STAR: Noisy Semi-Supervised Transfer Learning for Visual Classification

Hasib Zunair
Concordia University
Montreal, Canada

Yan Gobeil
Décathlon
Montreal, Canada

Samuel Mercier
Décathlon
Montreal, Canada

A. Ben Hamza
Concordia University
Montreal, Canada
hamza@ciise.concordia.ca

ABSTRACT

Semi-supervised learning (SSL) has proven to be effective at leveraging large-scale unlabeled data to mitigate the dependency on labeled data in order to learn better models for visual recognition and classification tasks. However, recent SSL methods rely on unlabeled image data at a scale of billions to work well. This becomes infeasible for tasks with relatively fewer unlabeled data in terms of runtime, memory and data acquisition. To address this issue, we propose noisy semi-supervised transfer learning, an efficient SSL approach that integrates transfer learning and self-training with noisy student into a single framework, which is tailored for tasks that can leverage unlabeled image data on a scale of thousands. We evaluate our method on both binary and multi-class classification tasks, where the objective is to identify whether an image displays people practicing sports or the type of sport, as well as to identify the pose from a pool of popular yoga poses. Extensive experiments and ablation studies demonstrate that by leveraging unlabeled data, our proposed framework significantly improves visual classification, especially in multi-class classification settings compared to state-of-the-art methods. Moreover, incorporating transfer learning not only improves classification performance, but also requires 6x less compute time and 5x less memory. We also show that our method boosts robustness of visual classification models, even without specifically optimizing for adversarial robustness.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence.

KEYWORDS

Semi-supervised learning, self-training, transfer learning, visual classification

MMSports 2021, October 20–24, 2021, Chengdu, China

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Hasib Zunair, Yan Gobeil, Samuel Mercier, and A. Ben Hamza. 2021. STAR: Noisy Semi-Supervised Transfer Learning for Visual Classification. In *MMSports 2021: 4th International ACM Workshop on Multimedia Content Analysis in Sports, October 20–24, 2021, Chengdu, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Visual classification is a fundamental problem in computer vision. It refers to the process of organizing a dataset of images into a known number of classes, and the task is to assign new images to one of these classes. Conventional machine learning models for visual classification rely on labeled data to train a classifier [8]. While unlabeled data can be generally obtained with minimal human labor, labeling data is often laborious, costly and requires the efforts of experienced human annotators. Semi-supervised learning (SSL) addresses this issue by leveraging large volumes of unlabeled data together with a relatively small amount of labeled data in a bid to learn better visual classifiers [20]. SSL has had a resurgence in recent years, in large part thanks to its ability to improve the accuracy of the model on important benchmarks. The objective of semi-supervised learning for the classification of images is to predict the most probable labels of images in a dataset by leveraging labeled and unlabeled data in order to train a classifier.

Some of the early approaches to semi-supervised visual classification have focused mostly on low-level computer vision frameworks, which involve hand-engineered features. These methods include co-training [1] and self-training [11] approaches, as well as transductive support vector machines and graph-based techniques [10]. Graph-based methods, for instance, capture the manifold structure of the data and encourage similar points to share labels. A major limitation with graph-based approaches is the need for similarity measures that create graphs with no inter-class connections. In many real-world applications, it is not easy to learn such a good visual similarity metric. Moreover, intra-class variations are often larger than inter-class variations, making pairwise similarity based methods of little utility. In addition, hand-crafted features often lead to unsatisfactory results on unseen data and do not generalize well across different tasks.

The advent of deep learning has sparked groundswell of interest in the adoption of deep neural networks (DNNs) for semi-supervised visual classification. A plethora of DNNs is based on convolutional neural networks, including pre-trained models such as

InceptionV3, ResNet, and EfficientNet [6, 14–16], as well as SSL models with large convolutional networks based on a teacher/student paradigm by training on the labeled data to get an initial teacher model, followed by training a student model [18]. These pre-trained models can be used for visual classification, feature extraction, and fine-tuning. Using a pre-trained network with transfer learning is typically much faster and easier than training a network from scratch.

In this paper, we introduce noisy semi-supervised transfer learning (STAR), an efficient SSL framework for visual classification with the goal of remedying the aforementioned issues. Our approach mitigates the issue of time required in training large models, while improving classification performance and reducing overfitting. The proposed approach is well suited for tasks that can leverage unlabeled image data on a scale of thousands, and is comprised of two main integrated stages. In the first stage, we train a supervised learning model, pre-trained on the ImageNet database [3], on the labeled data. This model is then used to generate pseudo-labels for the unlabeled data. In the second stage, a larger pre-trained model is trained on the combined labeled data and pseudo-labeled data. The larger model is noised using data augmentation and dropout [13] during training. The main contributions of this paper can be summarized as follows:

- We propose a noisy semi-supervised transfer learning framework by integrating transfer learning and noisy student training with the goal of improving visual classification performance and reducing computation overhead.
- We show through extensive experiments that our approach yields significant improvements over strong baseline methods on binary and multi-class classification tasks. These improvements are not only in terms of classification performance, but also in terms of computation time and memory required to train models for the desired tasks.
- We show that our method boosts robustness in visual classification models without specifically optimizing for adversarial robustness.

The rest of this paper is organized as follows. In Section 2, we review important relevant work. In Section 3, we present our STAR framework, which couples transfer learning and noisy student training to jointly improve classification performance and reduce computation overhead. In Section 4, we present experimental results to demonstrate the competitive performance of our approach on both binary and multi-class classification tasks. Finally, we conclude in Section 5 and point out future work directions.

2 RELATED WORK

Semi-supervised learning refers to the task of learning a prediction rule from a small amount of labeled data and a large amount of unlabeled data in order to improve model performance. It aims at bridging the gap between unsupervised learning, which uses unlabeled training data, and supervised learning, which uses labeled training data. Convolutional neural networks (CNNs) have become the de facto model for semi-supervised learning in image classification tasks. Sohn *et al* [12] introduce a semi-supervised learning method that combines pseudo-labeling and consistency regularization. Pseudo-labeling effectively uses the model's class prediction as a label to train against, while consistency regularization assumes that a model should output similar predictions when fed perturbed versions of the same image. Xie *et al* [17] present self-training with noisy student or noisy student training, which extends the idea of self-training and distillation with the use of equal-or-larger student models and noise added to the student during learning which has achieved state-of-the-art results on ImageNet [3]. Yalniz *et al* [18] proposed a learning method based on the teacher/student paradigm which leverages large collection of unlabeled images. Their method is not only improves standard architectures for image classification, but also improves performance for video classification and fine-grain recognition. SSL methods based on generative adversarial networks (GANs) have also received considerable attention [2, 9]. Dai *et al* [2] propose a semi-supervised learning method based on GANs. The method uses generated data to improve performance for the desired task. A key finding in their study is that a bad generator improves generalization. Li *et al* [9] consider GAN-based semi-supervised learning by comparing methods such as Triple GAN and Bad GAN. A generative network is trained on a set of data to produce or replicate similar examples using a generator and a discriminator. The goal of these methods is to predict the labels for unlabeled data, while generating new samples conditioned on these labels. They find that both methods can be used for different purposes. Triple GAN generates good image and label pairs, which can be used to train a classifier, whereas Bad GAN generates samples that force a shift in the decision boundary between the data manifold of the different classes.

Although this burgeoning literature has provided many useful insights, several gaps remain between model architecture design and training on image datasets at different scales (e.g. samples in thousands or billions). Recent state-of-the-art methods rely on unlabeled data at a scale of billions to work well. This becomes infeasible for tasks that can leverage unlabeled image data on a scale of thousands in terms of runtime, memory and data acquisition. In our work, we target the limitations of existing deep SSL approaches and aim to develop an integrated framework for visual classification by leveraging unlabeled data on a scale of thousands in order to learn better classification models.

3 METHOD

Problem Statement. Given the labels of a small subset of an image dataset, the objective of semi-supervised learning is to leverage unlabeled image data to improve model performance. More specifically, let $\mathcal{D}_\ell = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_\ell}$ be the set of labeled images $\mathbf{x}_i$ with associated known labels $y_i \in \mathcal{Y}_\ell$, and $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=N_\ell+1}^{N_\ell+N_u}$ be the set of unlabeled image data, where $N_\ell + N_u = N$. Then, the problem of semi-supervised image classification is to learn a classifier with the goal of predicting the labels of the set $\mathcal{D}_u$. It is important to note that for multi-class classification problems, the label of each image $\mathbf{x}_i$ in the labeled set $\mathcal{D}_\ell$ can be represented as a C-dimensional one-hot vector $\mathbf{y}_i \in \{0, 1\}^C$, where C is the number of classes.

We now present the main building blocks of our noisy semi-supervised transfer learning framework, which integrates transfer learning and noisy student training, as illustrated in Figure 1. Transfer learning is tailored for tasks that can leverage unlabeled data of a scale of thousands for visual classification, while noisy student training is

designed for training models at a very large scale (e.g. on ImageNet, noisy student training uses a teacher model to generate pseudo-labels for 300M unlabeled images). Our proposed STAR method first enables the teacher model to learn more effective representations with the knowledge from prior pre-training on ImageNet via transfer learning. This makes the student model learn even better representations than the teacher model by taking advantage of transfer learning, noise and the unlabeled data. The schematic layout and main steps of the proposed framework are illustrated in Figure 1.

Transfer Learning. Due to limited training data, it is standard practice to leverage deep learning models that were pre-trained on large datasets [19]. In our experiments, we use EfficientNets[16] pre-trained on ImageNet [3]. EfficientNets are a family of pre-trained convolutional neural networks that use a compound scaling method to uniformly scale the network width, depth, and resolution. We replace the final fully connected (FC) layer of the pre-trained model with a global average pooling (GAP) layer, which is widely used in classification tasks. It computes the average output of each feature map in the previous layer and helps minimize overfitting by reducing the total number of parameters in the model. GAP turns a feature map into a single number by taking the average of the numbers in that feature map. Similar to max pooling layers, GAP layers have no trainable parameters and are used to reduce the spatial dimensions of a three-dimensional tensor. The GAP layer is followed by a hidden layer and a single FC layer with a softmax function (i.e. a dense softmax layer of multiple units for the binary or multi-class classification case) that yields the probabilities of predicted classes. We train the models in two stages. First, we replace and train the newly added layers. Then, we also fine-tune the weights of the pretrained model by continuing the backpropagation. In the second stage, we only fine-tune the last two blocks of the pretrained model.

Noising Student. We use data augmentation and dropout as a form of input noise when training the student models. This strategy is usually carried out to improve generalization performance in classification tasks [4, 21]. It is often done by creating modified versions of the input images in a dataset through random transformations, including horizontal and vertical flip, Gaussian noise, brightness and zoom augmentation, horizontal and vertical shift, sampling noise once per pixel, color space conversion, and rotation. For model noise, we use dropout [13], which can be viewed as a way of regularizing a deep neural network by adding noise to its hidden units. Applying input noise and model noise to the unlabeled data enables the student model to treat the labeled and unlabeled data as a single distribution. Also, data augmentation enforces the student model to make consistent predictions across augmented versions of an image. While the teacher model generates pseudo-labels for clean images, the student model is trained to predict those labels for an augmented image as input. In other words, the student model is forced to retain the same category for the augmented image as the original image. When using dropout, the layer drops the neuron connection with a certain probability. This makes the teacher model like an ensemble of models when it generates pseudo-labels. We find that the effects of noising the student corroborate with the findings reported in [17].

Algorithm. The main algorithmic steps of our approach are summarized in Algorithm 1. The input is a set of labeled and unlabeled images for a particular visual classification task. The goal is to use the labeled and unlabeled data for the underlying task. A supervised learning model (i.e. teacher model), which has been pre-trained on ImageNet, is trained on the labeled images by minimizing the cross-entropy loss. Then, the teacher model is used to generate hard pseudo-labels (one-hot encodings) for the unlabeled data. Using both the labeled and pseudo-labeled data, a larger pre-trained model (e.g. student model) is trained to minimize the cross-entropy loss. During this training process, data augmentation and dropout are used as a form of input noise and model noise.

Algorithm 1 Noisy semi-supervised transfer learning (STAR)

Input: Labeled set $\mathcal{D}_\ell = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_\ell}$ of images $\mathbf{x}_i$ with associated known labels y_i, and unlabeled set $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=N_\ell+1}^{N_\ell+N_u}$ for a task $\mathcal{T}$.

Output: Learned parameters of the student model for task $\mathcal{T}$.

1: Initialize a supervised teacher model f_t with ImageNet weights.
2: Train f_t on the labeled image set by minimizing the cross-entropy loss.
3: Use f_t to generate pseudo-labels for the unlabeled image set.
4: Initialize a **larger** supervised student model f_s with ImageNet weights.
5: Train f_s with added **noise** on the combined labeled and pseudo-labeled images by minimizing the cross-entropy loss.

4 EXPERIMENTS

In this section, we conduct extensive experiments to assess the performance of the proposed visual classification framework in comparison with state-of-the-art methods on several datasets. The source code to reproduce the experimental results is made publicly available on GitHub[1].

4.1 Experimental Setup

Datasets. The summary descriptions of the datasets used to demonstrate and analyze the performance of the proposed STAR method are as follows:

- **Labeled datasets:** We conduct experiments on the Sport-or-not, Yoga-Pose, and Sport datasets. The Sport-or-not dataset is comprised of two classes with around 14K images being either of people practicing or not practicing sport. The Yoga-Pose dataset consists of almost 3K images of 18 different popular yoga poses such as *bridge*, *lotus*, and *tree*. The Sport dataset consists of 155 classes of popular sports such as *axe-throwing*, *basketball*, *cricket*, *kayaking* and many more. The labeled datasets are split into training, validation and test sets. For the Yoga-Pose dataset, the test set consists only of images that are in general harder to classify. A separate set is used as an additional test set, termed Test 2, only for the Sport-or-not dataset, as the classification task is much easier

[1] https://github.com/Decathlon/decavision

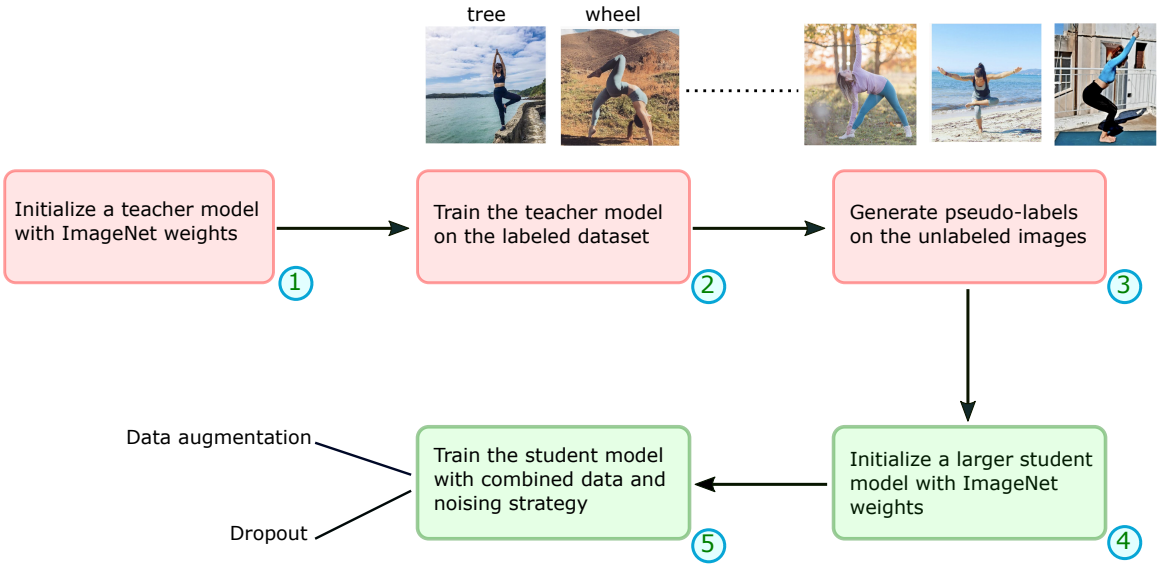

Figure 1: Schematic layout of the proposed noisy semi-supervised transfer learning (STAR) framework.

and less ambiguous. All datasets consist of training sets that are balanced (except for the Sport dataset) and test sets that are almost class balanced. Dataset statistics are summarized in Table 1.

- **Unlabeled datasets:** We obtain two sets of unlabeled data termed Unlb-700 and Unlb-120 from different sources, consisting of over 700K and 120K images, respectively. While the labeled data are collected from Instagram and Google Images, the unlabeled data are collected from Instagram using hashtag *decathlon* for sport images and various hashtags for yoga poses, and Getty Images for yoga poses using appropriate search keywords. Although these images are acquired using search keywords or labels, we ignore the labels and treat them as unlabeled data. For Sport-or-not and Sport classification tasks, we use the Unlb-700 dataset. For Yoga-Pose classification task, we use the Unlb-120 dataset.

Table 1: Dataset statistics for Sport-or-not, Yoga-Pose and Sport classification.

		Number of samples			
Dataset	Classes	Train	Validation	Test 1	Test 2
Sport-or-not	2	12,366	1,002	1,002	1,000
Yoga-Pose	18	2,855	196	214	-
Sport	155	74,961	12,830	3,591	-

Prior to training all models, the training and validation sets are converted to the TFRecords format with the desired input size (i.e. $299 \times 299 \times 3$) through image resizing. In order to achieve faster convergence, feature standardization is usually performed, i.e. we rescale the images to have values between 0 and 1. More specifically, given a data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, the standardized feature vector is given by

$$\mathbf{z}_i = \frac{\mathbf{x}_i - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)}, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i$ is the i-th input data point, denoting a row vector. It is important to note that in our approach, no post-processing is employed other than resizing and feature standardization.

Baselines. We evaluate the performance of the proposed method against EfficientNets [16] and Noisy Student Training [17]. A brief description of these state-of-the-art baselines can be summarized as follows:

- **EfficientNets** As the name suggests, EfficientNets is a group of computationally efficient convolutional neural networks that are trained on the ImageNet database, and they use a compound scaling method to uniformly scale the network width, depth, and resolution with a set of fixed scaling coefficients [16]. This group consists of eight models with configurations ranging from the baseline model EfficientNet-B0 to the larger model EfficientNet-B7, where each subsequent architecture refers to a model variant with more parameters and higher accuracy. Depth is the number of layers in the network, width is the number of filters in a convolutional layer, and resolution is simply the height and width of the input image. The main building block of EfficientNets is the mobile inverted bottleneck convolution (MBConv) together with a squeeze-and-excitation optimization.
- **Noisy Student Training (NST)** is a semi-supervised learning method for visual classification [17]. The method consists of two main stages. First, a supervised model is trained from scratch on the labeled images, and then then used as a teacher model to generate the pseudo-labels for the unlabeled data. Second, a larger model, known as student, is trained from scratch on the combined of labeled and pseudo-labeled image data. Noise is injected into the student model in the

What is the NST?
It describes here.

form of dropout and data augmentation with the goal of achieving better generalization than the teacher model. This process is then repeated until a desired level of classification performance is achieved. Unlike NST in which the student becomes the teacher durring the iterative training process, our STAR approach uses pre-training for both teacher and student models.

Implementation Details. All experiments are performed using the DecaVision library on Linux workstations running 4.8Hz and 64GB RAM with NVIDIA RTX 2080Ti and RTX 3080 GPUs. All models are based on EfficientNets [16], particularly EfficientNet-B3, and larger models such as EfficientNet-B5 and EfficientNet-B7 which have more learnable parameters to fit large number of unlabeled data. Pretrained ImageNet [8] models are trained for our tasks using the Adam optimization algorithm [7] to minimize the binary or categorical cross entropy losses, depending on the binary or multi-class setting. A factor of 0.1 is used to reduce the learning rate once the loss stagnates. Training is continued until the validation loss stagnates using an early stopping mechanism. For Sport-or-not and Yoga-Pose classification tasks, we use hard pseudo-labels (one-hot encoding). For the Sport classification task, we use all unlabeled data that has a confidence score higher than a threshold equal to 0.3.

The process of finding Hyper-parameters

Due to the nature of deep learning algorithms, we train every model in two stages. In the first stage, we perform a hyperparameter optimization process using the hypertuning feature in the DecaVision library, which is inspired by the scikit-optimize library[2]. We start by training a model 10 times with random combinations of hyperparameters, which are predefined in the search space (i.e. hidden size, learning rate, whether to skip or not the fine-tuning phase, learning rate in the fine-tuning phase, etc.). Then, we use what has been learned from these random combinations to find 15 better ones. In total, a single model configuration goes through 25 iterations of hyperparameter search to find the best model configuration possible. In the second stage, we train a new model configuration using the optimal hyperparameters. This stage starts by training an extra layer or layers, depending on the hyperparameter optimization, on top of the frozen pretrained model, followed by fine-tuning few blocks of the pretrained model by unfreezing them.

Evaluation metrics. In order to evaluate the performance of our proposed STAR framework against the baseline methods, we use average accuracy and F1 score as evaluation metrics. For the binary classification task (i.e. Sport-or-not classification), we measure performance using the average accuracy since the validation and test set are balanced, as shown in Table 1. For the multi-class classification tasks (i.e. Yoga-Pose and Sport classification), we report both average accuracy and per-class F1 score. For both metrics, a larger value indicates better classification performance.

4.2 Results and Analysis

In this subsection, we evaluate the proposed STAR approach on several datasets, demonstrating significant improvements over strong baseline methods.

[2]https://scikit-optimize.github.io/stable/

Sport-or-not classification results In Table 2, we report the classification results on the Sport-or-not dataset. The teacher model is the EfficientNet-B3 architecture trained on the labeled dataset. Pseudo-labels are then generated on the Unlb-700 dataset. This is followed by training an EfficientNet-B5 architecture on a combination of the labeled and pseudo-labeled datasets with data augmentation as noise. Both models were pre-trained on ImageNet. Notice that accuracy on the test set drops even after training on more than half a million images. This drop in accuracy is largely attributed to the fact that many images in the test set are very ambiguous and it is not easy for a human annotator to label such an image as *sport* or *not-sport*. A typical example is a picture of someone standing in front of the mountains, as shown in Figure 2 (left).

Figure 2: Example images of people not practicing sports, but have regions representative of sports and vice versa. It is unclear whether to label these images as *sport* or *not-sport* even by a human annotator.

Unclear Label's

Figure 2 (left) can be considered as *sport* if the person is mountain-climbing/hiking, or *not-sport* if the person is just taking a picture with mountains in the background. Figure 2 (center) shows a person not actually doing sport, but clearly was practicing sport moments earlier. Figure 2 (right) shows a person is biking, but not in the image. So the question is: can you tell if these images are *sport* or *not-sport*? This motivated us to build another unseen balanced test set, which we term Test 2, with only unambiguous images in an effort to properly evaluate our models. On the Test 2 set, the student model outperforms the teacher model by a relative improvement of 2.4% in terms of average accuracy. Also, our STAR approach yields a relative improvement of 1.36% over NST.

Table 2: Classification performance on the Sport-or-not datasets with two classes. For the teacher model, we train EfficientNet-B3 on the labeled training set. Both NST and our STAR method are trained on a combination of pseudo-labeled data, generated by EfficientNet-B3 on the unlabeled Unlb-700 dataset, and the labeled data. Boldface numbers indicate the best performance.

	Accuracy (%)		
Method	Validation	Test 1	Test 2
EfficientNet-B3 [16]	86.03	**86.58**	87.20
EfficientNet-B5 NST [17]	86.21	85.36	88.10
EfficientNet-B5 STAR	**86.53**	85.88	**89.30**

Yoga-Pose classification results. The overall performance results on the Yoga Pose dataset are summarized in Table 3. In this

setting, the teacher model is the EfficientNet-B5 architecture trained on the labeled data. Similar to the Sport-or-not dataset, we generate 120K pseudo-labels from the Unlb-120 dataset and train the EfficientNet-B7 architecture in a similar fashion. Notice that STAR yields significant performance improvements with almost a 5.88% relative improvement on the test set over the teacher model. Hence, the student model is able to make better predictions on both the validation and test set samples for individual classes, resulting in a much improved performance on the test set. This better performance is further illustrated in Figure 3, which shows the F1 score for each class of the Yoga-pose dataset on both validation and test sets. Figure 3 (top) shows the performance on the validation set, and we can see that EfficientNet-B7 STAR outperforms the teacher model by a large margin on classes such as *camel*, *chair*, *standing forward bend*, *warrior 1*, and *wheel*. We only see a drop in performance for the *child* and *twists* classes. In Figure 3 (bottom), the performance on the test set is shown. Notice that the largest improvement using EfficientNet-B7 STAR is seen for the *cobra* class with a F1-score of 100%, whereas EfficientNet-B5 yields a score of 50%. For the majority of the classes, a significant improvement is obtained using our STAR method compared to the baselines.

Table 3: Comparison of classification performance on the Yoga-Pose classification task. We train EfficientNet-B5 on the labeled training data. NST stands for Noisy Student Training. EfficientNet-B7 NST and STAR are trained with a combination of pseudo-labeled data, generated by EfficientNet-B5 on the unlabeled dataset, and the labeled data. Boldface numbers indicate the best performance.

	Accuracy (%)	
Method	Validation	Test
EfficientNet-B5 [16]	89.35	84.58
EfficientNet-B7 NST [17]	92.12	87.56
EfficientNet-B7 STAR	**93.06**	**89.55**

Sport classification results. Table 4 shows the comparison results with baselines on the Sport dataset. As can be seen, STAR outperforms both EfficientNet-B5 and EfficientNet-B7 NST baselines by relative improvements of 21.43% and 10.81%, respectively. Notice that the test accuracy of EfficientNet-B5 is low compared to the validation accuracy, due in large part to the fact that the test set is class-balanced and hence accuracy provides an unbiased representation of type I and type II errors. Also, notice the large gap between the accuracy values on the validation and test sets for the baselines. As for the proposed method the gap is much smaller, suggesting that there is less overfitting during the hyperparameter optimization process.

4.3 Adversarial Robustness

We also study the model performance against adversarial attacks. For each classification task (i.e. Sport-or-not, Yoga-Pose and Sport classification), we evaluate our best models both with and without NST or STAR against the fast gradient sign method (FGSM) attack [5], where the goal is to ensure misclassification. FGSM is a white box attack since it has complete access to the model being attacked. For a given input image, the method uses the gradient of the loss function with respect to the input image to create a new image that maximizes the loss, with the update on each pixel set to ϵ. The new image is called the adversarial image. We show an illustration of an FGSM attach in Figure 4, where the value of ϵ is set to 0.13.

Table 4: Comparison of classification performance on the Sport classification task. We train EfficientNet-B5 on the labeled training data. EfficientNet-B7 NST and STAR are trained with a combination of pseudo-labeled data, generated by EfficientNet-B5 on the unlabeled dataset, and the labeled data. Boldface numbers indicate the best performance.

	Accuracy (%)	
Model	Validation	Test
EfficientNet-B5 [16]	83.51	66.25
EfficientNet-B7 NST [17]	85.44	72.60
EfficientNet-B7 STAR	**89.28**	**80.45**

In Figure 5, we report the adversarial robustness results. As can be seen, our STAR method yields to significant improvements in terms of average accuracy despite the fact that no model optimization is performed for adversarial robustness. On the Yoga-Pose dataset, we observe the highest improvement of 3.1 percentage points over NST and 5% percentage points over the teacher model. For the Sport-or-not and Sport classification tasks, we see improvements of 2.7% and 0.89% over NST, and 4.1% and 0.98% over the teacher model.

4.4 Ablation Study

In this section, we study the importance of transfer learning in terms of runtime and memory, labeled data, and unlabeled data.

Importance of transfer learning in self-training. Table 5 shows the importance of transfer learning in self-training. As can be seen, training student models using transfer learning not only yields better accuracy, but also reduces training time and computation overhead by significantly reducing the number of learnable parameters. Notice that using transfer learning in self-training requires 6x less training time over the baseline. The baseline method here is similar to NST, where the networks are trained from scratch. We believe that our approach is more practical for the industrial perspective in the sense that we can quickly develop models that can leverage unlabeled data in a semi-supervised fashion.

Effect of varying labeled data in self-training. In order to understand the amount of required labeled data to achieve optimal performance, we vary the amount of labeled data available in the self-training process. As shown in Table 6, the performance of the supervised EfficientNet-B5 model suffers a relative drop of 8.5% when using only 1,000 samples for training. Interestingly, for EfficientNet-B7 STAR, even though there is a drop in performance, the classification performance is higher when using only 1,000

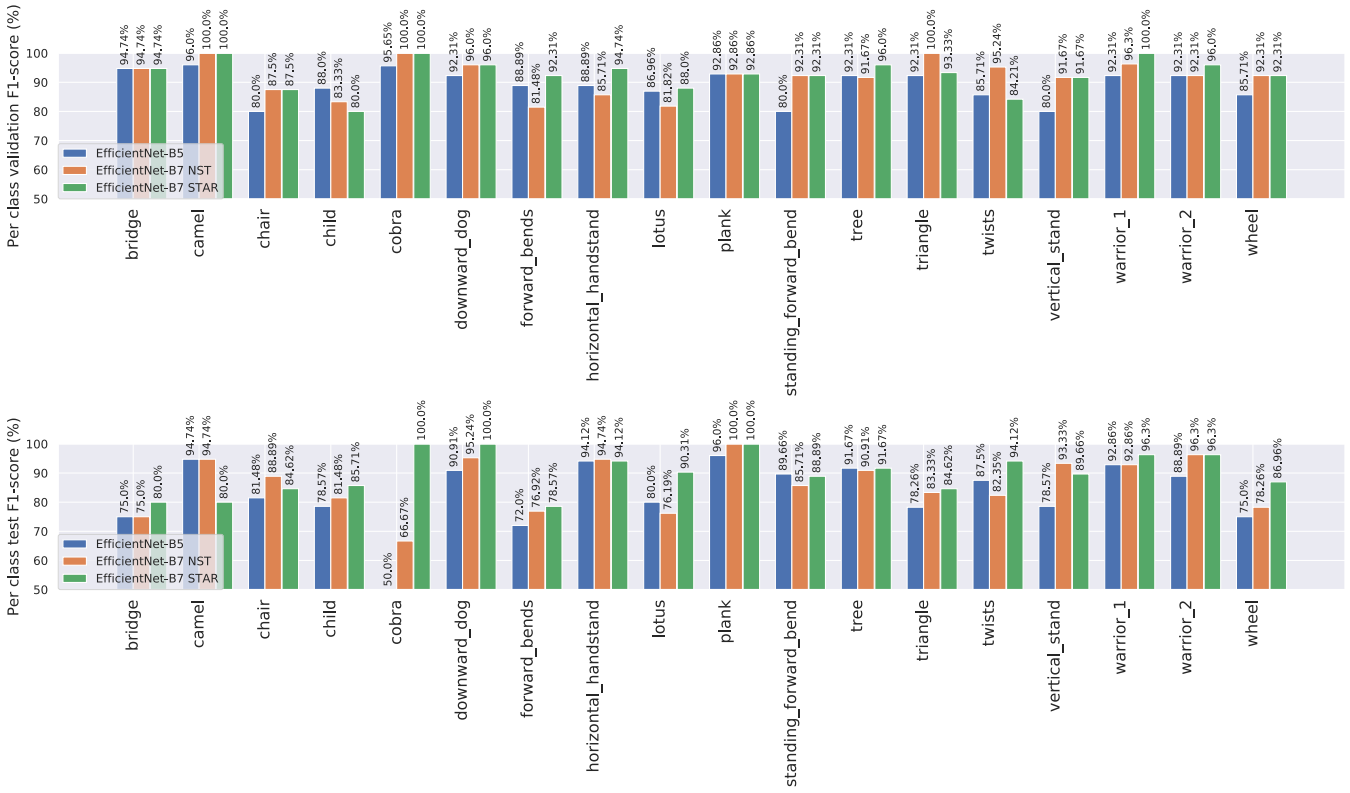

Figure 3: Comparison of validation(top) and test (bottom) classification performance on the Yoga-Pose dataset for each class. EfficientNet-B5 is trained on the available training data, while EfficientNet-B7 NST and STAR are trained on a combination of the available labeled training data and generated pseudo-labels data by EfficientNet-B5.

Figure 4: Input image (left), perturbed image using FGSM attack (center), and resulting adversarial image (right). An $\epsilon = 0.13$ is used to ensure perturbations are small.

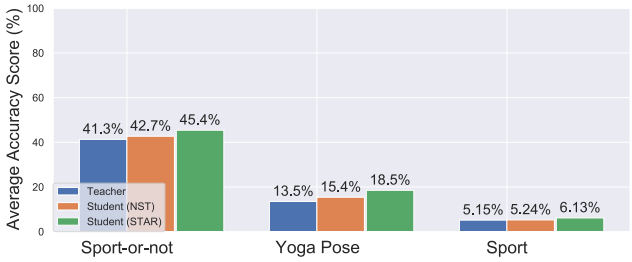

Figure 5: Adversarial robustness results with $\epsilon = 0.13$. STAR improves adversarial robustness against white box FGSM attack even though we do not optimize the model for adversarial robustness.

samples for training, where EfficientNet-B5 uses all available labeled data for training. This shows that in data scarcity scenarios, our STAR method can leverage unlabeled data to achieve optimal performance for the desired task.

Effect of varying pseudo-labels in self-training. We study how the number of pseudo-labels generated from the unlabeled samples can affect classification performance in self-training. In Figure 6 (top), we can see that as the number of pseudo-labels grows, the classification performance shows an increasing trend in both validation and test sets. Moreover, adding more pseudo-labels would potentially further improve performance since the trend has not saturated yet.

In Figure 6 (bottom), we show how varying the threshold value affects performance. A threshold of 0.7 means that if the predicted label is below such a threshold, it is discarded from the pseudo-labeled dataset. Notice an upward trend when decreasing the threshold. This trend can be attributed to the fact that even though increasing the threshold may lead to higher quality pseudo-labeled data, more images are being discarded, resulting in a drop in performance.

Table 5: Ablation study of transfer learning on the Sport-or-not dataset. For the two cases, we train the EfficientNet-B5 using the pseudo-labels from Unlb-700 and 14K labeled images with and without ImageNet pretraining. TL stands for transfer learning and size refers to the number of learnable parameters. Boldface numbers indicate the best performance

Method	Seconds per Epoch	Size
EfficientNet-B5 without TL [17]	250	30.6M
EfficientNet-B5 with TL (Ours)	**40**	**5.5M**

Table 6: Ablation study of labeled data for the Yoga-Pose classification task. We train the supervised classifier EfficientNet-B5 on the labeled data (1K, 2K, All), where All represents the entire training dataset. EfficientNet-B7 STAR is then trained on a combination of the labeled data and 120K pseudo-labeled data generated by EfficientNet-B5 from Unlb-120. Boldface numbers indicate the best performance.

	Test set accuracy (%)	
Labeled set size	EfficientNet-B5	EfficientNet-B7 STAR
1K	78.64	85.33
2K	81.52	87.24
All	**84.58**	**89.55**

Moreover, we found using hard pseudo-labels (i.e. *none*) on the unlabeled data yields better results compared to using soft pseudo-labels. This is also in line with findings in [17], where soft pseudo-labels are shown to work better for out-of-domain unlabeled images.

In Figure 7, we plot the raw probability scores by the teacher model on the unlabeled images used for the Yoga-Pose classification task. It is important to note that the plot shows that most of the pseudo-labels have high confidence due to a good teacher model. This is also attributed to the fact that the unlabeled images are acquired using search keywords or labels, which we ignored and treated them as unlabeled data. Hence, these images are somewhat relevant and not out-of-domain for the given task, and leaving relevant images out would degrade performance, as shown in Figure 6.

5 CONCLUSION

In this paper, we introduced STAR, a novel semi-supervised learning method that combines transfer learning and self-training with noisy student. The proposed framework is well suited for leveraging unlabeled images on a scale of thousands, and is efficient in terms of both runtime and memory, without compromising on accuracy. Extensive experiments show that STAR outperforms state-of-the-art methods, especially for multi-class classification tasks. Ablation studies also showed that leveraging transfer learning in the training process not only improves performance in terms of accuracy, but also requires 6x less compute time and 5x less memory. Moreover, we showed that STAR boosts robustness in visual classification models without specifically optimizing for adversarial

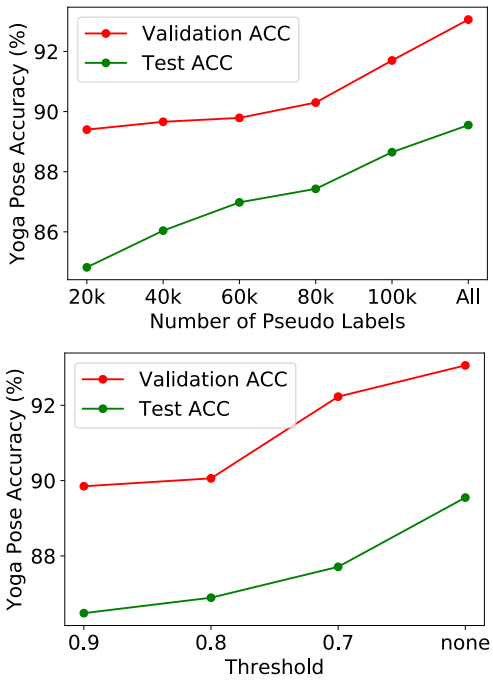

Figure 6: Sensitivity analysis of EfficientNet-B7 STAR on the Yoga-Pose classification task by varying the number of pseudo-labels (top) and threshold (bottom). *none* means that the lowest confidence predictions are taken in the pseudo-labeled data. In this case, all unlabeled images are part of the pseudo-labeled dataset.

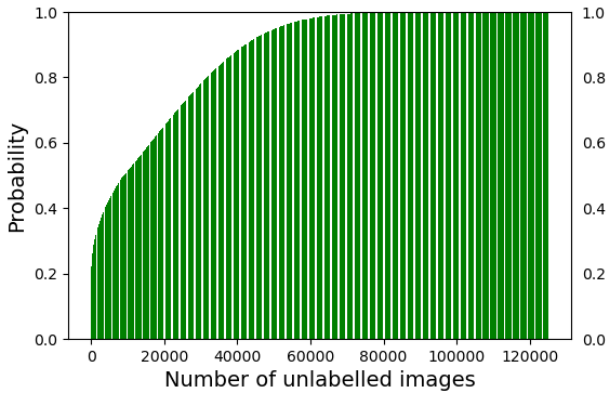

Figure 7: Confidence scores for the teacher model on the Unlb-120 dataset for Yoga-Pose classification task. For the majority of the images, we obtain high confidence scores, indicating that images are relevant to this task and not out-of-domain.

robustness. While our focus in this work was on binary and multi-class classification use-cases, the proposed STAR method can be easily extended to other downstream tasks such as segmentation and object detection, which we intent to explore as future work.

REFERENCES

[1] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proc. 11th Annual Conference on Computational Learning Theory*. 92–100.

[2] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan Salakhutdinov. 2017. Good semi-supervised learning that requires a bad GAN. In *Advances in Neural Information Processing Systems*.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[4] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *Proc. IEEE International Symposium on Biomedical Imaging*. 289–293.

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*. http://arxiv.org/abs/1412.6572

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[9] Wenyuan Li, Zichen Wang, Jiayun Li, Jennifer Polson, William Speier, and Corey W Arnold. 2019. Semi-supervised learning based on generative adversarial network: a comparison between good GAN and bad GAN approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[10] Wei Liu, Junfeng He, and Shih-Fu Chang. 2010. Large graph construction for scalable semi-supervised learning. In *International Conference on Machine Learning*.

[11] H Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.

[12] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[16] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. 6105–6114.

[17] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves ImageNet classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 10687–10698.

[18] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).

[19] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems*.

[20] Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey. (2005).

[21] Hasib Zunair and A Ben Hamza. 2020. Melanoma detection using adversarial training and deep transfer learning. *Physics in Medicine & Biology* 65, 13 (2020).