for humans to always label data. To minimize this need of labeling data, we try and implement a semi-supervised learning model. To progress with semi supervised learning firstly we must find a sizable unlabeled dataset, otherwise the problem can simply be addressed as a supervised learning problem. We must then implement a label guessing algorithm – to guess the labels of the unlabeled data, and a customized loss function - to understand the accuracy of our guessed labels for the unlabeled data. We will also perform current supervised learning practices of data augmentation – create new data points from existing data points, regularization - to prevent overfitting and help the semi-supervised learning smoothness assumption holds, entropy minimization – reduce the mix up/randomness of classes about a particular point, and hyper parameter tuning. We will then proceed to build, train and test our model, and compare our model to baseline models. Finally, we will try and improve the results of our model with the goal of getting a performance as good as possible, and a lower error rate.

3. Introduction to the Dataset:

We looked up many datasets for this project such as: MNIST, CIFAR-10, SVHN and STL-10. After some consideration, we decided to use the STL-10 dataset as it seemed to be the best for semi-supervised learning.

STL-10 is an image recognition dataset which inspired by the CIFAR-10 dataset. The difference is that it contains fewer labeled training examples and a larger number of unlabeled examples. The image resolution is higher as well.

In this dataset:

- Images are divided into 10 classes: bird, airplane, car, cat, dog, deer, horse, monkey, ship, and truck.
- Images are 96x96 pixels in resolution.
- There are 500 labeled training images and 100,000 unlabeled images.
- Data is divided into three files: train_image.zips, test_images.zips, and unlabeled_images.zips.

4. Plan:

Our plan between every milestone is:

Milestone1:

- Do literature research
- Learn about different approaches of SSL.
- Prepare the dataset

Milestone2:

- Design the appropriate model architecture (data augmentation, label guessing algorithm, loss function, regularization, entropy minimization, etc)

Milestone3:

- Implement the model and test it on different datasets

Final:

- Compare the result with existing models
- Add improvements

Reference:

[1] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-Supervised Learning. MIT Press, 2006.
[2] Nigam, Kamal, et al. "Text classification from labeled and unlabeled documents using EM." Machine learning 39.2-3 (2000): 103-134.
[3] Fujino, Akinori, Naonori Ueda, and Kazumi Saito. "A hybrid generative/discriminative approach to semi-supervised classifier design." AAAI. 2005.
[4] Rosenberg, Chuck, Martial Hebert, and Henry Schneiderman. "Semi-supervised self-training of object detection models." WACV/MOTION 2 (2005).
[5] Jones, Rosie. Learning to extract entities from labeled and unlabeled text. Diss. Carnegie Mellon University, Language Technologies Institute, School of Computer Science, 2005.
[6] Pu, Yunchen, et al. "Variational autoencoder for deep learning of images, labels and captions." Advances in neural information processing systems. 2016.
[7] Odena, Augustus. "Semi-supervised learning with generative adversarial networks." arXiv preprint arXiv:1606.01583 (2016).
[8] Vapnik, V. Statistic Learning Theory. Springer, 1998
[9] Lawrence, Neil D., and Michael I. Jordan. "Semi-supervised learning via Gaussian processes." Advances in neural information processing systems. 2005.
[10] Szummer, Martin, and Tommi S. Jaakkola. "Information regularization with partially labeled data." Advances in Neural Information processing systems. 2003.
[11] Zhu, Xiaojin, John Lafferty, and Ronald Rosenfeld. Semi-supervised learning with graphs. Diss. Carnegie Mellon University, language technologies institute, school of computer science, 2005.
[12] Berthelot, David, et al. "Mixmatch: A holistic approach to semi-supervised learning." Advances in Neural Information Processing Systems. 2019.
[13] Subramanya, Amarnag, and Partha Pratim Talukdar. "Graph-based semi-supervised learning." Synthesis Lectures on Artificial Intelligence and Machine Learning 8.4 (2014): 1-125.
[14] Liu, Bin, et al. "Deep metric transfer for label propagation with limited annotated data." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.
[15] Schuurmans, Dale, and Finnegan Southey. "Metric-based methods for adaptive model selection and regularization." Machine Learning 48.1-3 (2002): 51-84.
[16] Madani, Omid, David M. Pennock, and Gary W. Flake. "Co-validation: Using model disagreement on unlabeled data to validate classification algorithms." Advances in neural information processing systems. 2005.