Dimensionality-Driven Learning with Noisy Labels

Xingjun Ma [* 1] **Yisen Wang** [* 2] **Michael E. Houle** [3] **Shuo Zhou** [1] **Sarah M. Erfani** [1] **Shu-Tao Xia** [2]
Sudanthi Wijewickrema [1] **James Bailey** [1]

Abstract

Datasets with significant proportions of noisy (incorrect) class labels present challenges for training accurate Deep Neural Networks (DNNs). We propose a new perspective for understanding DNN generalization for such datasets, by investigating the dimensionality of the deep representation subspace of training samples. We show that from a dimensionality perspective, DNNs exhibit quite distinctive learning styles when trained with clean labels versus when trained with a proportion of noisy labels. Based on this finding, we develop a new dimensionality-driven learning strategy, which monitors the dimensionality of subspaces during training and adapts the loss function accordingly. We empirically demonstrate that our approach is highly tolerant to significant proportions of noisy labels, and can effectively learn low-dimensional local subspaces that capture the data distribution.

1. Introduction

Deep Neural Networks (DNNs) have demonstrated excellent performance in solving many complex problems, and have been widely employed for tasks such as speech recognition (Hinton et al., 2012), computer vision (He et al., 2016) and gaming agents (Silver et al., 2016). DNNs are capable of learning very complex functions, and can generalize well even for a huge number of parameters (Neyshabur et al., 2014). However, recent studies have shown that DNNs may generalize poorly for datasets which contain a high proportion noisy (incorrect) class labels (Zhang et al., 2017). It is important to gain a fuller understanding of this phenomenon, with a view to development of new training methods that can

achieve good generalization performance in the presence of variable amounts of label noise.

One simple approach for noisy labels is to ask a domain expert to relabel or remove suspect samples in a preprocessing stage. However, this is infeasible for large datasets and also runs the risk of removing crucial samples. An alternative is to correct noisy labels to their true labels via a clean label inference step (Vahdat, 2017; Veit et al., 2017; Jiang et al., 2017; Li et al., 2017). Such methods often assume the availability of a supplementary labelled dataset containing pre-identified noisy labels which are used to develop a model of the label noise. However, their effectiveness is tied to the assumption that the data follow the noise model. A different approach to tackle noisy labels is to utilize correction methods such as loss correction (Patrini et al., 2017; Ghosh et al., 2017), label correction (Reed et al., 2014), or additional linear correction layers (Sukhbaatar & Fergus, 2014; Goldberger & Ben-Reuven, 2017).

In this paper, we first investigate the dimensionality of the deep representation subspaces learned by a DNN and provide a dimensionality-driven explanation of DNN generalization behavior in the presence of (class) label noise. Our analysis employs a dimensionality measure called Local Intrinsic Dimensionality (LID) (Houle, 2013; 2017a), applied to the deep representation subspaces of training examples. We show that DNNs follow two-stage of learning in this scenario: 1) an early stage of *dimensionality compression*, that models low-dimensional subspaces that closely match the underlying data distribution, and 2) a later stage of *dimensionality expansion*, that steadily increases subspace dimensionality in order to overfit noisy labels. This second stage appears to be a key factor behind the poor generalization performance of DNNs for noisy labels. Based on this finding, we propose a new training strategy, termed *Dimensionality-Driven Learning*, that avoids the dimensionality expansion stage of learning by adapting the loss function. Our main contributions are:

- We show that from a dimensionality perspective, DNNs exhibit distinctive learning styles with clean labels versus noisy labels.

- We show that the local intrinsic dimensionality can

---

[*]Equal contribution [1]The University of Melbourne, Melbourne, Australia [2]Tsinghua University, Beijing, China [3]National Institute of Informatics, Tokyo, Japan. Correspondence to: Yisen Wang <wangys14@mails.tsinghua.edu.cn>, Xingjun Ma <xingjun.ma@unimelb.edu.au>.

be used to identify the stage shift from dimensionality compression to dimensionality expansion.

- We propose a Dimensionality-Driven Learning strategy (D2L) that modifies the loss function once the turning point between the two stages of dimensionality compression and expansion is recognized, in an effort to prevent overfitting.

- We empirically demonstrate on MNIST, SVHN, CIFAR-10 and CIFAR-100 datasets that our Dimensionality-Driven Learning strategy can effectively learn (1) low-dimensional representation subspaces that capture the underlying data distribution, (2) simpler hypotheses, and (3) high-quality deep representations.

2. Related Work

2.1. Generalization of DNNs

Zhang et al. (2017) showed that DNNs are capable of memorizing completely random labels and exhibit poor generalization capability. They argued that DNNs employ case-by-case memorization on training samples and their labels in this scenario. Krueger et al. (2017) highlighted that DNNs exhibit different learning styles on datasets with clean labels versus those on datasets with noisy inputs or noisy labels. They showed that DNNs require more capacity, longer training time to fit noisy labels and the learned hypothesis is more complex. Arpit et al. (2017) further substantiated this finding by identifying two stages of learning of DNNs with noisy labels: an early stage of simple pattern learning and refining, and a later stage of label memorization. They also showed that dropout regularization can hinder overfitting to noisy labels. Shwartz-Ziv & Tishby (2017) demonstrated that, on data with clean labels, DNNs with tanh layers undergo an initial label fitting phase and then a subsequent compression phase. They also argued that information compression is related to the excellent generalization performance of DNNs. However, Saxe et al. (2018) conducted experiments where information compression was not found to occur for ReLU (Glorot et al., 2011) DNNs.

While these works have studied the differences between learning with clean labels and learning with noisy labels, a full picture of this phenomenon and its implications for DNN generalization is yet to emerge. Our study adds another perspective based on subspace dimensionality analysis, and shows how this can lead to the development of an effective learning strategy.

2.2. Noisy Label Learning

A variety of approaches have been proposed to robustly train DNNs on datasets with noisy labels. One strategy is to explicitly or implicitly formulate the *noise model* and use a corresponding noise-aware approach. Symmetric label noise that is independent of the true label was modeled in (Larsen et al., 1998), and asymmetric label noise that is conditionally independent of individual samples was modeled in (Natarajan et al., 2013; Sukhbaatar et al., 2014). There are also more complex noise models for training samples where true labels and noisy labels can be characterized by directed graphical models (Xiao et al., 2015), conditional random fields (Vahdat, 2017), neural networks (Veit et al., 2017; Jiang et al., 2017) or knowledge graphs (Li et al., 2017). These methods aim to correct noisy labels to their true labels via a clean label inference step or by assigning smaller weights to noisy label samples. For the modeling of label noise, they often require an extra dataset with ground truth of pre-identified noisy labels to be available, or an expensive detection process. They may also rely on specific assumptions about the noise model. Another approach is to use a refined training strategy that utilizes correction methods to adjust the loss function to eliminate the influence of noisy samples (Wang et al., 2018). Backward and Forward are two such correction methods that use an estimated or learned factor to modify the loss function (Patrini et al., 2017). A linear layer is added on top of the network to further augment the correction architecture in (Sukhbaatar & Fergus, 2014; Goldberger & Ben-Reuven, 2017). Bootstrap replaces the target labels with a combination of raw target labels and their predicted labels (Reed et al., 2014).

Our proposed Dimensionality-Driven Learning strategy is also a loss correction method, one that avoids overfitting by using the estimation of the local intrinsic dimensionality of learned local subspaces to regulate the learning process. In Section 5 we empirically compare Dimensionality-Driven Learning with other loss correction strategies.

2.3. Supervised Learning and Dimensionality

The Local Intrinsic Dimensionality (LID) model (Houle, 2017a) was recently used for successful detection of adversarial examples for DNNs by (Ma et al., 2018). This work demonstrates that adversarial perturbations (one type of input noise) tend to increase the dimensionality of the local subspace immediately surrounding a test sample, and that features based on LID can be used for identifying such perturbations. However, in this paper we show how LID can be used in a new way, as a tool for assessing the learning behavior of a DNN, and developing an adaptive learning strategy against noisy labels.

Other works have also considered the use of dimensionality measures for regularization in manifold learning (Roweis & Saul, 2000; Belkin et al., 2004; 2006). For example, an intrinsic geometry regularization over Reproducing Kernel Hilbert Spaces (RKHS) was proposed in (Belkin et al., 2006)

to enforce smoothness of solutions relative to the underlying manifold, and a Laplacian-based regularization using the weighted neighborhood graph was proposed in (Belkin et al., 2004). In contrast to these works, which treated dimensionality as a characteristic of the global data distribution, we explore how knowledge of local dimensional characteristics can be used to monitor and modify DNN learning behavior for the noisy label scenario.

3. Dimensionality of Deep Representation Subspaces

We now briefly introduce the LID measure for assessing the dimensionality of data subspaces residing in the deep representation space of DNNs. We then connect dimensionality theory with the learning process of DNNs.

3.1. Local Intrinsic Dimensionality (LID)

Local Intrinsic Dimensionality (LID) is an expansion-based measure of intrinsic dimensionality of the underlying data subspace/submanifold (Houle, 2017a). In the theory of intrinsic dimensionality, classical expansion models (such as the expansion dimension (Karger & Ruhl, 2002) and generalized expansion dimension (Houle et al., 2012)) measure the rate of growth in the number of data objects encountered as the distance from the reference sample increases. Intuitively, in Euclidean space, the volume of an D-dimensional ball grows proportionally to r^D when its size is scaled by a factor of r. From the above rate of volume growth with distance, the dimension D can be deduced from two volume measurements as:

$$V_2/V_1 = (r_2/r_1)^D \Rightarrow D = \ln(V_2/V_1)/\ln(r_2/r_1). \quad (1)$$

The aforementioned expansion-based measures of intrinsic dimensionality would determine D by estimating the volumes in terms of the numbers of data points captured by the balls. Transferring the concept of expansion dimension from the Euclidean space to the statistical setting of continuous distance distributions, the notion of ball volume is replaced by the probability measure associated with the balls. This leads to the formal definition of LID (Houle, 2017a):

Definition 1 (Local Intrinsic Dimensionality).
Given a data sample $x \in X$, let $r > 0$ be a random variable denoting the distance from x to other data samples. If the cumulative distribution function $F(r)$ is positive and continuously differentiable at distance $r > 0$, the LID of x at distance r is given by:

$$\mathrm{LID}_F(r) \triangleq \lim_{\epsilon \to 0} \frac{\ln\left(F((1+\epsilon)r)/F(r)\right)}{\ln(1+\epsilon)} = \frac{rF'(r)}{F(r)}, \quad (2)$$

whenever the limit exists. The LID at x is in turn defined as the limit of the radius $r \to 0$:

$$\mathrm{LID}_F = \lim_{r \to 0} \mathrm{LID}_F(r). \quad (3)$$

LID_F describes the relative rate at which its cumulative distance function $F(r)$ increases as the distance r increases. In the ideal case where the data in the vicinity of x are distributed uniformly within a local submanifold, LID_F equals the dimension of the submanifold. Nevertheless, in more general cases, LID also provides a rough indication of the dimension of the submanifold containing x that would best fit the data distribution in the vicinity of x. We refer readers to (Houle, 2017a;b) for more details about LID.

Estimation of LID: Given a reference sample point $x \sim \mathcal{P}$, where $\mathcal{P}$ represents a global data distribution, $\mathcal{P}$ induces a distribution of distances relative to x — each sample $x_* \sim \mathcal{P}$ being associated with the distance value $d(x, x_*)$. With respect to a dataset X drawn from $\mathcal{P}$, the smallest k nearest neighbor distances from x can be regarded as extreme events associated with the lower tail of the induced distance distribution. From the statistical theory of extreme values, the tails of continuous distance distributions can be seen to converge to the Generalized Pareto Distribution (GPD), a form of power-law distribution (Coles et al., 2001; Hill, 1975). Several estimators of LID were developed in (Amsaleg et al., 2015; Levina & Bickel, 2005), of which the Maximum Likelihood Estimator (MLE) exhibited the best trade-off between statistical efficiency and complexity:

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(x)}{r_{max}(x)}\right)^{-1}. \quad (4)$$

Here, $r_i(x)$ denotes the distance between x and its i-th nearest neighbor, and $r_{max}(x)$ denotes the maximum of the neighbor distances. Note that the LID defined in Equation (3) is a *distributional* quantity, and the $\widehat{\mathrm{LID}}$ defined in Equation (4) is its *estimate*.

3.2. LID Estimation through Batch Sampling

Since computing neighborhoods with respect to the entire dataset X can be prohibitively expensive, we will estimate LID of a training example x from its k-nearest neighbor set within a *batch* randomly selected from X. Consider a L-layer neural network $h : \mathcal{P} \to \mathbb{R}^c$, where $h^{(i)}$ is the intermediate transformation of the i-th layer, and c is a positive number indicating the number of classes. Given a batch of training samples $X_B \subseteq X$, and a reference point $x \sim \mathcal{P}$ (not necessarily a training sample), we estimate the LID score of x as:

$$\widehat{\mathrm{LID}}(x, X_B) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(g(x), g(X_B))}{r_{max}(g(x), g(X_B))}\right)^{-1}, \quad (5)$$

where $g = h^{(L-1)}$ is the output of the second-to-last layer of the network, $r_i(g(x), g(X_B))$ is the distance of $g(x)$ to its i-th nearest neighbor in the transformed set $g(X_B)$, and r_{max} represents the radius of the neighborhood. $\widehat{\mathrm{LID}}(x, X_B)$
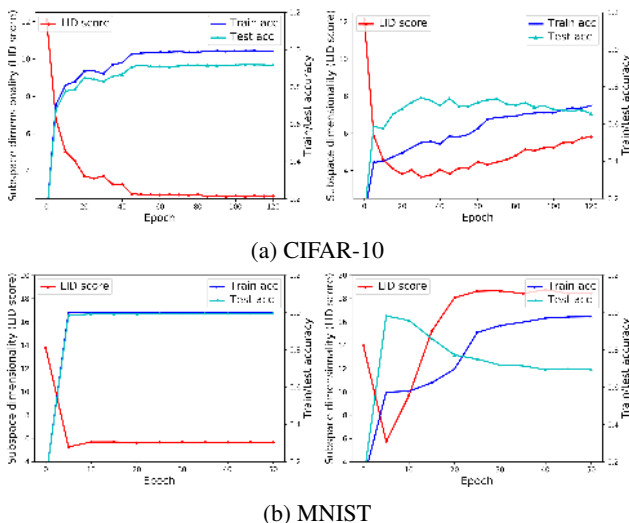
(a) CIFAR-10

(b) MNIST

Figure 1. The subspace dimensionality (average LID scores) and train/test accuracy throughout training for a 12-layer CNN on CIFAR-10 (a) and a 5-layer CNN on MNIST (b) dataset with clean (left subfigures) and noisy labels (right subfigures). The average LID scores were computed at layer 11 for CIFAR-10 and layer 4 for MNIST.

reveals the dimensional complexity of the *local subspace* in the vicinity of x, taken after transformation by g. Provided that the batch is chosen sufficiently large so as to ensure that the k-nearest neighbor sets remain in the vicinity of $g(x)$, the estimate of LID at $g(x)$ within the batch serves as an approximation to the value that would have been computed within the full dataset $g(X)$.

3.3. Subspace Dimensionality and Noisy Labels

We now show by means of an example how the subspace dimensionality of training and test examples is affected by the quality of label information, as the number of training epochs is increased. For our example, we trained a 5-layer Convolutional Neural Network (CNN) on MNIST (an image data set with 10 categories of handwritten digits (LeCun et al., 1998)) and a 12-layer CNN on CIFAR-10 (a natural image data set with 10 categories (Krizhevsky & Hinton, 2009)) using SGD, cross-entropy loss, and two different label quality settings: (1) clean labels for all training samples; (2) noisy labels for 40% of the training samples, generated by uniformly and randomly replacing the correct label with one of the 9 incorrect labels. LID values at layer 4 for MNIST and layer 11 for CIFAR-10 were averaged over 10 batches of 128 points each, for a total of 1280 test points. The resulting LID scores and the train/test accuracies are shown in Figure 1. When learning with clean labels, we observe a decreasing trend in LID score and an increasing trend in accuracy as the number of training epochs increases. However, when learning with noisy labels, we see a very different trend: first a decrease in LID followed by an increase,

accompanied by an initial increase in test accuracy followed by a decrease. We observed similar dimensionality trends for a 6-layer CNN on SVHN (Netzer et al., 2011) and a 44-layer ResNet (He et al., 2016) on CIFAR-100 (Krizhevsky & Hinton, 2009).

Clearly, in these two situations, the DNNs are exhibiting different learning styles. For training data with clean labels, the network gradually transforms the data to subspaces of low dimensionality. Once the subspaces of the lowest dimensionality has been found, the network effectively stops learning: the test accuracy stabilizes at its highest level and the dimensionality stabilizes at its lowest. On the other hand, for training data with noisy labels, the network initially learns a transformation of the data to subspaces of lower dimensionality, although not as low as when training on data with clean labels. Thereafter, the network progressively attempts to accommodate noisy labels by increasing the subspace dimensionality.

3.4. Two-Stage of Learning of DNNs on Noisy Labels

From the above empirical results, we find that DNNs follow two-stage of learning in the presence of label noise: 1) an early stage of *dimensionality compression*, in which the dimensionalities associated with the underlying data manifold are learned; and 2) a later stage of *dimensionality expansion*, in which the subspace dimensionalities steadily increase as the learning process overfits to the noisy data.

One possible explanation for this phenomenon can be found in the effect of transformation on the neighborhood set of test points. Given a training point $x \in X$, its initial spatial location (before learning) would relate to a low-dimensional local subspace determined by the underlying manifold (call this subspace A). Although the initial neighborhood of x would likely contain many data points that are also close to manifold A, the LID estimate would not necessarily be the exact dimension of A. LID reveals the growth characteristics of the distance distribution from x, which is influenced by — but not equal to — the dimension of the manifold to which x is best associated.

As the learning process progresses, the manifold undergoes a transformation by which it progressively achieves a better fit to the training data. If x is labeled correctly, and if many of its neighbors also have clean labels, the learning process can be expected to converge towards a local subspace of relatively low intrinsic dimensionality (as observed in the left-hand plot of Figure 1); however, it should be noted that the learning process still risks overfitting to the data, if carried out too long. With overfitting, the dimensionality of the local manifold would be expected to rise eventually.

If x is incorrectly labeled, each epoch in the learning process progressively causes x — or more precisely, its transform

(call it x') — to migrate to a new local subspace (call it A') associated with members of the same label that was incorrectly applied to x. During this migration, the neighborhood of x' tends to contain more and more points of A' that share the same label as x, and fewer and fewer points from the original neighborhood in A. With respect to the points of A', the mislabeled point x' is spatially an outlier, since its coordinates relate to A and not A'; thus, the presence of x' forces the local subspace around it to become more high-dimensional in order to accommodate (or compress) it. This distortion results in a *dimensionality expansion* in the vicinity of x' that would be expected to be reflected in LID estimates based at x'. Stopping the learning process earlier allows x' to find its neighborhood in A before the local subspace is corrupted by too many neighbors from A', which thus leads to better learning of the true data distribution and improved generalization to test data.

This explanation of the effect of incorrect labeling in terms of local subspaces is consistent with the one recently given in (Ma et al., 2018) for the effect of adversarial perturbation on DNN classification. In this situation, rather than directly assigning an incorrect label to the test item while leaving its spatial coordinates unchanged, the adversary must instead attempt to move a test point into a region associated with an incorrect class by means of an antagonistic learning process. In both cases, regardless of how the test point is modified, the neighborhoods of the transformed points are affected in a similar manner: as the neighborhood membership evolves, the local intrinsic dimensionality can be expected to rise. The associated changes in LID estimates have been used as the basis for the effective detection of a wide variety of adversarial attacks (Ma et al., 2018). Recent theoretical work for adversarial perturbation in nearest-neighbor classification further supports the relationship between LID and local transformation of data, by showing that the magnitude of the perturbation required in order to subvert the classification diminishes as the local intrinsic dimensionality and data sample size grow (Amsaleg et al., 2017).

4. Dimensionality-Driven Learning Strategy

In the previous section, we observed that learning in the presence of noisy labels has two stages: dimensional compression, followed by dimensional expansion. Motivated by these observations, we propose a Dimensionality-Driven Learning (D2L) strategy whose objective is to avoid the overfitting and loss of test accuracy associated with dimensional expansion.

Given a training sample x, we denote its raw label as y and its predicted label as $\widehat{y}$, where both y and $\widehat{y}$ are 'one-hot' indicator vectors. $(\widehat{LID}_0, \cdots, \widehat{LID}_i, \cdots, \widehat{LID}_T)$ is a sequence of LID scores, where $\widehat{LID}_i$ represents the LID score computed from the second-to-last DNN layer at the

i-th training epoch (T epochs in total). Each LID score is produced as follows. m batches of samples are randomly selected $X_B^1, \ldots, X_B^m$ and for each X_B^i and each of its members x, $\widehat{LID}(x, X_B^i)$ is computed. This gives $m \times |X_B^i|$ LID estimates, which are then averaged to compute the LID score for the epoch (later, in the experiments, we use $m = 10$ and $|X_B^i| = 128$

To avoid dimensionality expansion during training with noisy labels, we propose to reduce the effect of noisy labels on learning the true data distribution using the following adaptive LID-corrected labels:

$$y^* = \alpha_i y + (1 - \alpha_i)\widehat{y}, \qquad (6)$$

where α_i is a LID-based factor that updates at the i-th training epoch:

$$\alpha_i = \exp\left(-\lambda \frac{\widehat{LID}_i}{\min_{j=0}^{i-1} \widehat{LID}_j}\right), \qquad (7)$$

where $\lambda = i/T$ is a weighting that indicates decreasing confidence in the raw labels when the training proceeds to the dimensionality expansion stage (that is, when LID begins to increase). The training loss can then be refined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{y_n^*} y_n^* \log P(y_n^* | x_n), \qquad (8)$$

where N is the total number of training samples and $P(y_n^* | x_n)$ is the predicted class probability of y_n^* given x_n.

Interpreting Equations (6) - (8), we can regard D2L as a simulated annealing algorithm that attempts to find an optimal trade-off between subspace dimensionality and prediction performance. The role of α is an exponential decay factor that allows for interpolation between raw and predicted label assignments according to the degree of dimensional expansion observed over the learning history. Here, dimensional expansion is assessed in terms of the ratio of two average LID scores: the score observed at the current epoch, and the lowest score encountered at earlier epochs. As the learning enters the dimensional expansion stage, this ratio exceeds 1, and the exponential decay factor begins to favor the current predicted label. The complete D2L learning strategy is shown in Algorithm 1. Note that the computational cost of LID estimation through batch sampling is low compared to the overall training time ($t_{LID}/t_{training} \approx 1 - 2\%$), as it requires only the pairwise distances within a few batches.

To identify the turning point between the two stages of learning, we employ an epoch window of size $w \in [1, T-1]$ so as to allow w epochs of initialization for the network, and to reduce the variation of stochastic optimization. The turning point is flagged when the LID score of the current epoch is two standard deviations higher than the mean LID score of

Algorithm 1 Dimensionality-Driven Learning (D2L)

Input: dataset X, network $h(x)$, total epochs T, epoch window w, number of batches for LID estimation m.

Initialize: epoch $i \leftarrow 0$, $lids \leftarrow []$, $\alpha_0 \leftarrow 1$, turning epoch $u \leftarrow -1$.

repeat

  Train $h(x)$ for one epoch.

  $lid \leftarrow 0$, $\lambda \leftarrow i/T$.

  for $j = 1$ **to** m **do**

    Sample X_B from X.

    $lid \leftarrow lid + \frac{1}{|X_B|} \sum_{k=1}^{|X_B|} \widehat{\mathrm{LID}}(x, X_B)$.

  end for

  $lids[i] \leftarrow lid/m$.

  if $i \geq w$ **and** $u = -1$ **and**
  $lid - \mathrm{mean}(lids[i - w : i - 1]) > 2 \cdot \mathrm{std}(lids[i - w : i - 1])$ **then**

    $u \leftarrow i - 1$.   *# turning point found*

    Rollback $h(x)$ to the u-th epoch.

  end if

  if $u > -1$ **then**

    $\alpha_i = \exp\left(-\lambda \cdot lids[i]/\min(lids[0 : i - 1])\right)$.

  else

    $\alpha_i = \alpha_0$

  end if

  $y^* = \alpha_i y + (1 - \alpha_i)\widehat{y}$.

  Update loss to $\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{y_n^*} y_n^* \log P(y_n^*|x_n)$.

  $i \leftarrow i + 1$.

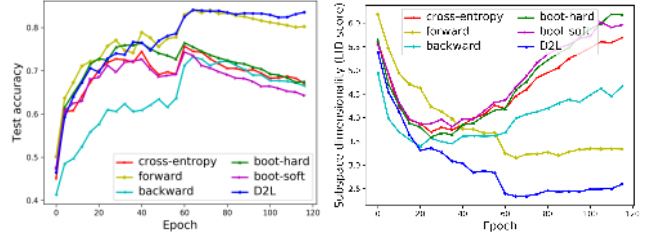**until** $i = T$ **or** early stopping.

(a) CIFAR-10 with 40% noisy labels.

(b) CIFAR-10 with 60% noisy labels.

Figure 2. The trend of test accuracy and subspace dimensionality on CIFAR-10 with 40% and 60% noisy labels.

the w preceding epochs, until which the D2L loss is equivalent to the cross-entropy loss (enforced by setting α equal to 1). The epoch at which the turning point is identified can be regarded as the first epoch at which overfitting occurs; for this reason, we roll the model state back to that of the previous epoch, and begin the interpolation between the raw and predicted label assignments. Although we find in the experimental results of Section 5 that this strategy works consistently well for a variety of datasets, further variations upon this basic strategy may also be effective. The D2L code is available at https://github.com/xingjunm/dimensionality-driven-learning.
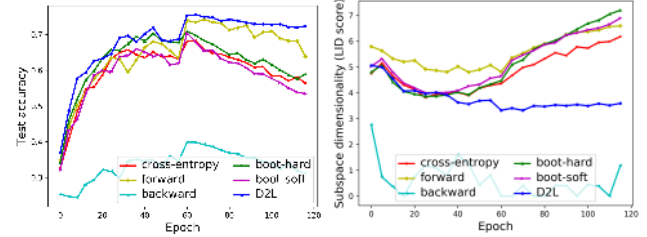
5. Experiments

We evaluate our proposed D2L learning strategy, comparing the performance of our model with state-of-the-art baselines for noisy label learning.

5.1. Empirical Understanding of D2L

We first provide an empirical understanding of the proposed D2L learning strategy on subspace learning, hypothesis learning, representation learning and model analysis.

Experimental Setup: The experiments were conducted on the benchmark dataset CIFAR-10 (Krizhevsky & Hinton, 2009). We used a 12-layer CNN architecture. All networks were trained using SGD with momentum 0.9, weight decay 10^{-4} and an initial learning rate of 0.1. The learning rate was divided by 10 after epochs 40 and 80 ($T = 120$ epochs in total). Simple data augmentations (width/height shift and horizontal flip) were applied. Noisy labels were generated by introducing symmetric noise, in which the labels of a given proportion of training samples are flipped to one of the other class label, selected with equal probability. In (Vahdat, 2017) this noisy label generation scheme has been verified to be more challenging than that of restricted (asymmetric) label noise, which assumes that mislabelling only occurs within a specific set of classes (Reed et al., 2014; Patrini et al., 2017).

Competing Strategies: 1) Backward (Patrini et al., 2017): training via loss correction by multiplying the cross-entropy loss by a noise-aware correction matrix; 2) Forward (Patrini et al., 2017): training with label correction by multiplying the network prediction by a noise-aware correction matrix; 3) Boot-hard (Reed et al., 2014): training with new labels generated by a convex combination (the "hard" version) of the noisy labels and their predicted labels; 4) Boot-soft (Reed et al., 2014): training with new labels generated by a convex combination (the "soft" version) of the noisy labels and their predictions; and 5) Cross-entropy: the conventional approach of training with cross-entropy loss.

The parameters of the competitors were configured according to their original papers. For our proposed D2L, we set $k = 20$ for LID estimation, and used the average LID score over $m = 10$ random batches of training samples as the

overall dimensionality of the representation subspaces.

Effect on Subspace Learning: We illustrate the effect of D2L on subspace learning by investigating the dimensionality (measured by LID) of the deep representation subspaces learned by DNNs and the test accuracy throughout training. The results are presented in Figure 2 for the CIFAR-10 dataset, with noisy label proportions set to 40% and to 60%. First, examining the test accuracy (the left-hand plots), we see that D2L can stabilize the test accuracy after around 60 epochs regardless of the noise rate, whereas the competitors experience a substantial decrease in test accuracy. This indicates the effectiveness of D2L in limiting the overfitting to noisy labels. Second, we focus on the dimensionality of the representation subspaces learned by different models (the right-hand plots). We observe that D2L is capable of learning representation subspaces which have significantly lower dimensionality than other models. It can also be noted that lower-dimensional subspaces lead to better generalization and higher test accuracy. This supports our claim that the true data distribution is of low dimensionality, and that D2L is capable of learning the low-dimensional true data distribution even with a large proportion of noisy labels. Note that for the case of 60% label noise, the low test accuracy of the 'backward' model, as well as the low dimensionality of the learned subspaces, together show that this competitor suffered from underfitting.
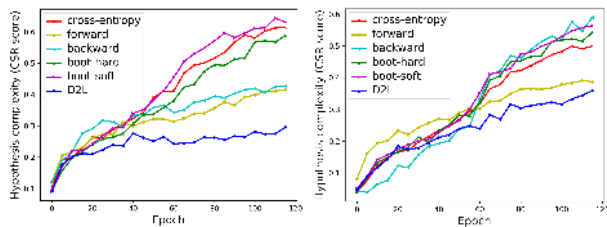

Figure 3. The hypothesis complexity (measured by CSR) on CIFAR-10 with 40% (left) and 60% (right) noisy labels.

Effect on Hypothesis Learning: We investigate the complexity of the hypotheses learned from different models. Given a hypothesis space $\mathcal{H}$, a learned hypothesis $h \in \mathcal{H}$ from a DNN with lower complexity is expected to generalize better. Here, we use the recently proposed Critical Sample Ratio (CSR) (Arpit et al., 2017) as the measure for hypothesis complexity. CSR measures the density around the decision boundaries, where a high CSR score indicates a complex decision boundary and hypothesis.

As shown in Figure 3, the complexity of the learned hypothesis from D2L is significantly lower than that of its competitors. Recalling the results from Figure 2, where D2L achieved the highest test accuracy, we conclude that a simpler hypothesis does lead to better generalization, and that D2L is capable here of learning smoother decision
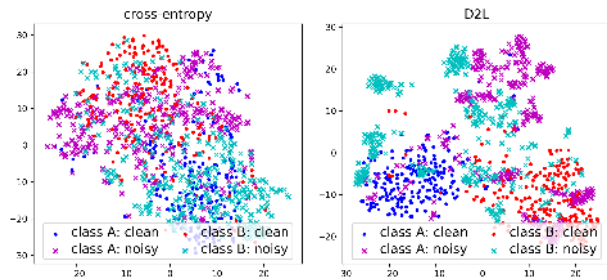

Figure 4. Representations (t-SNE 2D embeddings) of two CIFAR-10 classes, 'airplane' (A) and 'cat' (B), learned by cross-entropy (left) and our D2L model (right), with 60% of the class labels set to noise.
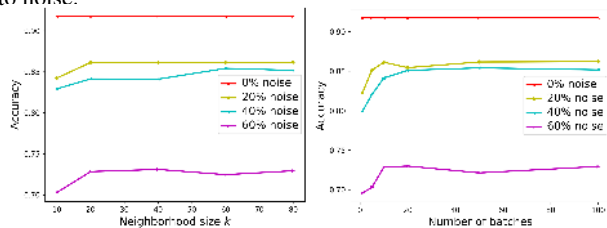

Figure 5. Grid searching neighborhood size k (left) and number of batches m (right) for the estimation of LID on CIFAR-10 with various noise rate.

boundaries and a simpler hypothesis than its competitors.

Effect on Representation Learning: To analyze the effectiveness of D2L for representation learning, we visualize dataset representations in 2-dimensional embeddings using t-SNE (Maaten & Hinton, 2008), a commonly-used dimensionality reduction technique for the visualization of high-dimensional data (LeCun et al., 2015). Figure 4 presents the reduced 2D embeddings of 500 randomly selected samples from each of two classes on CIFAR-10. For each class, 40% of the samples were assigned correct labels (the 'clean' samples), and 60% were assigned incorrect labels chosen uniformly at random from the 9 other classes (the 'noisy' samples). We see that D2L (the right-hand plot) can learn high-quality representations that accurately separate the two classes of objects (blue vs red), and can effectively isolate noisy samples (magenta/cyan) from clean samples (blue/red). However, for both classes, representations learned by cross-entropy (the left-hand plot) suffer from significant overlapping between clean and noisy samples. Note that the representations of noisy samples learned by D2L are more fragmented, since the noisy labels are from many different classes. Overall, D2L is able to learn a high-quality representation from noisy datasets.

Parameter Sensitivity: We assess the sensitivity of D2L to the neighborhood size k and the number of batches m used to compute the mean LID. Figure 5 shows that D2L is relatively insensitive to these two hyper-parameters on the CIFAR-10 dataset. We observed similar behavior with the other three datasets.

Table 1. Test accuracy (%) of different models on MNIST, SVHN, CIFAR-10 and CIFAR-100 with varying noise rates (0% – 60%). The mean accuracy (±std) over 5 repetitions of the experiments are reported, and the best results are highlighted in **bold**.

Dataset / Noise Rate		cross-entropy	forward	backward	boot-hard	boot-soft	D2L
MNIST	0%	99.24±0.0	**99.30±0.0**	99.23±0.1	99.13±0.2	99.20±0.0	99.28±0.0
	20%	88.02±0.1	96.45±0.1	90.12±0.1	87.69±0.2	88.50±0.1	**98.84±0.1**
	40%	68.46±0.1	94.90±0.1	70.89±0.1	69.49±0.2	70.19±0.2	**98.49±0.1**
	60%	45.51±0.2	82.88±0.1	52.83±0.2	50.45±0.1	46.04±0.1	**94.73±0.2**
SVHN	0%	90.12±0.0	90.22±0.1	90.16±0.1	89.47±0.0	89.26±0.0	**90.32±0.0**
	20%	79.10±0.1	85.51±0.1	79.61±0.2	81.21±0.1	79.26±0.2	**87.63±0.1**
	40%	62.92±0.1	79.09±0.2	64.15±0.1	63.25±0.2	64.30±0.2	**82.68±0.1**
	60%	38.54±0.2	62.57±0.2	53.14±0.1	47.61±0.2	39.21±0.2	**80.92±0.2**
CIFAR-10	0%	89.31±0.1	**90.27±0.1**	89.03±0.2	89.06±0.3	89.46±0.2	89.41±0.2
	20%	81.52±0.1	84.61±0.3	79.41±0.1	81.19±0.4	79.21±0.2	**85.13±0.2**
	40%	73.51±0.3	82.84±0.2	74.69±0.2	76.67±0.2	73.81±0.1	**83.36±0.3**
	60%	67.03±0.3	72.41±0.4	45.42±0.4	70.57±0.3	68.12±0.2	**72.84±0.3**
CIFAR-100	0%	68.20±0.2	68.54±0.3	68.48±0.3	68.31±0.2	67.89±0.2	**68.60±0.3**
	20%	52.88±0.2	60.25±0.2	58.74±0.3	58.49±0.4	57.32±0.3	**62.20±0.4**
	40%	42.85±0.2	51.27±0.3	45.42±0.2	44.41±0.1	41.87±0.1	**52.01±0.3**
	60%	30.09±0.2	41.22±0.3	34.49±0.2	36.65±0.3	32.29±0.1	**42.27±0.2**

5.2. Robustness against Noisy Labels

Finally, we evaluate the robustness of D2L against noisy labels under varying noise rates (0%, 20%, 40%, and 60%) on several benchmark datasets, comparing to state-of-the-art baselines for noisy label learning.

Experimental Setup: Experiments were conducted on several benchmark datasets: MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), CIFAR-10 (Krizhevsky & Hinton, 2009) and CIFAR-100 (Krizhevsky & Hinton, 2009). We used a LeNet-5 network (LeCun et al., 1998) for MNIST, a 6-layer CNN for SVHN, a 12-layer CNN for CIFAR-10 and a ResNet-44 network (He et al., 2016) for CIFAR-100. All networks were trained using SGD with momentum 0.9, weight decay 10^{-4} and an initial learning rate of 0.1. The learning rate is divided by 10 after epochs 20 and 40 for MNIST/SVHN (50 epochs in total), after epochs 40 and 80 for CIFAR-10 (120 epochs in total), and after epochs 80, 120 and 160 for CIFAR-100 (200 epochs in total) (Huang et al., 2016). Simple data augmentations (width/height shift and horizontal flip) were applied on CIFAR-10 and CIFAR-100. Noisy labels were generated as described in Section 5.1. On a particular dataset, the compared methods differ only in their loss functions — they share the same CNN architecture, regularizations (batch normalization and max pooling), and the number of training epochs. We repeated the experiments 5 times with different random seeds for network initialization and label noise generation.

Results: We report the mean test accuracy and standard deviation over 5 repetitions of the experiments in Table 1. D2L outperforms its competitors consistently across all datasets and across all noise rates tested. In particular, the performance gap between D2L and its competitors increases

as the noise rate is increased from 20% to 60%. We also note that as the noise rate increases, the accuracy drop of D2L is the smallest among all models. Even with 60% label noise, D2L can still obtain a relatively high classification accuracy, which indicates that D2L may have the potential to be an effective strategy for semi-supervised learning.

6. Discussion and Conclusion

In this paper, we have investigated the generalization behavior of DNNs for noisy labels in terms of the intrinsic dimensionality of local subspaces. We observed that dimensional compression occurs early in the learning process, followed by dimensional expansion as the process begins to overfit. Employing a simple measure of local intrinsic dimensionality (LID), we proposed a Dimensionality-Driven Learning (D2L) strategy for avoiding overfitting that identifies the learning epoch at which the transition from dimensional compression to dimensional expansion occurs, and then suppresses the subsequent dimensionality expansion. D2L delivers very strong classification performance across a range of scenarios with high proportions of noisy labels.

We believe that dimensionality-based analysis opens up new directions for understanding and enhancing the behavior of DNNs. Theoretical formulation of DNN subspace dimensionality, and investigation of the effects of data augmentation and regularization techniques such as batch normalization (Ioffe & Szegedy, 2015) and dropout (Srivastava et al., 2014) are possible directions for future research. Another open issue is the investigation of how other forms of noise such as adversarial or corrupted inputs and asymmetric label noise can affect local subspace dimensionality and DNN learning behavior.

References

Amsaleg, Laurent, Chelly, Oussama, Furon, Teddy, Girard, Stéphane, Houle, Michael E., Kawarabayashi, Ken-ichi, and Nett, Michael. Estimating local intrinsic dimensionality. In *SIGKDD*, 2015.

Amsaleg, Laurent, Bailey, James, Barbe, Dominique, Erfani, Sarah M., Houle, Michael E., Nguyen, Xuan Vinh, and Radovanović, Miloš. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *WIFS*, 2017.

Arpit, Devansh, Jastrzebski, Stanisaw, Ballas, Nicolas, Krueger, David, Bengio, Emmanuel, Kanwal, Maxinder S., Maharaj, Tegan, Fischer, Asja, Courville, Aaron, Bengio, Yoshua, and Lacoste-Julien, Simon. A closer look at memorization in deep networks. In *ICML*, 2017.

Belkin, Mikhail, Matveeva, Irina, and Niyogi, Partha. Regularization and semi-supervised learning on large graphs. In *COLT*, 2004.

Belkin, Mikhail, Niyogi, Partha, and Sindhwani, Vikas. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434, 2006.

Coles, Stuart, Bawa, Joanna, Trenner, Lesley, and Dorazio, Pat. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

Ghosh, Aritra, Kumar, Himanshu, and Sastry, PS. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

Goldberger, Jacob and Ben-Reuven, Ehud. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *CVPR*, 2016.

Hill, Bruce M. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3 (5):1163–1174, 1975.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E., Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N., and Kingsbury, Brian. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29 (6):82–97, 2012.

Houle, Michael E. Dimensionality, discriminability, density & distance distributions. In *ICDMW*, pp. 468–473, 2013.

Houle, Michael E. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In *SISAP*, 2017a.

Houle, Michael E. Local intrinsic dimensionality II: multivariate analysis and distributional support. In *SISAP*, 2017b.

Houle, Michael E., Kashima, Hisashi, and Nett, Michael. Generalized expansion dimension. In *ICDMW*, 2012.

Huang, Gao, Sun, Yu, Liu, Zhuang, Sedra, Daniel, and Weinberger, Kilian Q. Deep networks with stochastic depth. In *ECCV*, 2016.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Jiang, Lu, Zhou, Zhengyuan, Leung, Thomas, Li, Li-Jia, and Fei-Fei, Li. Mentornet: Regularizing very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

Karger, David R. and Ruhl, Matthias. Finding nearest neighbors in growth-restricted metrics. In *STOC*, 2002.

Krizhevsky, Alex and Hinton, Geoffrey. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Krueger, David, Ballas, Nicolas, Jastrzebski, Stanislaw, Arpit, Devansh, Kanwal, Maxinder S, Maharaj, Tegan, Bengio, Emmanuel, Fischer, Asja, and Courville, Aaron. Deep nets don't learn via memorization. In *ICLR*, 2017.

Larsen, Jan, Nonboe, L., Hintz-Madsen, Mads, and Hansen, Lars Kai. Design of robust neural network classifiers. In *ICASSP*, volume 2, 1998.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521(7553):436–444, 2015.

Levina, Elizaveta and Bickel, Peter J. Maximum likelihood estimation of intrinsic dimension. In *NIPS*, 2005.

Li, Yuncheng, Yang, Jianchao, Song, Yale, Cao, Liangliang, Luo, Jiebo, and Li, Jia. Learning from noisy labels with distillation. In *ICCV*, 2017.

Ma, X., Li, B., Wang, Y., Erfani, S., Wijewickrema, S. Schoenebeck, G., Houle, M. E. Song, D., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, 2018.

Maaten, Laurens van der and Hinton, Geoffrey. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Natarajan, Nagarajan, Dhillon, Inderjit S., Ravikumar, Pradeep K., and Tewari, Ambuj. Learning with noisy labels. In *NIPS*, 2013.

Netzer, Yuval, Wang, Tao, Coates, Adam, Bissacco, Alessandro, Wu, Bo, and Ng, Andrew Y. Reading digits in natural images with unsupervised feature learning. In *NIPSW*, 2011.

Neyshabur, Behnam, Tomioka, Ryota, and Srebro, Nathan. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Patrini, Giorgio, Rozza, Alessandro, Menon, Aditya, Nock, Richard, and Qu, Lizhen. Making neural networks robust to label noise: a loss correction approach. In *CVPR*, 2017.

Reed, Scott, Lee, Honglak, Anguelov, Dragomir, Szegedy, Christian, Erhan, Dumitru, and Rabinovich, Andrew. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Roweis, Sam T. and Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Saxe, Andrew M., Bansal, Yamini, Dapello, Joel, Advani, Madhu, Kolchinsky, Artemy, Tracey, Brendan D., and Cox, David D. On the information bottleneck theory of deep learning. In *ICLR*, 2018.

Shwartz-Ziv, Ravid and Tishby, Naftali. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Silver, David, Huang, Aja, Maddison, Chris J., Guez, Arthur, Sifre, Laurent, Van Den Driessche, George, Schrittwieser, Julian, Antonoglou, Ioannis, Panneershelvam, Veda, Lanctot, Marc, Dieleman, Sander, Grewe, Dominik, Nham, John, Kalchbrenner, Nal, Sutskever, Ilya, Lillicrap, Timothy, Leach, Madeleine, Kavukcuoglu, Koray, Graepel, Thore, and Hassabis, Demis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

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

Sukhbaatar, Sainbayar and Fergus, Rob. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.

Sukhbaatar, Sainbayar, Bruna, Joan, Paluri, Manohar, Bourdev, Lubomir, and Fergus, Rob. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Vahdat, Arash. Toward robustness against label noise in training deep discriminative neural networks. In *NIPS*, 2017.

Veit, Andreas, Alldrin, Neil, Chechik, Gal, Krasin, Ivan, Gupta, Abhinav, and Belongie, Serge. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, 2017.

Wang, Yisen, Liu, Weiyang, Ma, Xingjun, Bailey, James, Zha, Hongyuan, Song, Le, and Xia, Shu-Tao. Iterative learning with open-set noisy labels. In *CVPR*, 2018.

Xiao, Tong, Xia, Tian, Yang, Yi, Huang, Chang, and Wang, Xiaogang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.