STACI: Adapting the Soft Teacher Approach for Image Classification

Written by Saadhana Srinath

CIS 5700 - Advanced Data Mining
University of Michigan

Abstract

This study introduces the idea of adapting the end-to-end semi-supervised approach of Soft Teacher for Image Classification. It explores the application of the semi-supervised learning approach – Soft Teacher, originally designed for object detection, to the task of image classification using the CIFAR-10 benchmark. STACI essentially aims to improve classification accuracy with limited labeled data. It leverages a student-teacher framework with data augmentation and an exponential moving average (EMA) update technique to enforce consistency between the predictions of the student and the teacher model. The results highlight the model's efficacy in improving classification accuracy under a semi-supervised setting, providing insights into the utility of soft labels and consistency regularization.

Introduction

Supervised learning techniques for image classification often require large amounts of labeled data to achieve high accuracy. However, manually labeling data can be expensive and time-consuming. Semi-supervised learning has emerged as a promising approach to alleviate the reliance on labeled data by leveraging large amounts of unlabeled data in conjunction with a small labeled dataset for training. This approach can be particularly advantageous in scenarios where labeled data is scarce, but a significant amount of unlabeled data is readily available.

This project explores the versatility of the "Soft Teacher" method by applying it to image classification on the CIFAR-10 dataset. Originally designed for semi-supervised object detection by [Li et al., 2020] (cite the original paper here), this technique leverages a student-teacher framework with consistency regularization. The student model is trained on both labeled and unlabeled data, while the teacher model guides the student's learning by generating soft pseudo-labels for the unlabeled data. Consistency regularization encourages the student model to produce similar outputs for the same unlabeled data even with different augmentations. The key contribution lies in adapting and evaluating this method specifically for image classification, and the results demonstrate its effectiveness in bridging the gap between supervised learning performance and what can be achieved with limited labeled data.

Related Work

Several methods have been explored for semi-supervised learning. Discussed here are three seminal semi-supervised learning methods: the Mean Teacher (Tarvainen Valpola, 2017), MixMatch (Berthelot et al., 2019), and FixMatch (Sohn et al., 2020). The Mean Teacher method utilizes consistency regularization between a student model and a teacher model which is an exponential moving average (EMA) of the student model, achieving good performance on various datasets. MixMatch, meanwhile, further expands the consistency principle by combining it with Mixup and entropy minimization, creating an effective algorithm for semi-supervised learning. FixMatch on the other hand, leverages strong augmentations on unlabeled data and selects only the most consistent examples for pseudo-labeling, demonstrating effectiveness in reducing the impact of noisy labels. STACI inherits the main idea of ensemble knowledge from the Mean Teacher approach, incorporating an EMA-based teacher model and employing consistency regularization. It then utilizes augmentations for pseudo-labeling just like how FixMatch does.

Problem Formulation

Semi-supervised image classification requires a model to learn from both labeled and unlabeled data. The proposed approach includes a teacher model (an Exponentially Moving Average of the student model) and a student model. The training involves two key ingredients: classification loss, computed on labeled data, and consistency loss, encouraging the student model's predictions on augmented unlabeled data to match the soft pseudo-labels generated by the teacher model. STACI encapsulates this process into a self-trainable pipeline that promotes consistency across augmented perturbations of the same image.

Experiments

Experimental Methodology

The experimental methodology begins with the CIFAR-10 dataset and a division into labeled and unlabeled subsets. The ResNet18 architecture is employed as the backbone

architecture for both the student and teacher models, utilizing SGD with momentum for optimization. Baseline methods include the aforementioned FixMatch algorithm for performance comparison. The primary performance metric comprises model accuracy on the CIFAR-10 test set, which indicates the quality of semi-supervised learning.

The key components of the method are as follows:

- **Dataset:** CIFAR-10 dataset containing 60,000 images for training and 10,000 images for testing.
- **Data Augmentation:** Weak augmentation (random horizontal flip with a 50 percent probability and random translation vertically and horizontally up to 12.5 percent) and strong augmentation (aspect resizing, random horizontal flip, random geometrical transformations followed by some random erasing and cutout) are employed for unlabeled data.
- **Student Model:** A convolutional neural network with the ResNet-18 architecture is employed as the student model.
- **Teacher Model:** A convolutional neural network with the ResNet-18 architecture that is pre-loaded with ImageNet weights, is used as the teacher model. This teacher model is updated using an EMA of the student model parameters in each iteration. This enforces a slower update for the teacher compared to the student, providing more stable guidance during training.
- **Loss Functions:** Two loss functions are combined for training:
  - Classification Loss: Cross-entropy loss is used to measure the discrepancy between the student model's predictions and the true labels for the labeled data. Minimising the classification loss would result in the model matching the true labels better.
  - Consistency Loss: Kullback-Leibler divergence compares the predictions of the student model to those of the teacher model on the unlabeled data. This loss is employed to encourage the student model to produce similar outputs for the same unlabeled data even under different augmentations (weak and strong) during training. Minimising the consistency loss would mean that the student matches the teacher's predictions better promoting consistent behavior.

Training Procedure:

As shown in Figure 1, the student model is trained on both labeled and unlabeled data. The student model directly predicts on the labeled data and a cross entropy loss (supervised loss) is computed on these predictions to guide the supervised learning component. The unlabeled data is first strongly (to generate different versions of the same image) and weakly augmented. The strongly augmented labeled data is supplied to the student model and the weakly augmented labeled data is supplied to the teacher model. Both the models predict probability distributions for each of the classes. From the teacher's predictions, only the highly confident scores are considered to become the "soft" labels.

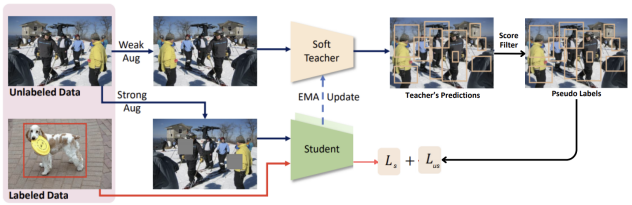


Figure 1: Shows the workflow of STACI

These soft labels are then compared to the predictions of the student model using the KL-divergence loss (unsupervised loss). Both the supervised and unsupervised losses are combined to complete the training process. In each iteration, once the student model updates it's weights based on the combined loss, the teacher model is updated using an EMA of the student model parameters.

Baselines

The performance of the proposed semi-supervised approach, STACI, is compared against the following baselines:

- Supervised Learning with ResNet-18 trained only on labeled data. Here, a simple residual neural network with 18 layers is trained on the CIFAR-10 dataset with a 100 percent labeled data ratio.
- FixMatch – a semi-supervised learning technique where a model is trained on a combination of labeled and unlabeled data. Here, a weakly-augmented image is fed into the model to obtain predictions. When the model assigns a probability to any class which is above a threshold, the prediction is converted to a one-hot pseudo-label. Then, the model predicts on a strong augmentation of the same image. The model is then trained to make its prediction on the strongly augmented version match the pseudo-label via a cross-entropy loss.

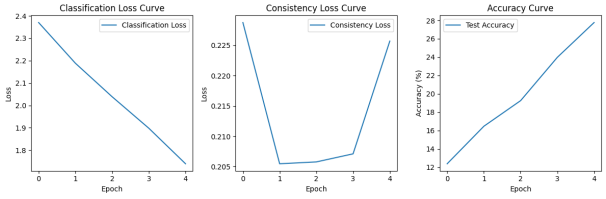
Performance Metric

Classification accuracy is used as the primary metric to evaluate the performance of all models. The accuracy is calculated against different ratios of labeled and unlabeled data.

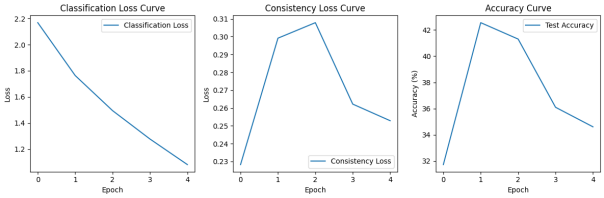
Results and Discussion

This study mainly evaluates the model's performance over varying ratios of labeled data following the base paper's methodology, over multiple runs for each ratio with 5 epochs.

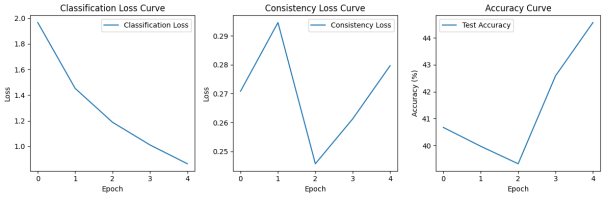
- With 1 percent Labeled Data:



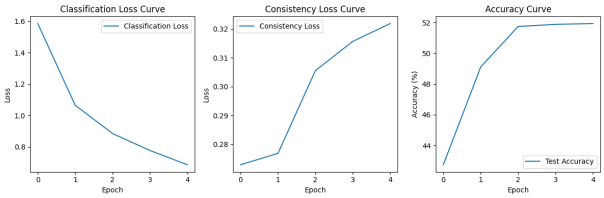
- With 5 percent Labeled Data:



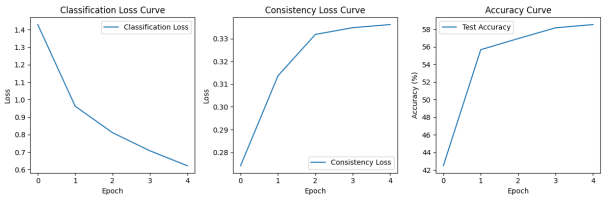
- With 10 percent Labeled Data:



- With 30 percent Labeled Data:



- With 45 percent Labeled Data:



Surely, as the percentage of labeled data increases, the performance also increases. Here, the fact that the model is trained only for 5 epochs owing to the unavailability of computational resources is also a major factor. Training the model over more number of epochs improves the model performance as shown below.

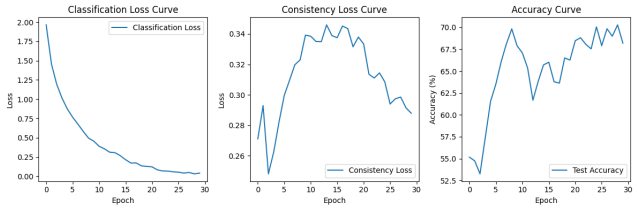


Figure 2: Labeled Ratio=10 Perc; Epochs=30

These results suggest that leveraging unlabeled data through the Soft Teacher approach can improve classification performance, especially when labeled data is limited.

Comparison with Baselines

Supervised Learning Baseline: The ResNet-18 model trained only on labeled data serves as a baseline for performance. Since this is a simple supervised approach, it's expected to perform well on labeled data but may not utilize additional unlabeled data effectively.

FixMatch: When comparing with FixMatch, note the role of weak and strong augmentations and the process of generating pseudo-labels. FixMatch typically outperforms purely supervised approaches when a significant amount of unlabeled data is available since it effectively leverages unlabeled instances.

STACI: The semi-supervised approach in uses both a student and a teacher model, utilizing unlabeled data for learning. It would be expected to and in fact does better than the purely supervised approach, but its performance relative to FixMatch would depend on how effectively the pseudo-labels and consistency loss are utilized.

Running FixMatch for varying ratios of labeled and unlabeled data is out of the scope of this project owing to time constraints but the comparison on STACI with the Supervised Learning approach over 30 epochs is detailed below.

Labeled Data Ratio	Supervised Learning	Semi-Supervised Learning (Soft Teacher)
45%	~70	~71
30%	~65	~67
10%	~58	~60
5%	~53	~56
1%	~48	~54

Figure 3

Hyperparameter Analysis

In any deep learning system, hyperparameters play a critical role in controlling the behavior of the training algorithm and the performance of the trained model. The effectiveness of this adaptation of "Soft Teacher" from semi-supervised object detection to image classification is in part contingent on careful selection and tuning of hyperparameters. Here, we discuss key hyperparameters included in the code and their impact on the training process and results.

- **Batch Size:** Batch size (64) affects memory utilization and convergence properties. Larger batch sizes can lead to more stable and reliable gradient estimates leading to an increase in performance.
- **Learning Rate:** The initial learning rate for Stochastic Gradient Descent (0.001) is a modest choice that strikes a balance between convergence speed and stability. However, for this particular problem I found that the value of 0.01 works better after a little bit of hyperparameter tuning.

- **SGD Momentum:** Momentum (0.7) helps accelerate SGD in the relevant direction and dampens oscillations, improving convergence. A momentum of 0.9 is usedd normally but 0.7 works better in this case.
- **Decay for EMA (Exponential Moving Average):** 'decay' (0.9), determined after some amount of tuning, is used in updating the teacher model, determines how much the teacher's weights are influenced by the current student's weights. A higher decay value places more trust in the historical student model weights, providing a stabilizing effect.
- **Number of Epochs:** 'numOfepochs' (5), representing the number of complete passes through the dataset, was chosen to demonstrate the effectiveness of the approach without extensive training.

By tuning these hyperparameters, we can ensure the learning process efficiently leverages both labeled and unlabeled data for improving overall classification performance.

Although not hyperparameters per say, the factors discussed below also play an important role in improving the performance of the model.

- **Seed values for reproducibility:** Seed values for NumPy, PyTorch, CUDA, and the random library are set to a specific number to ensure reproducibility of results.
- **Model Architecture:** ResNet18, pre-loaded with ImageNet weights, is used as the backbone. This choice stems from ResNet's robust performance across various vision tasks and its manageable computational load. Using different architectures with different initial weights can result in drastic changes in performance.
- **Data Augmentation Techniques:** Different Augmentation techniques work for different datasets. Here, STACI follows the weak augmentations from Fixmatch and strong augmentations from the base paper. Changing the augmentation techniques can result in drastic changes in performance.
- **Normalization Constants:** Mean and standard deviation constants normalize the input data, ensuring it has zero mean and unit variance. This helps accelerate training by standardizing the input distribution. These values depend entirely on the dataset in question.

Future Work

The Supervised approach used to compare STACI with, is state-of-art, developed over several years. STACI on the other hand is a novel approach with has a lot of potential particularly in situations where labeled data is not readily available. Even without proper research, STACI performs decently with more amounts of unlabeled data. However, it still requires alot of experimentation and development. Future work could involve utilizing hyperparameter optimization techniques such as grid search, random search, or Bayesian optimization to find the optimal set of hyperparameters and achieve even better performance.

Conclusion

This study confirms the viability of applying soft pseudo-labeling strategies from semi-supervised object detection to image classification. Future extensions could explore the integration of contrastive learning for better representation learning in the unlabeled data space, or the use of meta-learning approaches to dynamically adjust the consistency regularization based on model confidence.

References

Xu, Mengde, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. "End-to-end semi-supervised object detection with soft teacher." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 3060-3069. 2021.

Berthelot, David, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. "Mixmatch: A holistic approach to semi-supervised learning." Advances in neural information processing systems 32 (2019).

Sohn, Kihyuk, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence." Advances in neural information processing systems 33 (2020): 596-608.