Semantic Clustering based Deduction Learning for Image Recognition and Classification

Wenchi Ma[1], Xuemin Tu[2], Bo Luo[1], Guanghui Wang[3]

Abstract

The paper proposes a semantic clustering based deduction learning by mimicking the learning and thinking process of human brains. Human beings can make judgments based on experience and cognition, and as a result, no one would recognize an unknown animal as a car. Inspired by this observation, we propose to train deep learning models using the clustering prior that can guide the models to learn with the ability of semantic deducing and summarizing from classification attributes, such as a cat belonging to animals while a car pertaining to vehicles. The proposed approach realizes the high-level clustering in the semantic space, enabling the model to deduce the relations among various classes during the learning process. In addition, the paper introduces a semantic prior based random search for the opposite labels to ensure the smooth distribution of the clustering and the robustness of the classifiers. The proposed approach is supported theoretically and empirically through extensive experiments. We compare the performance across state-of-the-art classifiers on popular benchmarks, and the generalization ability is verified by adding noisy labeling to the datasets. Experimental results demonstrate the superiority of the proposed approach.

Keywords: Deduction learning, clustering prior, semantic space, smooth

[1] W. Ma and B. Luo are with the Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS, 66045 USA e-mail: wenchima@ku.edu; bluo@ku.edu.

[2] X. Tu is with the Department of Mathematics, The University of Kansas, Lawrence, KS, 66045 USA e-mail: xuemin@ku.edu.

[3] G. Wang is with the Department of Computer Science, Ryerson University, Toronto, ON, M5B 2K3 Canada. e-mail: wangcs@ryerson.ca

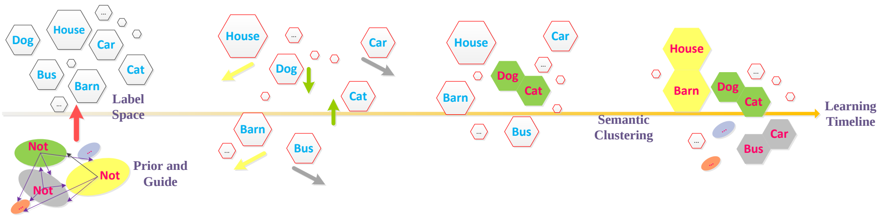

Figure 1: The deduction progress of semantic clustering. Prior and Guide works as the prior information that is combined with the labels as the labeling input data. The Learning Timeline is the same as the normal classification learning process, but our model provides the possibility of doing high-level semantic clustering by the deduction progress as the aid for the classification task. The model, at the end of the learning timeline, is expected to provide better classification accuracy.

semantic clustering.

1. Introduction

The powerful ability for feature expression and semantic extraction of deep Convolutional Neural Networks (CNNs) has dramatically pushed the flourishing development of computer vision [1] [2] [3]. At the same time, large-scale labeled data samples ensure the effectiveness of supervised learning, which enables the deep learning models to efficiently extract abstract but highly-semantic information for complicated vision tasks [4] [5] [6]. Undoubtedly, future learning models should be complex, robust, knowledge-driven, and cognition-based [7] [8]. This defines them with the cognitive ability of self-enhancing, synthesizing knowledge from multiple sources, and deducing based on knowledge and experiences [7].

Some complementary and weak supervision information has been exploited to boost the learning performance of models [9] [10]. Such complementary supervision includes early side information[11], privileged information [12], and weak supervision based on semi-supervised data [13] [14], noisy labeled data [15] [16], and complementary labels [17] [18] [19]. Most of these methods supplement extra direct labeling information or replace expensive accurate labels with cheap labeling information. These complementary labels, in fact, increase the labeling cost as a direct mapping from label space to sample space, named as "hard labeling" in the later sections. Most importantly, these methods are unable to equip deep models with the ability of self-enhancement, synthesis, and deduction.

In this paper, we leverage the wide-applied but fundamental supervised image classification and propose deduction learning by semantic clustering. We introduce semantic prior, high-level clustering information, represented by different colors in Figure 1, although no names are given for each color. For example, we expect the model know the cat and the dog should be closed to each other, though the model would never know they should be called "animal". Semantic prior (Prior and Guide to Label Space in Figure 1) is thus introduced into the classification learning models, guiding them to form effective semantic clustering so that they are able to deduce high-level semantic expression (e.g. Same color cells go attached together in Figure 1), as shown in Figure 1.

Inspired by the idea of negative learning [19] [18], we propose to guide the classifier to learn the opposite class that does not belong to the same cluster with the accurate label. For example, if a sample is labeled by "cat", then our algorithm will tell the classifier that the image is not "car" or any other random label that belongs to a different cluster other than "cat", during one learning shot.

This random search for the opposite label is in accordance with the semantic prior that is fed into the model along with other inputs that specifically refer to the images and their corresponding labels in this work. Statistically, the opposite semantic labels corresponding to a certain accurate label should be chosen with equal probability given the number of learning periods (epochs) is large enough. Theoretically, this proposed method enables a smooth clustering in the semantic space and an effective deduction, which makes the model able to deduce that "cat" should be one element of an abstract cluster, although the model would never know it can be called "animal", as shown in the second stage of Figure 1, where the colors "Green, Grey, Yellow", each represents a higher hierarchical category. Each specific class, like the cat, would be learning that if it belongs to “Green”, then it is totally on the opposite side of other classes that belong to “Yellow” and “Grey”.

Finally, it is noticed that the proposed method does not give up the conventional label learning by introducing one composite loss function. This ensures the label learning and the semantic clustering in the same timeline during the learning process. It conforms to the requirement of cognition learning [7]. By this setting, the model could finish high-level semantic expressions, capturing the concepts, similar to “animal”, “vehicle”, “buildings”, etc., as shown in the third stage in Figure 1, where sample classes accomplish clusters.

The major contributions of this paper are summarized below:

- Semantic Clustering: We propose a high-level semantic mapping within semantic space, enhancing the semantic expression and providing a certain level of independence for overcoming the limitation of convolution operation at the pixel level. It is realized by introducing a semantic prior which could guide the model to find the opposite semantic label that is not from the same semantic colony with the given true label.

- Deduction Learning: Deduction learning is realized by the semantic prior and the proposed random search for opposite semantic, which ensures the smoothness of semantic clustering and the robustness of classification. It could be implemented as a plug-in module that could play in arbitrary classification models by introducing a composite loss function.

- Robust Improvement: We achieved stable convergence and robust classification performance on mainstream classification models. It is also verified by working on noisy data environment where there exists a certain ratio of incorrect labels.

- Wide Applicability: In the proposed method, label learning and semantic clustering follow the same learning timeline, equipping the model with the ability of deduction and cognition. It can be taken as a plug-in module for broad deep learning applications, such as few-shot learning, zero-shot learning or even semi-supervised learning.

The functional source code of the paper can be accessed from the link https://github.com/rucv/deduction-learning.

2. Related Work

2.1. Hierarchical Semantic Information

At first, the research in this field focuses on exploring or utilizing the inherent relations among label classes, or looking for the intermediate representations between classes. [20] formed a label-embedding problem where each class is embedded in the space of attribute vectors so that the attributes act as intermediate representations that enable parameter sharing between classes. Another research in [21] uses a label relation graph to encode flexible relations between class labels by building the rich structure of real-world labels. The idea of incremental learning by hierarchical label training has been explored recently by a few other papers. Progressive Neural Networks [22] learn to solve complex sequences of task by leveraging prior knowledge with lateral connections. "iCaRL" allows learning in a class incremental way: only the training data for a small number of classes is present at the same time and new classes can be added progressively [23]. Tree-CNN [24], proposes training root network by general classes and then learning the fine classes by corresponding growth-network (mainly learned by leaf structure of the network). While this research direction solves hierarchical semantic learning based on an independent timeline for each stage. Our proposed idea shares the same timeline with the normal classification task throughout the entire learning process which works as an exploration towards cognitive learning. At the same time, the methods above directly provide concrete class relation structure on the basis of the original class labels for training, without exploring the deduction ability of the networks.

Learning with real, concrete complementary labeling information was proposed by [17] for the image classification task. It was based on an assumption that the transition probability for complementary labels is equal to each other.

It modified the traditional one-versus-all (OVA) and pairwise-comparison (PC) losses so that it is suitable for the uniform probability distribution, working as an unbiased estimator for the expected risk of true-labeled classification. Later on, the work [18] argued that there are two unsolved problems in the previous work. The first one lies in the fact that the complementary labels tend to be affected by annotators' experience and limited cognition. The other one is the proposed modified OVA and PC losses can not be generalized to more popular losses, such as the cross-entropy loss. Thus, they proposed the transition matrix setting to fix the bias from the biased complementary labels. At the same time, they provided intensive mathematical analysis to prove their proposed setting can be generalized to many losses which directly provides an unbiased estimator for minimizing expectation risk. These works expect better semantic learning by introducing intensive complementary labeling while they do not explore the deduction ability of the networks themselves as well. They are essentially regular label learning. The work in [19] automatically generated complementary labels from the given noisy labels and utilized them for the proposed negative learning, incorporating the complementary labeling into noisy label learning.

2.2. Semantic Labeling in Noisy Cases

Some researchers attempt to aid learning in noisy cases by introducing effective semantic label learning. Some attempt to create noise-robust losses by introducing transition probabilities to the field of classification and transfer learning [25] [26]. Some propose to use the transition layer to modify deep neural network [27]. In other studies, researchers try to re-weight the training sample based on the reliability of the given label [28] [29]. Some other approaches try to prune the correct samples from the softmax outputs [30] [31]. Different from them, this paper dedicates to the research on how self-clustering and deduction learning ability of networks would influence the robustness in noisy labeling cases.

This paper tries to explore the self-deduction ability of networks in the semantic space and focuses on guiding the models to fetch effective hierarchical semantic information in a self-learning way by semantic clustering and cognitive accumulation. First, it could completely free the confinement problem of transition probabilities. The proposed semantic prior based random search for opposite semantic ensures the equal probability, providing the mapping independence in semantic space. Second, the semantic clustering boosts positive label learning. For example, if the sample “cat” has a low classification probability, the semantic clustering could help enhance this confidence by guiding this model to realize that the object is at least an animal, not a “car”. Third, our proposed method shares the same timeline with conventional label learning, enabling effective cognitive accumulation. Moreover, there is no need for specifically defining loss functions for the proposed models. Following the loss formations of the original label learning in specific models is all we need, potentially leading to better generalization.

3. Problem Setup

People can make deduction independent of the actual vision behavior. Thus, in deep learning, we expect the model with similar independence to ensure the realization of high-level mapping in semantic space.

Semantic Space for Image Classification. Semantic space is originally proposed in the natural language domain, aiming to create representations of natural language that are capable of capturing meanings [32]. In computer vision, the concept of semantic space is much more abstract. Current semantic extraction is limited both by spatial size and by the individual data sample. However, it should aim to overcome the limitations of convolution-based or receptive-field based approaches operating at the pixel level. Convolution-based deep learning models are fixed at the pixel level and are poor for generalization, which would easily break down if the individual image differs from or is strange to those in the training materials used for the statistical models. Compared to spatial feature learning that performs at the pixel level, semantic learning should be a relatively independent process that works on the semantic element, which is the common description for a class of objects. Moreover, the semantic expression could have multi-levels that describe the relevant or diverging characters of semantic elements. For example, the "cat" as a semantic element could be clustered to the high-level semantic expression, something similar to an "animal".

Definition 1 (Semantic Space)**.** *Without loss of generality, let $\mathcal{C}$ be the semantic space, $c \in \mathbb{Z}^+$ be the semantic element in $\mathcal{C}$ that appears as one semantic label indicating a specific object class. The semantic relation of different c is defined by r. $[c] = \{1,...,c\}$ signifies the set of semantic labels. Then, we have*

$$\mathcal{C} \overset{def}{=} \langle [c], r \rangle \tag{1}$$

where element c is uniformly sampled from $\mathcal{C}$. Tuple $\langle [c], r \rangle$ expresses the fact that semantic elements $c \in [c]$ are linked to each other by the relation r, forming the abstract spatial distribution in $\mathcal{C}$.

Semantic Cell. In order to better describe the abstract relation distribution in $\mathcal{C}$, we propose *Semantic Cell* as the semantic unit that could label a group of objects that have similar features in feature space $\mathcal{X}$, which corresponds to the element $c \in [c]$ in Definition 1. It realizes a multi-to-one mapping that bridges the link between feature space $\mathcal{X}$ and semantic space $\mathcal{C}$.

Definition 2 (Semantic Mapping)**.** *Let $g(\mathbf{x})$ be the mapping function of a given multi-class classification learning model that estimates the classification probabilities based on the input sample $\mathbf{x}$ in feature space $\mathcal{X}$. $f(\mathbf{x})$ predicts the classification label y based on the maximum probability principle, mapping the feature sample $\mathbf{x}$ to the corresponding semantic cell c in $\mathcal{C}$.*

$$f(\mathbf{x}) \overset{def}{=} \arg\max_{i \in [c]} g_i(\mathbf{x}) \tag{2}$$

where $f : \mathcal{X} \to \mathcal{C}$, the maximum probability of g and $f(\mathbf{x}) \in \mathcal{C}$. $g_i(\mathbf{x})$ realizes the estimation towards $P(y = i|\mathbf{x})$.

Semantic Colony. Semantic Colony θ takes semantic cell c as individual sample. It clusters $c \in \mathcal{C}$ that hold related semantic information as θ. Based on which, it defines the intra-class relation and inner-class differentiation to realize clustering in semantic space $\mathcal{C}$ with high-order semantic expression.

Definition 3 (Semantic Clustering)**.** *Without loss of generality, let Θ be the distribution of semantic colonies θ in $\mathcal{C}$. H conducts clustering for semantic cell $c \in \mathcal{C}$ into semantic colony $\theta \sim \Theta$. $\mathbf{c}$ is the vector with the elements of semantic cells $c \in [c]$. Then, we have*

$$\theta \overset{def}{=} H(\mathbf{c}, r_{\mathbf{c}}) \tag{3}$$

where $H : [c] \to \Theta$, $\mathbf{c}$ consists of semantic cells c in [c] that are semantically related, and H maps $\mathbf{c}$ to $\theta \sim \Theta$ in accordance with the corresponding semantic relation $r_{\mathbf{c}}$.

4. Methodology

In this section, we first introduce the general approach that deep neural networks learn optimal classification with hard labels. Then, we discuss the learning with semantic deduction and propose corresponding training and test model.

4.1. Conventional Classification Learning

In multi-class classification, we aim to learn a classifier $f(\mathbf{x})$ that predicts the classification label y for a given observation sample $\mathbf{x}$. Typically, the classifier directly maps $\mathbf{x}$ into the label space $\mathcal{Y}$ by the following function:

$$f(\mathbf{x}) = \arg\max_{i \in \mathcal{Y}} W_i^T \mathbf{x} \tag{4}$$

where $f : \mathcal{X} \to \mathcal{Y}$ and W_i refers to the learning parameters of the classifier f, with the estimation of $P(y = i|\mathbf{x})$.

In supervised learning, loss functions are proposed to measure the expectation of the predicting $f(\mathbf{x})$ for y [33]. It is typically defined as the expected risk [18] for various loss functions.

$$R(g) = \mathbb{E}_{\mathbf{x}, y \sim P(\mathbf{x}, y)}[\ell(f(\mathbf{x}), y)] \tag{5}$$

A well-trained classifier f^* minimizes this expected risk $R(g)$,

$$f^* = \arg\min_{f \in \mathcal{F}} R(f) \tag{6}$$

where $\mathcal{F}$ is the distribution space of f.

4.2. Learning with Semantic Deduction

In semantic space, the description of hard labels towards objects is limited. To better describe an object or a scene, people usually enumerate related features and associate their prior cognition and experience for a reasonable deduction. Current deep learning models realize feature sensing and learning but lack the proper deduction that could enrich the description of objects. Our previous analysis shows that hard labels in semantic space could potentially build more links, as the discussion in Section 2.1. We introduce the semantic prior, guiding the model to learn the semantic links by deduction. The overview of our method is depicted in Figure 2. The overall inputs include training sample images, corresponding labels, and the semantic prior information which provides the high-level semantic hierarchy of current classification labels. The classification model is trained in the same way as the original network. For the green part in Figure 2, given label y, the model finds the corresponding opposite semantic label for the sample image according to the semantic prior by an equal-probability random search, shown as the yellow block. Then both the true label and the opposite semantic label are fed into the composite loss we defined. The output of the proposed method is expected of better classification performance in the way of classification accuracy.

First, the semantic prior works as the criterion for colonies' formation in semantic space $\mathcal{C}$. For example, a cat labeled by $c_i \in [c]$ should be grouped into "animal" colony, if denoted by θ_m. Similarly, a car labeled by c_j could be grouped into the "vehicle" colony θ_n. Second, the semantic deduction is fully performed in semantic space $\mathcal{C}$, instead of defining complementary labels as weak supervision. Thus, we do not need any tedious and laborious labeling work, which would avoid labeling bias from human beings' bias [18], and the

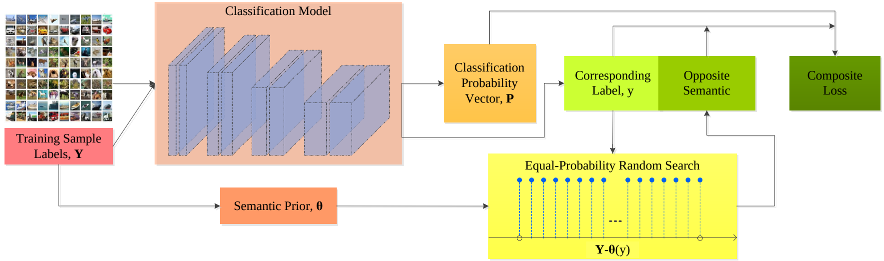

Figure 2: An overview of the proposed method. We use semantic prior based random search to produce opposite semantic so as to form the composite loss function, guiding the model to form semantic colonies.

problem that the complementary labeling is essentially non-uniformly selected from the $c-1$ classes other than the true label class $(c>2)$.

4.3. Equal-Probability Search for Opposite Semantic.

We assume that the variables $(\mathbf{x}, c, \theta)$ are defined in the space $(\mathcal{X} \times [c] \times \Theta)$, with the joint probability measure $P(\mathbf{x}, c, \theta)$.

Given a sample $(\mathbf{x}, c, \theta) \in (\mathcal{X} \times [c] \times \Theta)$, its opposite classification label $\bar{c}$ is randomly selected from $[c]\backslash\theta$. When the sampling frequency in a complete learning period is greatly larger than the class number $n_{[c]}$, the probability for each $\bar{c} \in [c]\backslash\theta$ that indicates how likely it is selected can be expressed as

$$P_i(\bar{Y}=\bar{c}|X=\mathbf{x}, Y=c) = \frac{1}{n_{([c]\backslash\theta)}} \tag{7}$$

where $n_{([c]\backslash\theta)}$ is the number of semantic cells in $[c]\backslash\theta$. This conclusion verifies that the proposed semantic-prior based random search method for the opposite semantic label $\bar{c}$ is statistically consistent, and it realizes the independency of $\bar{c}$ with respective to feature space $\mathcal{X}$ conditioned on c and θ. Thus we have,

$$P(\bar{Y}=\bar{c}|X=\mathbf{x}, Y=c) = P(\bar{Y}=\bar{c}|Y=c) \tag{8}$$

The optimal classifier can be found under the uniform assumption, which has been proven in previous work [17]. Meanwhile, the uniform selection means equal probability, ensuring the smooth clustering and the stability and robustness of the learning process. While for man-made complementary labels, they are confined by the fact that $\bar{Y}$ is assumed to be independent of feature $\mathcal{X}$ [18] [17].

Based on the exist of independence, the complete mapping from $\mathbf{x}$ to $\bar{y}$ can be set up as the following formula, $\forall i, j \in [c]$,

$$\begin{aligned} P(\bar{y}|\mathbf{x}) &= \sum_{i \in \theta_i, j \notin \theta_i} P(\bar{y} = j, y = i|\mathbf{x}) \\ &= \sum_{i \in \theta_i, j \notin \theta_i} P(\bar{y} = j|y = i, \mathbf{x}) P(y = i|\mathbf{x}) \\ &= \sum_{i \in \theta_i, j \notin \theta_i} P(\bar{y} = j|y = i) P(y = i|\mathbf{x}) \end{aligned} \tag{9}$$

4.4. Learning with Smooth Semantic Clustering

Conventionally, the classifier is trained to learn that the input image belongs to a specific, single class label. Let $\mathbf{x} \in \mathcal{X}$ be the input image, $y \in [c]$ denotes its label. $f(\mathbf{x}, W)$ maps the input $\mathbf{x}$ to the score space: $\mathcal{X} \rightarrow \mathbb{R}^c$, as equation (4) shows. The training process is guided by the cross entropy loss (most popular classification cost function) of f as

$$\mathcal{L}_{\mathbb{P}}(f, y) = -\sum_{m=1}^{c} \mathbf{y}_m \log \mathbf{p}_m \tag{10}$$

where $\mathbf{y} \in \{0, 1\}^c$ is the one-hot vector form of y. $\mathbf{p}_m$ is the m^{th} element of probability vector $\mathbf{p}$. The conventional learning process is to optimize the probability $\mathbf{p}_m$ according to the given exact label $\mathbf{y}_m$ so that $\mathbf{p}_m \rightarrow 1$. Based on which, we propose a learning algorithm with smooth high-level clustering by guiding f to learn the semantic prior from the opposite label. Inspired by [19], the opposite semantic should push f to optimize the corresponding classification probability $\bar{\mathbf{p}}_m \rightarrow 0$.

$$\mathcal{L}_{\mathbb{O}}(f, y) = -\sum_{m=1}^{c} \bar{\mathbf{y}}_m \log(1 - \bar{\mathbf{p}}_m) \tag{11}$$

where $\mathbf{y}_m \in \theta_m$, $\bar{\mathbf{y}}_m \in [c]$ and $\bar{\mathbf{y}}_m \notin \theta_m$. $\bar{\mathbf{p}}_m$ is the corresponding classification possibility of label $\bar{\mathbf{y}}_m$ in vector $\mathbf{p}$. Thus, the random selection of $\bar{\mathbf{y}}_m$ comes from $[c]\backslash\theta$ in every iteration during the training process, shown in Algorithm 1.

Algorithm 1 Smooth Semantic Clustering

```
Input: Training label $y \in \mathcal{Y} = [c]$, semantic prior $\hat{\theta} \sim \hat{\Theta}$
while iteration do
    if $y \in \hat{\theta}_i$ then
        $\bar{y}$ = Select randomly from $[c]\backslash\hat{\theta}_i$
    There exists another semantic colony $\theta_j$
    if $\bar{y} \in \theta_j$ then
        $y \notin \theta_j$
Output: Opposite semantic label $\bar{y}$ and the learned semantic colony $\theta \sim \Theta$
```

From Algorithm 1, we can observe that the learning for clustering in the semantic space $\mathcal{C}$ is synchronous with image classification. Thus, we can define a composite loss function for an end-to-end semantic clustering classifier.

$$\begin{aligned} \mathcal{L} &= \alpha_1 \mathcal{L}_{\mathbb{P}} + \alpha_2 \mathcal{L}_{\mathbb{O}} \\ &= -\alpha_1 \sum_{m=1}^{c} \mathbf{y}_m \log \mathbf{p}_m - \alpha_2 \sum_{m=1}^{c} \bar{\mathbf{y}}_m \log(1 - \bar{\mathbf{p}}_m) \end{aligned} \tag{12}$$

where α_1 and α_2 are weights defining the ratio of $\mathcal{L}_{\mathbb{P}}$ and $\mathcal{L}_{\mathbb{O}}$ respectively.

For a specific input image, there is not only a semantic label y but also other semantic description $\theta \sim \Theta$, and θ is the high-level semantic expression corresponding to y, which builds a new semantic attribute with a larger range. Since the opposite semantic is randomly selected with equal probability, the clustering hyperplane in $\mathcal{C}$ can be smooth.

4.5. Optimal Learning

In the case of $\mathcal{L}$, we define the expected risk $\bar{R}(f)$ with the mapping $f : \mathcal{X} \rightarrow \{[c], \Theta\}$. If we can find an optimal f^* such that $f^* = P(Y = i|X), \forall i \in [c]$, then in theory, we expect that we can find the optimal $\bar{f}^*$ such that $\bar{f}^* = P(\bar{Y} = i|X), \forall i \in [c]$, where $P(\bar{Y}|X) = \sum_{i \in \theta_i, j \notin \theta_i} P(\bar{Y} = j, Y = i|X)$ according to equation (9). If the above idea can be proved, with sufficient training samples, the proposed algorithm with $\bar{R}(f)$ is capable of simultaneously learning a good classification and clustering for (X, Y, θ).

Following [18], we will prove that the proposed semantic clustering learning with its corresponding loss function $\mathcal{L}$ is able to identify the optimal classifier. First, we introduce the following assumption [18],

Assumption 1. *The optimal learning with mapping f^* satisfies $f_i^*(X) = P(Y = i|X), \forall i \in [c]$ by minimizing the expected risk $R(f)$.*

Based on this assumption, we are able to prove that $\bar{f}^* = f^*$ following the theorem below [18].

Theorem 1. *Suppose that Assumption 1 is satisfied, then the minimum solution $\bar{f}^*$ of $\bar{R}(f)$ is also the minimum solution f^* of $R(f)$, i.e., $\bar{f}^* = f^*$.*

Proof. Based on Assumption 1, loss function $\mathcal{L}$, and function (9) for the learning in the proposed smooth semantic clustering, we have

$$\begin{aligned} f_i^*(X) &= P(\bar{Y} = j|X) \\ &= \sum_{i \in \theta_i} P(\bar{Y} = j, Y = i|X), \forall i, j \in [c], j \notin \theta_i \end{aligned} \tag{13}$$

Let $\bar{\mathbf{s}}(X) = [P(\bar{Y} = 1|X), \cdots, P(\bar{Y} = c)|X)]$ and $\mathbf{s}(X) = [P(Y = 1|X), \cdots, P(Y = c)|X)]$. According to the discussion of [18], we rewrite $\bar{R}(f)$ as

$$\begin{aligned} \bar{R}(f) &= \int_X \sum_{j=1}^{c} P(\bar{Y} = j) P(X|\bar{Y} = j) \mathcal{L}(f(X), \bar{Y} = j) dX \\ &= \sum_{j=1}^{c} P(\bar{Y} = j) \int_X P(X|\bar{Y} = j) \mathcal{L}(f(X), \bar{Y} = j) dX \\ &= \sum_{j=1}^{c} P(\bar{Y} = j) \bar{R}_j(f) \end{aligned} \tag{14}$$

where $P(\bar{Y} = j)$ is given when we have $Y = i$, distributed as $P(\bar{Y} = j|Y = i)$ according to Algorithm 1. $\bar{R}_j(f) = \int_X P(X|\bar{Y} = j) \mathcal{L}(f(X), \bar{Y} = j) dX$. Thus, if we use $\mathbf{C}$ to denote the operation form of $P(\bar{Y} = j|Y = i)$, according to function (9) and the above convergence analysis, we have

$$\bar{\mathbf{s}}(X) = \mathbf{C}^T \mathbf{s}(X) \tag{15}$$

where $P(\bar{Y} = j|Y = i)$ is realized based on the random search with semantic prior. Equation (15) ensures that

$$\bar{f}^*(X) = \arg\max_i \mathbf{C}^T \mathbf{s}_i(X) = \mathbf{C}^T \arg\max_i \mathbf{s}_i(X) = \mathbf{C}^T f^*(X) \tag{16}$$

where $i \in [1, c]$. Thus, we have $\bar{f}^* \iff f^*$. The proof is completed. □

5. Experiment

In this section, we study the impact of the proposed semantic deduction algorithm on popular image classifiers using mainstream benchmark datasets. In order to show that our algorithm is able to generalize to complex or disordered data environment with better robustness, we follow each specific experimental setting of the baseline methods, and only vary the data environment by producing noisy labels at certain ratios.

Learning Scenarios. To identify the gain of the proposed deduction learning algorithm, we design fairly comparable learning scenarios where only the deduction related hyper-parameters are changed from the default original setting while keeping all the rest unchanged. The assignment for the weights of α_1 and α_2 in equation12 is based on the experiment performance. We introduce the most core algorithm idea of the current state-of-the-art works of complementary supervision information designed for various fields [19] [18] [17] into our experiment setting as one of the baselines. Details are listed below:

- Default Setting (OT): In this setting, we train the original baseline classification models and keep all the hyper-parameters unchanged as in the corresponding published papers and public code. We take both classical and state-of-the-art CNN classifier networks into consideration, including Multilayer Perceptron (MLP)[34], VGG [34], ResNet[2], DenseNet [1], Wresnet [35], ResNext[36]. All of them are trained and compared with our proposed methods fairly.

- Random Opposite Semantic (RT): Under this setting, we exploit the opposite semantic label $\bar{y} \in [c]$ that corresponds to the original accurate label $y \in [c]$, satisfying $\bar{y} \neq y$. We use random search for the opposite label in the label pools $[c]$ [19] instead of hard labeling so as to avoid bias [18] [19]. Thus, this setting does not refer to the semantic prior when looking for the opposite semantic label $\bar{y}$. All other settings follow the Default Setting.

- Semantic Deduction (SD): We implement the proposed deduction learning by semantic clustering. The opposite semantic label $\bar{y}$ is randomly selected from $[c]\backslash\hat{\theta}_i$, where $[c]$ is the set of semantic labels. $\hat{\theta}_i$ is the i_th semantic colony (details in Algorithm 1). Thus, it naturally satisfies $\bar{y} \neq y$, y referring to the original accurate label $y \in [c]$. It strictly follows the training setting with the identical hyper-parameters to those in the Default Setting.

Data Sets.

- Fashion-MNIST: Fashion-MNIST is a new image classification benchmark with different data classes of clothing[37]. The dataset has an image size of 28×28, input channels of 1, and the number of classes of 10. In our SD setting, we provide the semantic prior for it to group the 10 classes fashion clothing into three groups: "clothes", "shoes", and "bags".

- CIFAR10: CIFAR10 consists of $50,000$ training images and $10,000$ test images of dimension 32×32. It has a total of 10 general classes[38]. In the SD setting, we group the 10 classes into two groups, "vehicles" and "animals".

- CIFAR100: CIFAR100 has $50,000$ training images and $10,000$ test images of the resolution of 32×32. It has a total of 100 classes, with 500 training images in each class [38]. For the SD setting, we provide

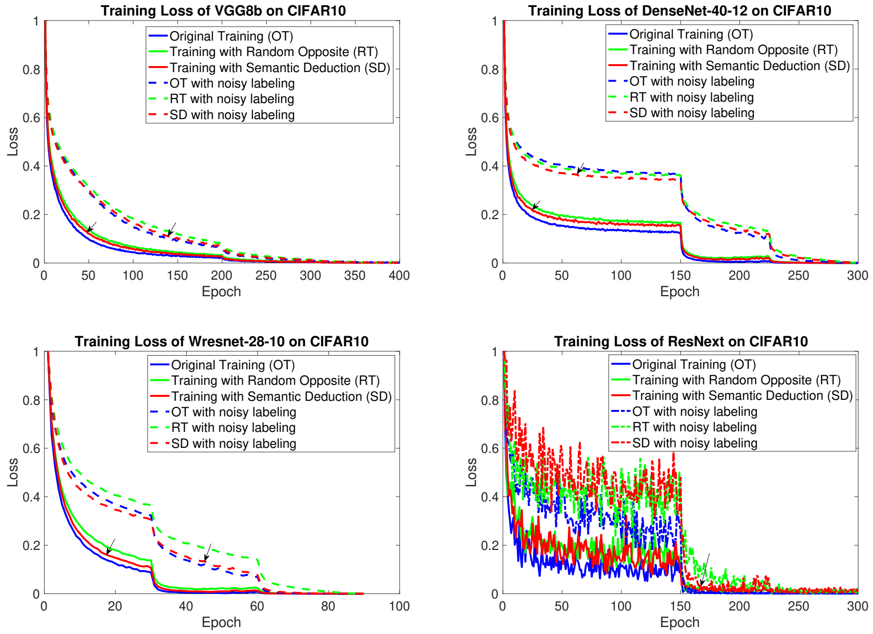

Figure 3: Convergence performance of different models by training loss on CIFAR10.

two schemes, "SD_v1" and "SD_v2. The former one divides classes into "7" groups, including "people", "animal", "man-made stuff", "transportation", "plants", "building", and "nature". The latter contains 8 groups: "people", "animal", "life appliances", "transportation", "food", "plants", "building", and "nature", isolating "food" from the "man-made" as an independent expression.

5.1. Results in Original Data Environment

We first evaluate our proposed algorithm in the original data environment, directly using the images from the data sources. From the mathematical analysis in Section 4, the identification of optimal learning depends on stable convergence performance. Thus, we summarize the learning behaviors of each approach on CIFAR10 and CIFAR100 in Figure 3 and Figure 4, respectively.

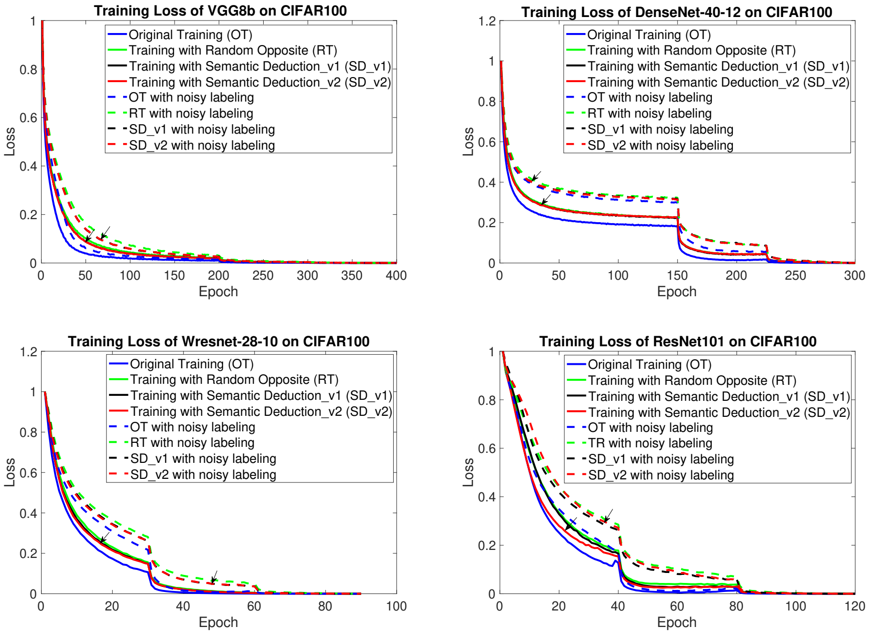

Figure 4: Convergence performance comparison by training loss on CIFAR100.

Convergence Performance. To obtain a fair comparison, we normalize the loss distribution to $[0, 1]$ for all scenarios. **(a)** Our algorithm generally shows consistent convergence with different classifiers, as shown in the red or yellow solid lines in Figure 3 and 4. We can see that SD usually converges faster than RT as the black arrows shown in almost every case. This consistent performance verifies that the proposed self-clustering learning process helps speed up convergence, assisting the classifier to execute the right decision, although there is no additional labeling information fed into these models. **(b)** From all the sub-figures in both Figure 3 and 4, although SD converges a little bit slower than the original baseline (solid blue line) at the first stage, they finally obtain similar stability. This is due to the introduction of the additional learning process, semantic clustering. **(c)** Although we design two semantic prior schemes, SD_v1 and SD_v2, they both show very consistent convergence, where the red and black solid lines even overlap with each other in Figure 4. **(d)** The fluctu-

Method	Solver	α_1	α_2	OT	RT	SD(ours)
MLP-3[34]	adam	1	0.5	91.5	91.77	**91.78**
VGG8b [34]	adam	1	0.3	95.45	95.47	**95.53**
VGG8b(multi=2.0) [34]	adam	1	0.3	95.33	95.52	**95.54**

Table 1: Classification accuracy on FashionMNIST.

Method	Solver	α_1	α_2	OT	RT	SD(ours)
VGG8b [34]	adam	1	0.5	94.12	94.14	**94.32**
ResNet18[2]	adam	1	0.5	93.45	93.57	**93.62**
DenseNet-40-12 [1]	sgd	1	0.5	94.68	94.79	**94.92**
Wresnet-28-10 [35]	sgd	1	0.5	94.52	94.58	**94.80**
ResNext [36]	sgd	1	0.5	96.16	96.26	**96.30**

Table 2: Classification accuracy on CIFAR10.

ation in ResNext is due to the non-averaged loss value in the original code for each epoch. From the above observation, it is evident that the introduction of semantic clustering achieves stable and fast convergence, theoretically qualified to yield an optimal classification mapping.

Classification Accuracy. We summarize the classification accuracy in Table 1, 2, and 3. **(a)** Generally, SD obtains almost the highest classification accuracy across the three benchmarks for all the compared classifiers. These classifiers include two mainstream solvers, adam [39] and sgd [40], but SD leads the performances in both situations. **(b)** Although the improvement brought by SD is limited for Fashion MNIST, this is mainly due to the relatively simple classification task and the limited number of classes. When it comes to CIFAR100 as shown in Table 3, SD always yields 1-3% increase in accuracy compared with OT. **(c)** We can observe that RT in some special situations achieves high performance, such as RT winning SD in the case of Wresnet-28-10. However, its performance is not as stable as SD, which even yields lower classification accuracy than OT, such as that in the case of ResNet101. These observations imply

Method	Solver	α_1	α_2	OT	RT	SD_v1	SD_v2
VGG8b [34]	adam	1	0.5	73.85	74.78	**74.95**	74.83
ResNet50 [2]	sgd	1	0.5	73.78	76.36	75.59	**76.64**
DenseNet-40-12 [1]	sgd	1	0.5	74.89	75.82	**76.26**	75.73
Wresnet-28-10 [35]	sgd	1	0.5	76.98	**77.62**	77.54	77.59
ResNet101 [2]	sgd	1	0.5	75.3	74.45	75.51	**76.29**
ResNet152 [2]	sgd	1	0.3	72.21	73.25	74.38	**74.40**

Table 3: Classification accuracy on CIFAR100.

that the proposed smooth semantic clustering algorithm can effectively enhance the performance of state-of-the-art classifiers, preserving a very stable learning state at the same time, potentially leading to its broader applicability.

Compared with the recent publication [24], which proposes a network learning algorithm that organizes the incrementally learning data into feature-driven super-class and improves upon existing hierarchical CNN models by introducing the capability of self-growth, so that the finer classification is done. This idea, to a certain degree, shares a similar concept with our idea, except that we do not need to label data with super-class and keep the same hierarchical structure during the learning process. We compare its results with ours in Table 4 and Table 5, respectively. It can be seen from them that, although the Tree-CNN models provide a competitive accuracy with its base network VGG-11, it shows no advantages over our SD models. SD models obtain a more than 4% advantage over incremental learning methods (VGG11 and Tree-CNN in Table 4) on CIFAR 10 and averagely 5% higher than incremental learning methods on CIFAR100 considering the test classification accuracy. It demonstrates that our proposed high-level semantic clustering algorithm, in a direct supervised learning, could further improve the adaptive ability towards data, and keep a stable learning process, which is further verified in the following sections. Most importantly, we focus on the exploration towards the self-deducing ability of CNN models, which is different from all the above-mentioned ideas.

Model	VGG11	Tree-CNN	VGG8b	ResNet18-SD	DenseNet-SD	WresNet-SD
Test Accuracy	90.51	86.24	94.32	93.62	94.92	94.80

Table 4: Comparison with Tree-CNN on Cifar10, where SD refers to models that are applied with our proposed algorithm. VGG11 and Tree-CNN are trained by ”old” and ”new” data in an incremental way [24].

Model	VGG11	Tree-CNN5	Tree-CNN10	Tree-CNN20	VGG8b-SD	Wresnet-28-10-SD
Test-Acc	72.23	69.85	69.53	68.49	74.95	77.54

Table 5: Comparison with Tree-CNN on Cifar100, where Test-Acc stands for the Test Accuracy. SD refers to the corresponding models that are applied with our proposed algorithm. VGG11 and Tree-CNN are trained by ”old” and ”new” data in an incremental way [24].

5.2. Results in Noisy Data Environment

In this section, we evaluate the proposed algorithm in noisy data environments. We produce a noisy data environment by adding noise labels to the original data sources. Specifically, we implement this operation on CIFAR10 and CIFAR100, where 10% of the training data in each data set are randomly labeled by incorrect labels that belong to the same colony with the correct labels. For example, if the image is labeled correctly by “cat”, then we randomly search another class label in the “animal” cluster such as “dog” as the replacement of the label “cat”.

Convergence Performance. The comparative results are shown in Figures 3 and 4, from which we can see that **(a)** SD maintains the same learning stability as that in original environment. It even surpasses the baseline OT by convergence speed in some cases, such as DenseNet-40-12 and Wresnet-28-10 on CIFAR10. **(b)** SD generally converges faster than RT, especially in the case of Wresnet-28-10. It shows SD works better assisting the classifier to execute reasonable classification decisions in noisy situations, which exhibits good robustness of the proposed algorithm. **(c)** SD with the composite loss function “$\mathcal{L}$” shows perfect robustness across both shallow and deep networks. Thus, SD is expected to

Method	Solver	α_1	α_2	OT	RT	SD(ours)
VGG8b [34]	adam	1	0.3	89.71	**90.52**	90.33
ResNet18 [2]	adam	1	0.3	89.22	**90.71**	90.32
DenseNet-40-12 [1]	sgd	1	0.5	91.47	92.16	**92.25**
wresnet-28-10[35]	sgd	1	0.5	89.07	87.67	**89.21**
ResNext[36]	sgd	1	0.3	91.29	92.17	**92.53**

Table 6: Classification on CIFAR10 with noisy labels.

Method	Solver	α_1	α_2	OT	RT	SD_v1	SD_v2
VGG8b [34]	adam	1	0.5	67.68	68.72	68.89	**68.95**
ResNext [2]	sgd	1	0.5	75.48	74.51	75.03	**75.65**
DenseNet-40-12 [1]	sgd	1	0.5	70.25	**72.80**	72.61	72.09
wresnet-28-10 [35]	sgd	1	0.5	71.42	71.79	71.60	**72.59**
ResNet101[2]	sgd	1	0.5	68.93	67.97	68.71	**69.75**

Table 7: Classification on CIFAR100 with noisy labels.

identify the optimal classification theoretically.

Classification Accuracy. The comparative results are shown in Tables 6 and 7. We can observe that **(a)** SD, in general, surpasses OT by 1-2%. **(b)** Although RT surpasses SD in some cases, their results are very close. SD is always consistent for all the compared models. **(c)** RT is less robust than SD for its poor performance in some cases with much lower accuracy than OT, such as Wresnet-28-10 on CIFAR10, and ResNext and ResNet101 on CIFAR100.

These observations indicate that the proposed deduction learning by semantic clustering not only enhances the classification performance but also improves the generalization for a given classifier. From the above experiments, it is evident that the proposed semantic clustering method can help the model achieve more accurate classification decisions. Although the semantic prior-based opposite label search provides rough information, it can aid the model to deduce high-level semantic expression along with the entire learning process, realizing the experience accumulation and basic cognitive learning. Thus, it could be an excellent plug-in module that could be applied in other supervised learning, few-shot learning, zero-shot learning, or even semi-supervised learning where each learning stage could be a better fit, generalized, and becoming much more robust. In the meanwhile, from the perspective of calculation, the proposed mechanism of deduction learning by the opposite semantic constraint only introduces one more loss item, which is only the tenth level of the order of magnitudes. Compared with matrix multiplication of any two layers during the training process which has the million level of the order of magnitudes, our proposed model is capable of keeping the time complexity of calculation, while its superior stability and robustness make it easy to be generalized to other computer vision tasks.

6. Conclusion

In this paper, we have proposed a deduction learning approach to boost the gain of high-level semantic clustering. We have demonstrated that if a classifier can perform further independent mapping in the semantic space, it will help the model achieve higher classification performance with better generalization ability and robustness. The proposed smooth semantic clustering algorithm ensures label learning and semantic deduction being processed in the same timeline so as to form a basic cognition. Extensive experiments across various classifiers on different datasets demonstrate the superiority of the proposed method toward further enhancing state-of-the-art classification performance.

Acknowledgement

The work was supported in part by The National Aeronautics and Space Administration (NASA) under the grant no. 80NSSC20M0160, and the Nvidia GPU grant.

References

[1] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] L. He, G. Wang, Z. Hu, Learning depth from single images with deep neural network embedding focal length, IEEE Transactions on Image Processing 27 (9) (2018) 4676–4689.

[4] W. Ma, Y. Wu, F. Cen, G. Wang, Mdfn: Multi-scale deep feature learning network for object detection, Pattern Recognition 100 (2020) 107149.

[5] W. Xu, K. Shawn, G. Wang, Toward learning a unified many-to-many mapping for diverse image translation, Pattern Recognition 93 (2019) 570–580.

[6] Z. Zhang, W. Ma, Y. Wu, G. Wang, Self-orthogonality module: A network architecture plug-in for learning orthogonal filters, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1050–1059.

[7] G. Marcus, The next decade in ai: Four steps towards robust artificial intelligence, arXiv preprint arXiv:2002.06177.

[8] F. Cen, X. Zhao, W. Li, G. Wang, Deep feature augmentation for occluded image classification, Pattern Recognition 111 (2020) 107737.

[9] K. Li, N. Y. Wang, Y. Yang, G. Wang, Sgnet: A super-class guided network for image classification and object detection, arXiv preprint arXiv:2104.12898.

[10] W. Ma, K. Li, G. Wang, Location-aware box reasoning for anchor-based single-shot object detection, IEEE Access 8 (2020) 129300–129309.

[11] E. P. Xing, M. I. Jordan, S. J. Russell, A. Y. Ng, Distance metric learning with application to clustering with side-information, in: Advances in neural information processing systems, 2003, pp. 521–528.

[12] V. Vapnik, A. Vashist, A new learning paradigm: Learning using privileged information, Neural networks 22 (5-6) (2009) 544–557.

[13] M. T. Law, Y. Yu, R. Urtasun, R. S. Zemel, E. P. Xing, Efficient multiple instance metric learning using weakly supervised data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 576–584.

[14] P. Haeusser, A. Mordvintsev, D. Cremers, Learning by association–a versatile semi-supervised training method for neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 89–98.

[15] C. Gong, H. Zhang, J. Yang, D. Tao, Learning with inadequate and incorrect supervision, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 889–894.

[16] I. Misra, C. Lawrence Zitnick, M. Mitchell, R. Girshick, Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2930–2939.

[17] T. Ishida, G. Niu, W. Hu, M. Sugiyama, Learning from complementary labels, in: Advances in neural information processing systems, 2017, pp. 5639–5649.

[18] X. Yu, T. Liu, M. Gong, D. Tao, Learning with biased complementary labels, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 68–83.

[19] Y. Kim, J. Yim, J. Yun, J. Kim, Nlnl: Negative learning for noisy labels, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 101–110.

[20] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE transactions on pattern analysis and machine intelligence 38 (7) (2015) 1425–1438.

[21] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, H. Adam, Large-scale object classification using label relation graphs, in: European conference on computer vision, Springer, 2014, pp. 48–64.

[22] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, arXiv preprint arXiv:1606.04671.

[23] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, icarl: Incremental classifier and representation learning, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2001–2010.

[24] D. Roy, P. Panda, K. Roy, Tree-cnn: a hierarchical deep convolutional neural network for incremental learning, Neural Networks 121 (2020) 148–160.

[25] A. Ghosh, H. Kumar, P. Sastry, Robust loss functions under label noise for deep neural networks, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[26] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in: Advances in neural information processing systems, 2018, pp. 8778–8788.

[27] D. Hendrycks, M. Mazeika, D. Wilson, K. Gimpel, Using trusted data to train deep networks on labels corrupted by severe noise, in: Advances in neural information processing systems, 2018, pp. 10456–10465.

[28] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 4334–4343.

[29] K.-H. Lee, X. He, L. Zhang, L. Yang, Cleannet: Transfer learning for scalable image classifier training with label noise, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5447–5456.

[30] Y. Ding, L. Wang, D. Fan, B. Gong, A semi-supervised two-stage approach to learning from noisy labels, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1215–1224.

[31] D. Tanaka, D. Ikami, T. Yamasaki, K. Aizawa, Joint optimization framework for learning with noisy labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5552–5560.

[32] M. Baroni, A. Lenci, Distributional memory: A general framework for corpus-based semantics, Computational Linguistics 36 (4) (2010) 673–721.

[33] P. L. Bartlett, M. I. Jordan, J. D. McAuliffe, Convexity, classification, and risk bounds, Journal of the American Statistical Association 101 (473) (2006) 138–156.

[34] A. Nøkland, L. H. Eidnes, Training neural networks with local error signals, in: International Conference on Machine Learning, PMLR, 2019, pp. 4839–4850.

[35] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146.

[36] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[37] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, arXiv preprint arXiv:1708.07747.

[38] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images.

[39] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[40] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.