Leveraging Auxiliary Data for Semi-Supervised Learning

Author: Ross Briden
Collaborators: Wasu Piriyakulkij, Cristina Menghini, Nihal V. Nayak, Jeffrey Zhu, and Elaheh Raisi
Faculty Sponsor: Stephen Bach

Abstract

State-of-the-art semi-supervised learning algorithms leverage both labeled data, which is often scarce, and unlabeled data, which is often abundant, to learn a given task. However, little attention has been paid to semi-supervised learning with *auxiliary data*, labeled datasets from other tasks. In this paper, we explore strategies for designing, building, and evaluating semi-supervised learning algorithms that leverage auxiliary data. This work is part of the TAGLETS project, which is available as an open-source system at github.com/BatsResearch/taglets.

Introduction

Over the past decade, deep neural networks — a class of algorithms which utilize labeled data to make inferences about unseen data — have achieved state-of-the-art performance in many domains, ranging from image classification to protein folding. Training deep neural networks, however, requires large labeled datasets, with algorithms often utilizing millions of training examples. In many domains, unlabeled data is widely available while acquiring large amounts of labeled data is prohibitively expensive. Semi-supervised learning (SSL) leverages this disparity by utilizing unlabeled data for training. Recently, the popularity of SSL has exploded, with many SSL algorithms matching — and sometimes exceeding — the performance of their fully-supervised counterparts while requiring only a fraction of the labeled data. However, state-of-the-art SSL algorithms fail to leverage another abundant source of supervision available to many machine learning practitioners: auxiliary data.

In TAGLETS, we demonstrate that, if structured intelligently, auxiliary data is a useful and cheap source of supervision for SSL algorithms, achieving state-of-the-art results in several image classification benchmarks. We organize auxiliary data using a Structured Collection of Annotated Datasets, or SCADS, enabling auxiliary data to be efficiently queried based semantic embeddings. Given this abstraction, we can then turn to the issue of how to best use auxiliary data with existing state-of-the-art SSL algorithms.

Contributions

In our experiments, we examine three state-of-the-art SSL algorithms: FixMatch [1], SimCLRv2 [2], and Meta Pseudo Labels [3]. First, we reimplemented each of these algorithms in PyTorch and verified their performance. Unfortunately, we found that the performance of SimCLRv2 deteriorates significantly when trained on smaller datasets; consequently, we do not include this

Method	Auxiliary Data	OfficeHome-Product			OfficeHome-Clipart		
		1-shot	5-shot	20-shot	1-shot	5-shot	20-shot
FixMatch	Yes	56.87 ± 3.29	82.77 ± 1.01	90.05 ± 0.96	29.90 ± 1.34	61.49 ± 4.12	75.28 ± 0.22
Meta Pseudo Label	Yes	47.38 ± 2.98	76.67 ± 1.75	87.38 ± 1.75	22.77 ± 3.99	57.25 ± 4.64	75.17 ± 1.22
FixMatch	No	39.28 +/- 3.34	72.87 ± 1.45	85.03 ± 1.81	20.51 ± 0.96	49.69 ± 3.31	66.56 ± 1.23
Meta Pseudo Label	No	41.84 ± 3.02	73.10 ± 1.29	87.08 ± 1.64	21.57 ± 3.99	50.88 ± 4.76	70.17 ± 1.43

Table 1: The performance of FixMatch, Meta Pseudo Labels, and other algorithms on the OfficeHome-Product and OfficeHome-Clipart datasets. For these experiments, FixMatch and Meta Pseudo Labels both use strategy 1 to train with auxiliary data.

method in our results. Next, we retrofitted each of these algorithms to support auxiliary data, exploring two different strategies:

1. Pretraining the encoder for each SSL algorithm on auxiliary data.
2. Self-supervising the encoder for each SSL algorithm on auxiliary data.

Though simple, we achieve the best performance using strategy 1 (see Table 1). We hypothesize that supervised pretraining with auxiliary data is helpful because these samples contain useful low-level and high-level features, though future work is needed to confirm this. Strategy 2 was inspired by the architecture of SimCLRv2, which performs unsupervised pretraining on unlabeled data. However, this approach results in poor performance since all label information is discarded without increasing the amount of auxiliary data we sample.

After selecting an auxiliary data learning strategy, we evaluated the performance of FixMatch and Meta Pseudo Labels on the OfficeHome-Product and OfficeHome-Clipart datasets, with and without auxiliary data (see Table 1). Using auxiliary data, we observe a significant boost in accuracy across both benchmarks, particularly in the low-shot regime. Thus, our experiments demonstrate that auxiliary data, when chosen selectively, can improve the performance of state-of-the-art SSL algorithms.

Conclusion and Future Work

In this paper, we explore the applications of auxiliary data for semi-supervised learning. We observe that auxiliary data can improve the performance of state-of-the-art semi-supervised learning algorithms, especially when labeled data for a given task is scarce. Though we focused on existing semi-supervised learning algorithms, we hope future works will explore the possibility of creating novel semi-supervised learning algorithms that utilize auxiliary data more directly and intelligently.

References

[1] Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence.ArXiv, abs/2001.07685, 2020

[2] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervisedlearners.arXiv preprint arXiv:2006.10029, 2020

[3] Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V. Meta pseudo labels. In IEEE Conference on Computer Vision and Pattern Recognition, 2021.