ICML | 2018
Thirty-fifth International Conference on Machine Learning

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, Sudanthi Wijewickrema[1], James Bailey[1]
(* Equal Contribution)

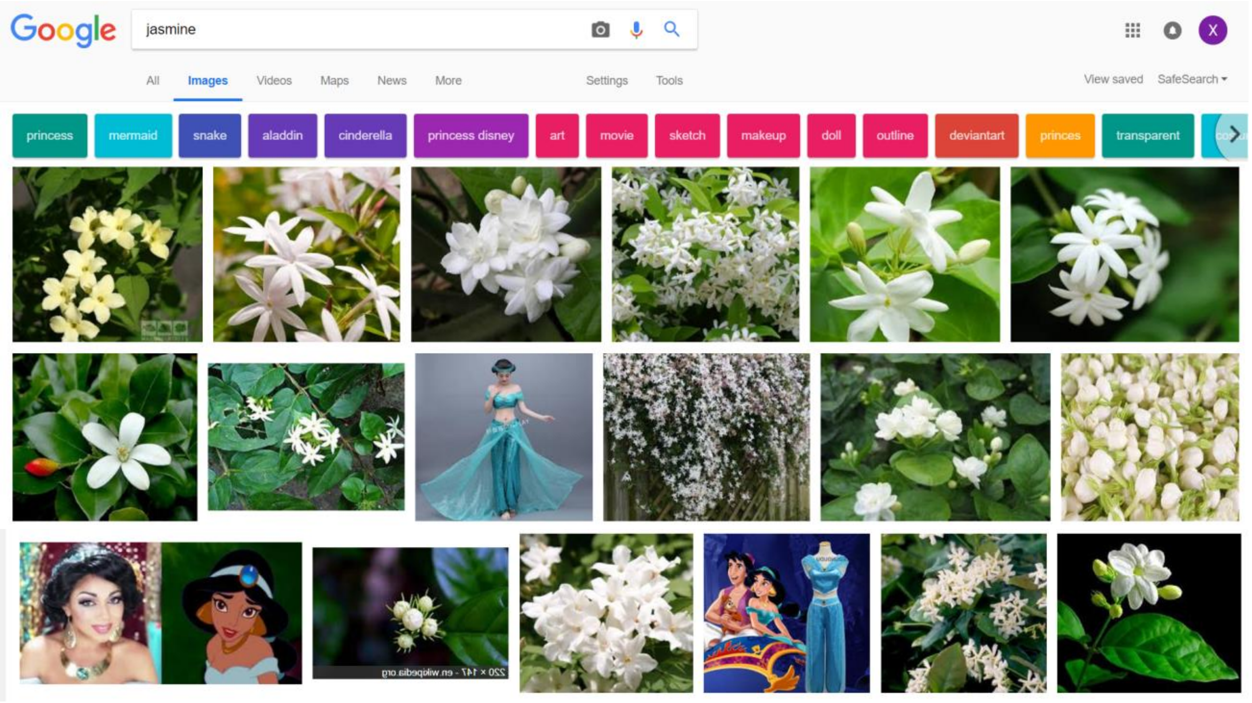
[1]The University of Melbourne; [2]Tsinghua University;
[3]National Institute of Informatics, Japan

Purpose of this paper:

- Investigating learning behaviours of deep neural networks (DNNs) on data with noisy (incorrect) labels.
- Exploring learning strategies that can robustly train DNNs on data with noisy labels.

Noisy label learning:

- Large-scale annotated datasets are important for deep learning.
- Data labelling can be costly, time-consuming and error-prone.
- Webly-searched and crowd-sourcing annotated data often contain noisy labels.



Related work:

- Understanding learning behaviours of DNNs
  - Zhang et al. 2017
    - DNNs overfit to random labels, by case-by-case memorization.
  - Krueger et al. 2017
    - DNNs exhibit different styles on clean *vs* noisy labels, and they do not learn by memorization.
  - Arpit et al. 2017
    - DNNs learn by: 1) simple pattern learning, then 2) label memorization.

Related work:

- DNNs and noisy label learning
  - Probabilistic modelling of label noise:
    Larsen et al. 1998, Natarajan et al. 2013, Sukhbaatar et al. 2014
  - Label inferring or propagation:
    Xiao et al. 2015, Vahdat 2017, Veit et al. 2017, Li et al. 2017.
  - Loss correction:
    Patrini et al. 2017, Sukhbaatar et al. 2014, Goldberger et al. 2017, Reed et al. 2014.
  - Sample reweighting (ICML 2018): Jiang et al. 2018, Ren et al. 2018.
  - Contrastive learning (CVPR 2018): Wang et al. 2018.

Challenges of noisy label learning:

- Difficult to determine whether or not the learning process is noisy.
- Difficult to train DNNs of good generalization with noisy labels.

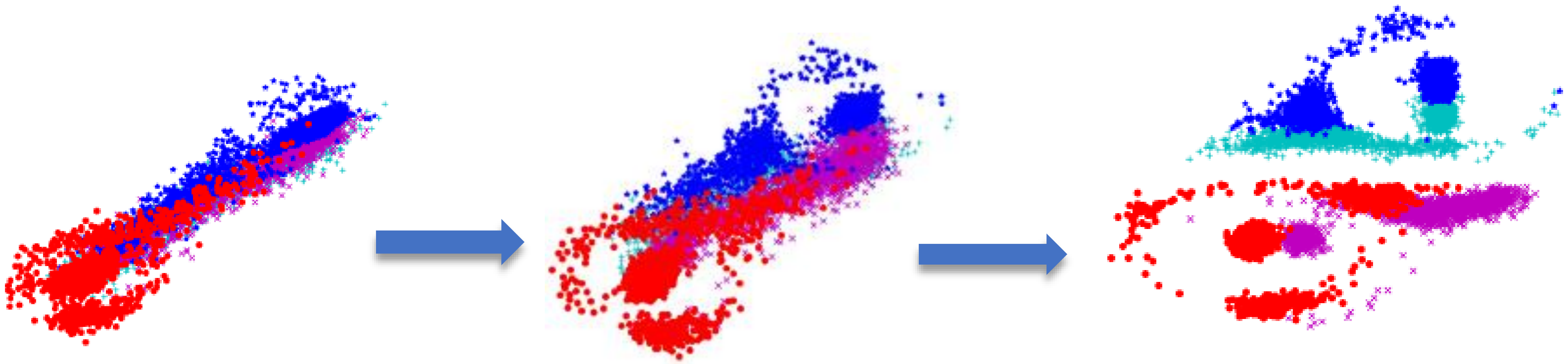
Our contributions:

- We identify two distinctive learning behaviours of DNN throughout training:
  a. Clean labels: **dimensionality compression;**
  b. Noisy labels: **dimensionality shift**, from **compression** to **expansion**.

- We propose Dimensionality-Driven Learning (D2L) to avoid dimensionality expansion, so as to avoid overfitting to noisy labels.

Dimensionality of DNN feature spaces:

We investigate the relation between dimensionality and noisy label overfitting.



1st layer — 10th layer — 20th layer

Our Intuition:

If learning is a compression/fitting process, then

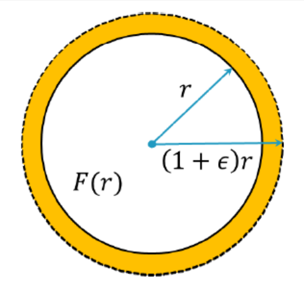
a) clean classes of data can be easily compressed to simpler manifold with lower intrinsic dimensionality;
b) noisy classes of data are hard to compress.

Local Intrinsic Dimensionality (LID):

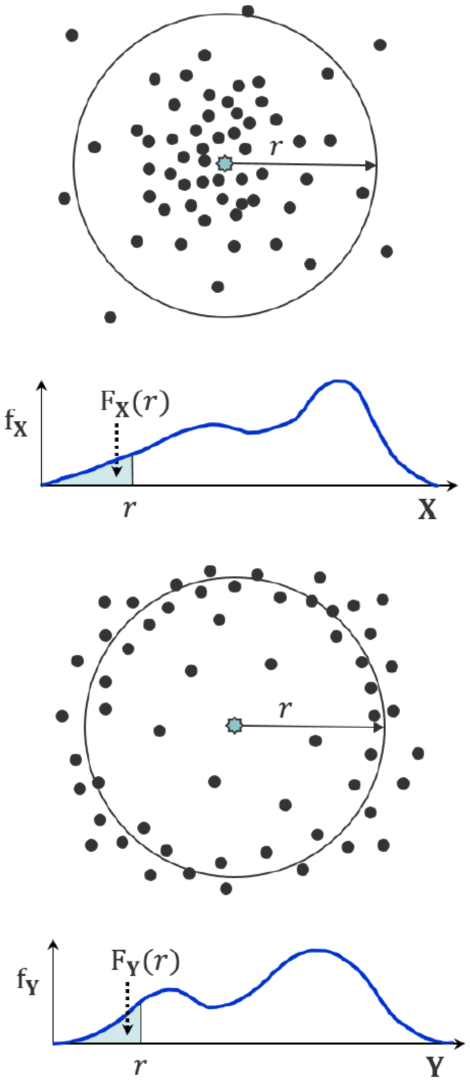
Definition (Local Intrinsic Dimensionality)

Given a data sample $x \in X$, let $r > 0$ be a random variable denoting the distance from x to other data samples. The *local intrinsic dimension* of x at distance r is

$$\mathrm{LID}_F(r) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln(F((1+\epsilon)\cdot r)/F(r))}{\ln(1+\epsilon)} = \frac{r \cdot F'(r)}{F(r)},$$



wherever the limit exists.



- $F(r)$: **cdf** of the distribution of distances to data from a given reference location.
- $\mathrm{LID}_F(r)$: measures growth rate of $F(r)$ as the radius r expands (*Houle 2017a*).

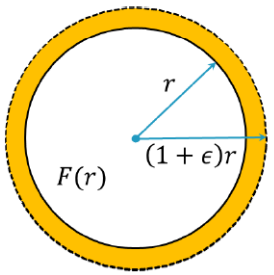
Estimation of LID:

Estimators of LID already available:

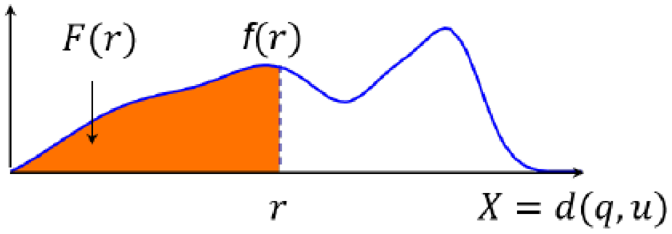
- Hill (MLE) estimator (*Hill 1975, Amsaleg et al. 2015*):

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{r_i(x)}{r_k(x)}\right)^{-1},$$

r_i is the distance of x to its i^{th} nearest neighbour.

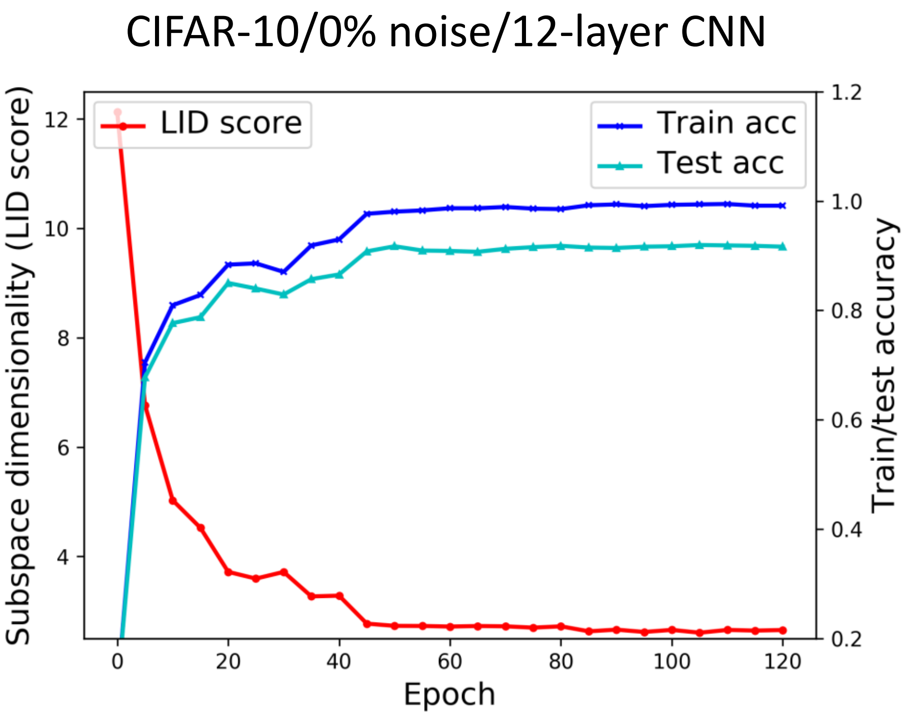


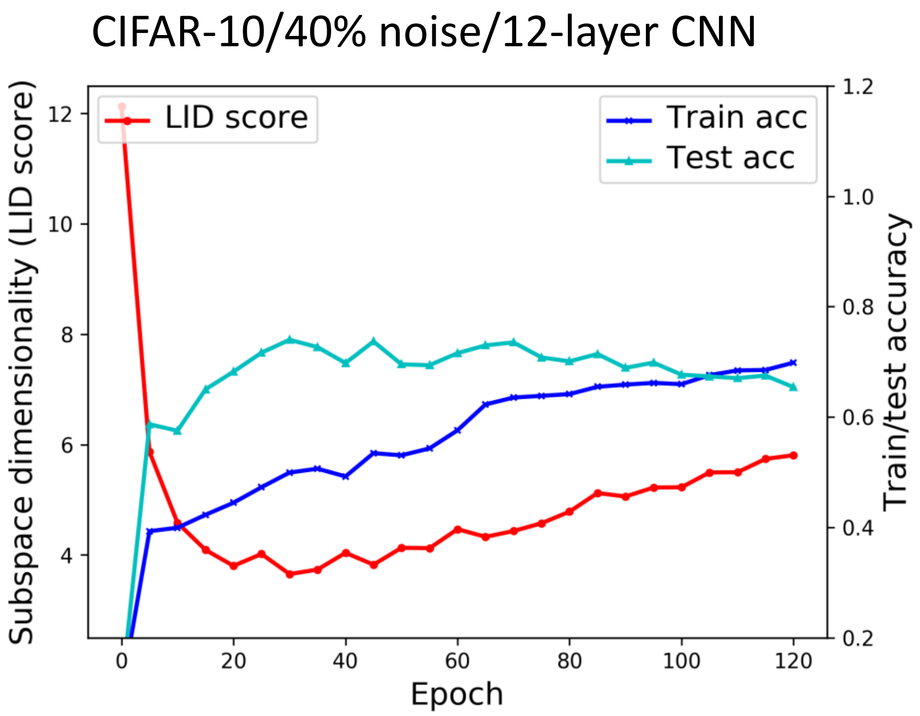
- Based on Extreme Value Theory:
  - Nearest neighbor distances are extreme events.
  - Lower tail distribution follows Generalized Pareto Distribution (GPD).



- Other estimators: e.g. *Amsaleg et al. 2015, Levina & Bickel 2005*.

Learning with clean *vs* noisy labels (CIFAR-10):





- ❑ Clean labels: decreasing subspace dimensionality: **compression**.
- ❑ Noisy Labels: dimensionality shift from **compression** to **expansion.**
- ❑ Dimensionality expansion indicates overfitting to noisy labels.

Dimensionality-Driven Learning (D2L):

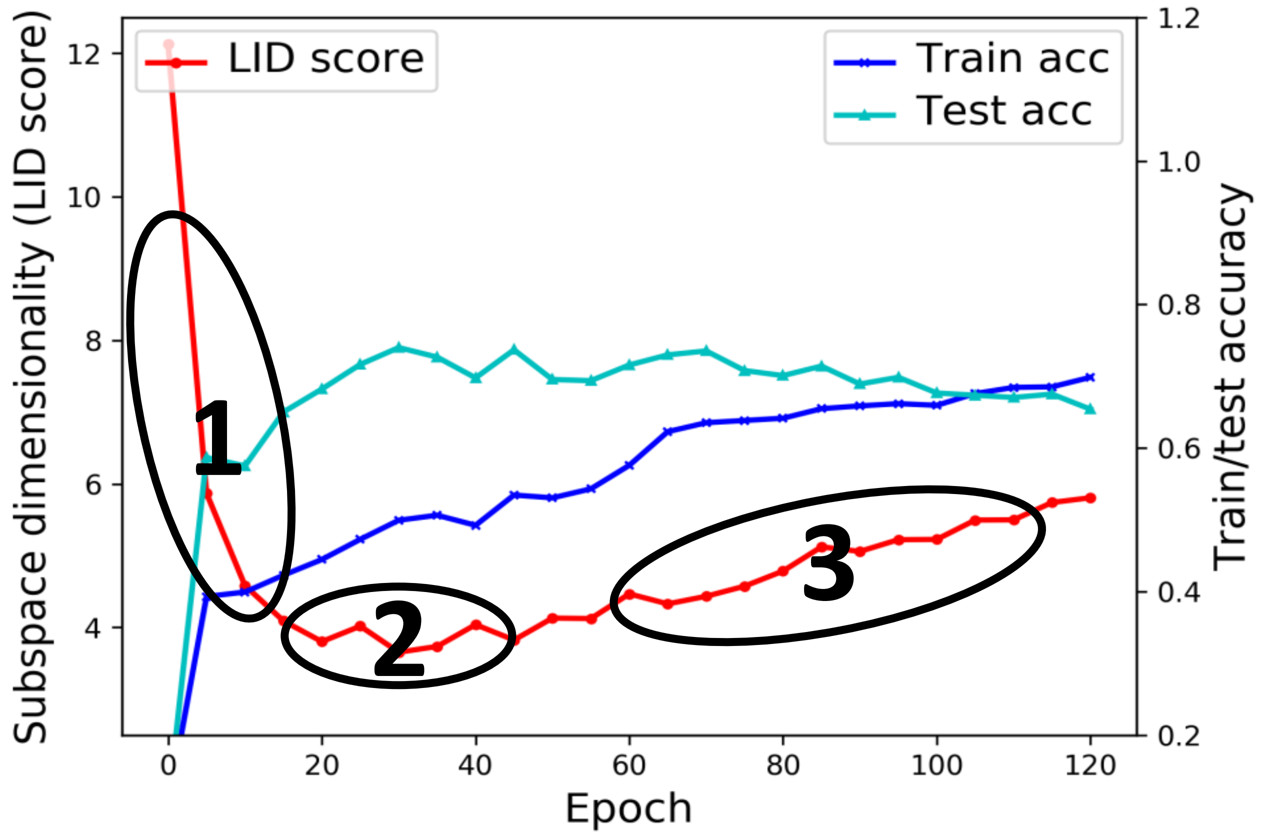
We avoid dimensionality expansion phase by using **LID-adapted labels.**

- Our proposed loss function kicks in only after dimensional shift to expansion has been detected.
- Our observation: The higher the dimensionality, the more noisy the labels.
- Original labels should not be fully trusted.

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{y_n^*} y_n^* \log P(y_n^*|x_n)\,, \qquad y^* = \alpha_i y + (1-\alpha_i)\hat{y}, \qquad \alpha_i = \exp\left(-\lambda \frac{\widehat{\mathrm{LID}}_i}{\min_j^{i-1}\widehat{\mathrm{LID}}_j}\right)$$

- ○ y_n^*: blending of original (y) and predicted ($\hat{y}$) label values.
- ○ α_i: LID-based weighting for the label interpolation.
- ○ $\widehat{\mathrm{LID}}_i$: average of LID scores over 10 batches, at i^{th} epoch.
- ○ $\lambda = i/T$: training progress (T: total number of epochs).

Work flow of D2L:



1. Early stage of compression: rely on raw labels.
2. Turing point: dimensionality shift from **compression** to **expansion.**
3. Later stage of dimensionality expansion: rely on predicted labels.

Empirical evaluation of D2L:

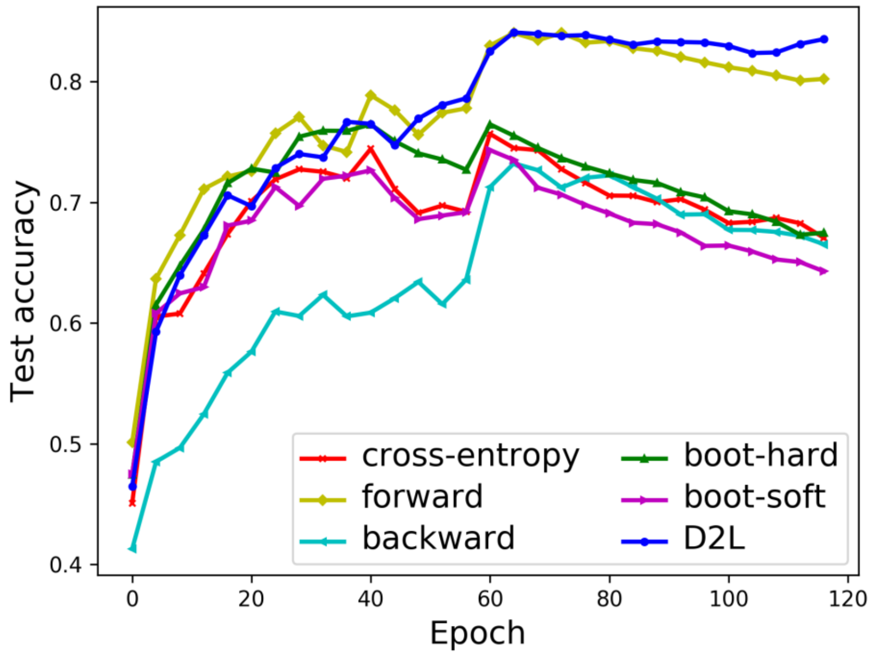
- Setting:

  A 12-layer CNN on CIFAR-10 with 40%/60% random label noise, SGD trained for 120 epochs.

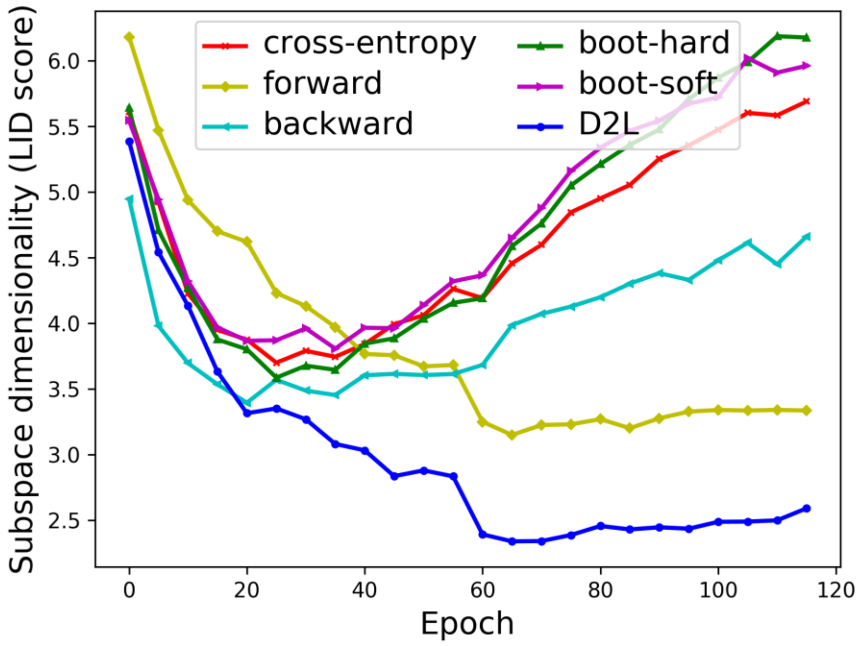
- Compared training strategies:
  a) Backward/Forward (Patrini et al. 2017).
  b) Boot-hard/Boot-soft (Reed et al. 2014).
  c) Cross entropy (standard definition).

- Understand different training methods from 3 viewpoints:
  a) Dimensional complexity of the learned subspaces (measured by average LID score).
  b) Complexity of the learned hypothesis (measured by Critical Sample Ratio, Arpit et al. 2017).
  c) Quality of learned representation (by visualization).

Empirical evaluation of D2L – subspace dimensionality:



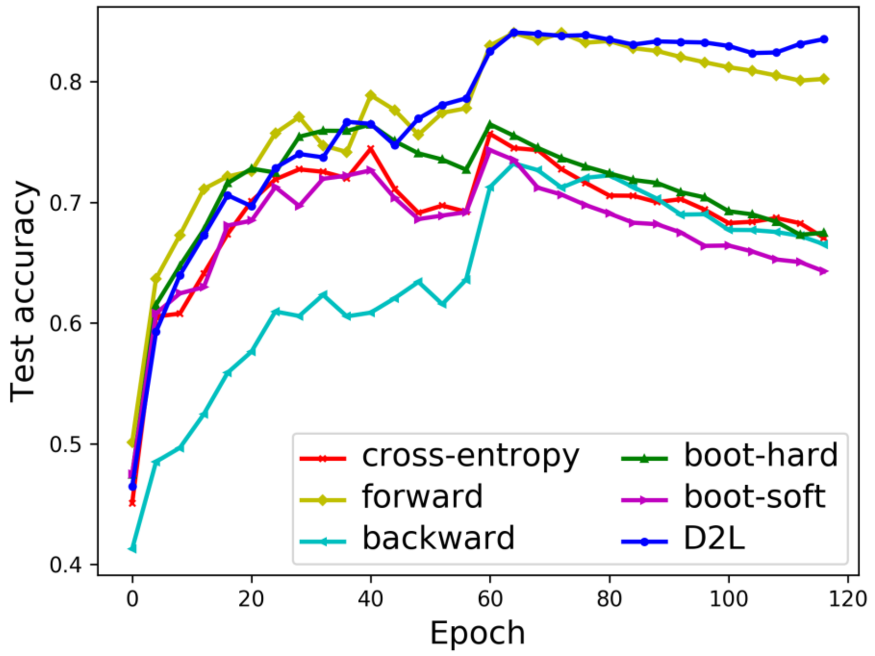
Test accuracy: CIFAR-10/40% noise



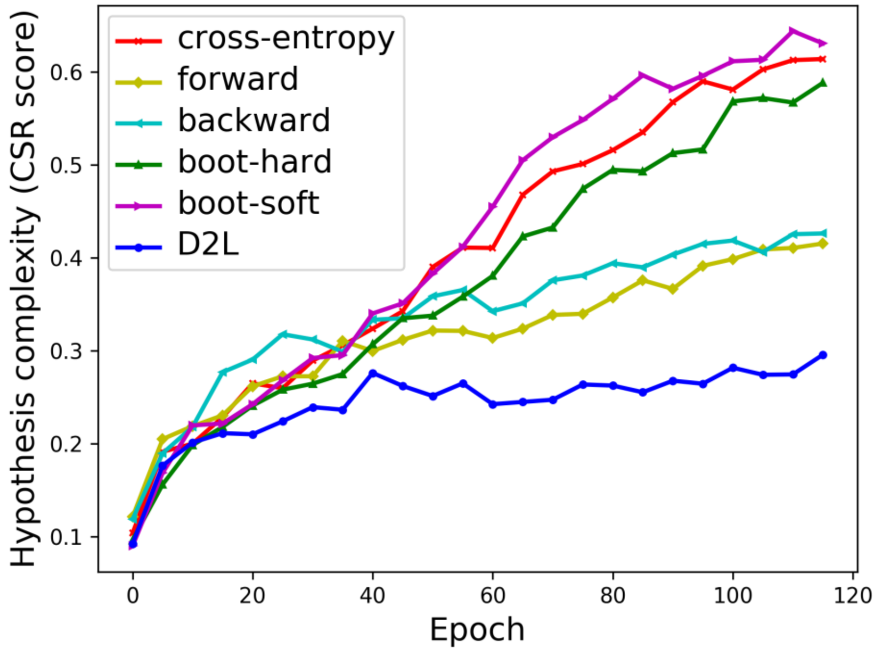
Subspace Complexity: CIFAR-10/40% noise

- D2L learns simpler subspaces with better test accuracy.

Empirical evaluation of D2L – hypothesis complexity:



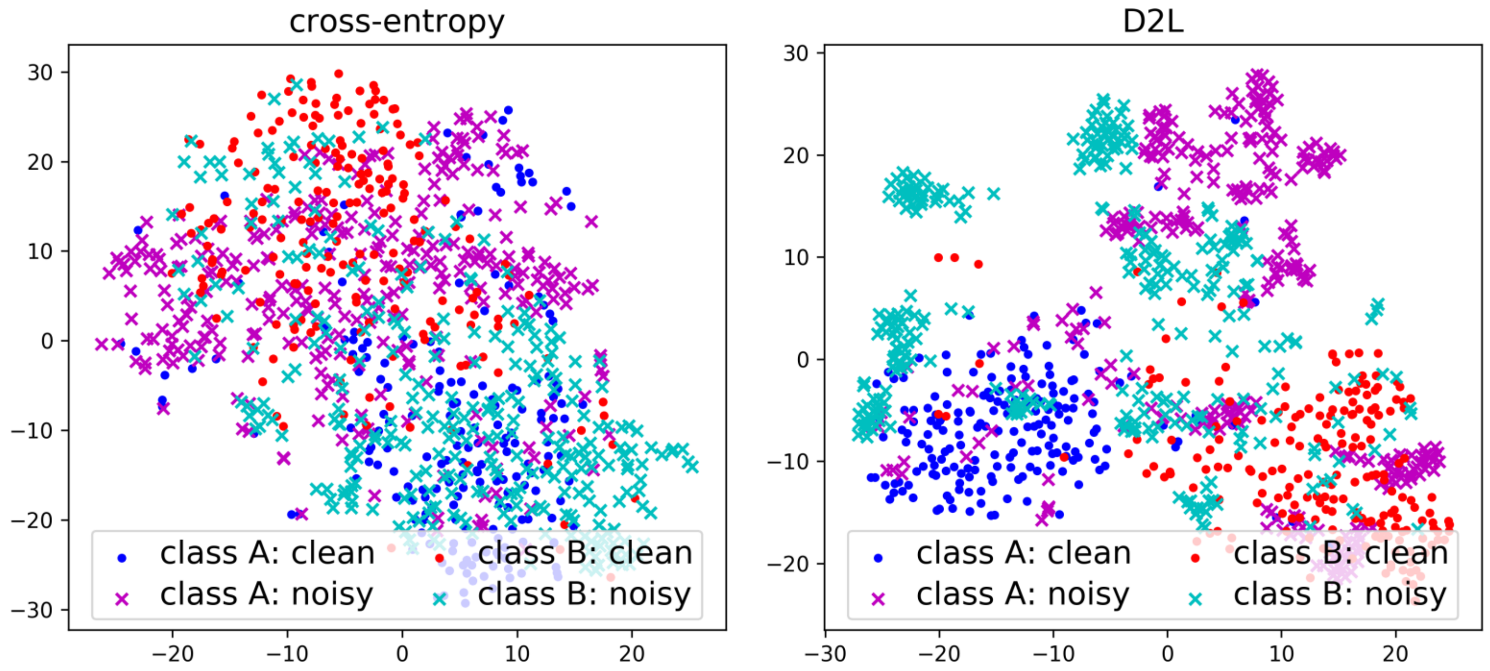
Test accuracy: CIFAR-10/40% noise



Hypothesis complexity: CIFAR-10/40% noise

- D2L learns simpler hypothesis.
- The Critical Sample Ratio (Arpit et al. 2017) also indicates adversarial robustness.

Empirical evaluation of D2L – representation:



Class A: 'airplane'
Noisy 'airplane': 9 other classes

Class B: 'cat'
Noisy 'cat': 9 other classes

Noise rate: 60% random.

- ❑ D2L learns more fragmented representation, globally scattered, locally clustered.
- ❑ Small scattered clusters indicate noisy samples from 9 different classes.

Robustness to noisy labels:

Table 1: Test accuracy (%) ± std on MNIST, SVHN, CIFAR-10 and CIFAR-100.

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%	82.66±1.8	96.45±0.4	84.69±1.2	80.69±2.2	83.50±1.2	**98.84±0.0**
	40%	60.14±3.9	88.90±0.9	64.89±1.0	60.49±1.6	59.19±1.8	**98.49±0.0**
	60%	38.51±3.7	72.88±1.6	42.83±3.3	40.45±1.6	39.04±3.0	**94.73±1.2**
SVHN	0%	90.12±0.3	90.22±0.1	90.16±0.2	89.47±0.0	89.26±0.0	**90.32±0.0**
	20%	76.10±0.9	85.51±0.7	74.61±0.5	76.10±0.3	75.26±0.2	**87.63±0.1**
	40%	57.92±1.4	74.09±0.7	59.15±0.8	58.25±0.2	58.30±0.3	**84.68±0.6**
	60%	36.54±0.62	60.57±0.6	50.54±0.7	42.51±1.2	37.21±0.9	**80.92±0.8**
CIFAR-10	0%	90.39±0.6	**90.27±0.0**	89.03±1.2	89.06±0.9	89.46±0.6	89.41±0.2
	20%	73.12±1.3	84.61±0.3	79.41±0.1	79.19±0.4	82.21±0.4	**85.13±0.6**
	40%	65.07±3.3	81.84±0.1	74.69±1.3	76.67±0.8	75.81±0.3	**83.36±0.5**
	60%	52.55±1.6	72.41±0.7	40.42±0.4	70.57±0.3	68.32±0.6	**72.84±0.6**
CIFAR-100	0%	68.20±0.4	68.54±0.1	68.48±0.2	68.31±0.2	67.89±0.2	**68.60±0.3**
	20%	52.88±0.5	60.25±0.2	58.74±0.3	58.49±0.4	57.32±1.1	**62.20±0.5**
	40%	42.85±0.3	51.27±0.3	45.42±0.6	46.44±0.7	45.77±1.1	**53.01±0.7**
	60%	30.09±0.2	44.22±0.7	34.49±1.1	42.65±0.9	40.29±1.2	**45.21±0.4**

- D2L demonstrated strong performance for different noise rates.

Conclusion:

- We identify distinctive DNN learning behaviours on clean *vs* noisy labels.
- Subspace dimensionality expansion is associated with overfitting to noisy labels.
- D2L can learn low-dimensional subspaces, simpler hypotheses and high-quality representations.

Future work:

- Different theoretical formulations of subspace dimensionality.
- Explore D2L on other forms of noise, other network architectures.
- Investigation of the effects of data augmentation and regularization techniques such as batch normalization and dropout.

References:

[1] Michael E. Houle. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In SISAP, pp. 64–79, 2017a.

[2] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In SIGKDD, pp.29–38. ACM, 2015.

[3] 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.

[4]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.

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

[6] 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.

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

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

[9] 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.

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

Thank you!

Poster: this afternoon.
Thu Jul **12th** 06:15 -- 09:00 PM