LEARNING FROM NOISY LABELS WITH DEEP NEURAL NETWORKS

SAINBAYAR SUKHBAATAR
ROB FERGUS
NEW YORK UNIVERSITY

CONTENTS

INTRODUCTION

Supervised deep networks work very well If you have huge labeled data

- Ex) ImageNet

However, hand labeling is expensive and noisy

Noisy/weak labels are easy to obtain

- Billions of images from image search engines
- user hashtags

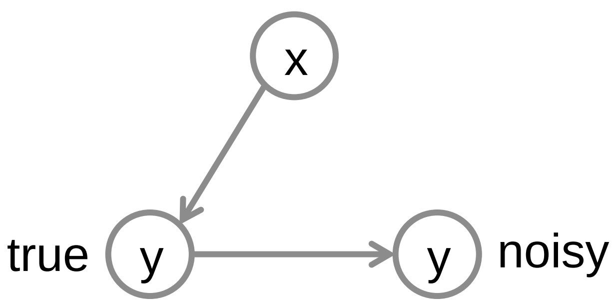
How to train deep networks on noisy labels?

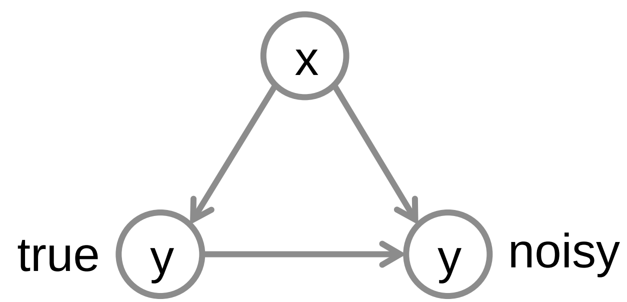
- Besides from data cleaning methods

LABEL NOISE

Assumption on noise:

1. **Most simple: noise is completely random**
2. **In the middle: noise is random given the true label**
3. **Most complex: noise depends on actual input itself**





BOTTOM-UP NOISE MODEL

Noisy label

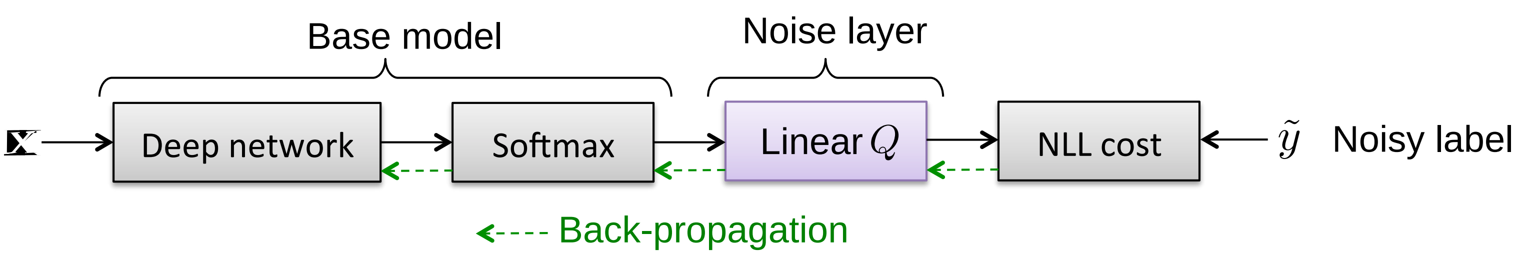
Model parameters (weights, biases)

True label (unknown)

Noise distribution (also unknown)

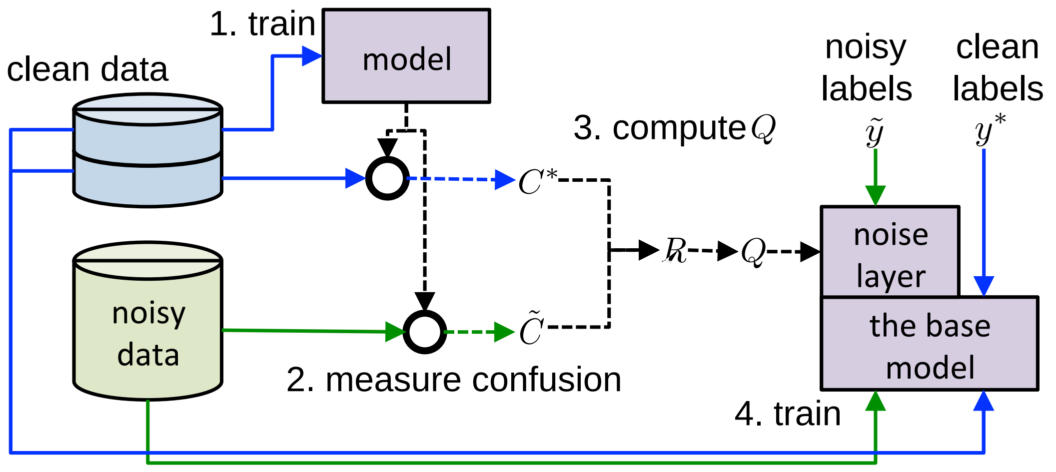
$$p(\tilde{y} = j|\mathbf{x}, \theta) = \sum_i p(\tilde{y} = j|y^* = i)p(y^* = i|\mathbf{x}) = \sum_i q_{ji} p(y^* = i|\mathbf{x}, \theta)$$

- Noise distribution adapts network's output so it would better match to noisy labels in training data.
- It can be implemented by a simple linear layer on top of the softmax



Mnih et al, 2012 (binary classification); Bootkrajang et al, 2012 (logistic regression)

ESTIMATING NOISE DISTRIBUTION USING CLEAN DATA



Confusion on noisy data

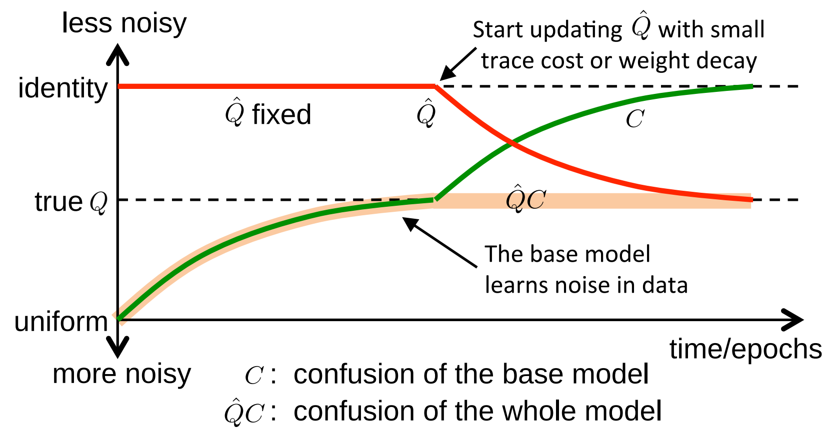
Confusion on clean data

$$C^* R = \tilde{C} \implies R = {C^*}^{-1} \tilde{C}$$

$$p(\tilde{y} = j | y^* = i) = \frac{p(y^* = i | \tilde{y} = j) p(\tilde{y} = j)}{p(y^* = i)} \implies q_{ji} = \frac{r_{ij} p(\tilde{y} = j)}{p(y^* = i)}$$

LEARNING NOISE DISTRIBUTION

- Q is a linear layer → can use backprop
- After each update, project Q back to probability matrix (column sums to 1)
- How to prevent the base model from learning noise?
- Make Q noisy → pushes confusion of the base model to identity



Theorem 1. *In the following optimization problem, the only global minimum is $Q = Q$ and $C = I$. (where Q, $\hat{Q}$ and C are probability matrices).*

$$\underset{\hat{Q},C}{\text{minimize}} \quad tr(\hat{Q}) \quad \text{subject to} \quad \hat{Q}C = Q,\ \hat{q}_{ii} > \hat{q}_{ij},\ q_{ii} > q_{ij} \quad \text{for } \forall i, j \neq i.$$

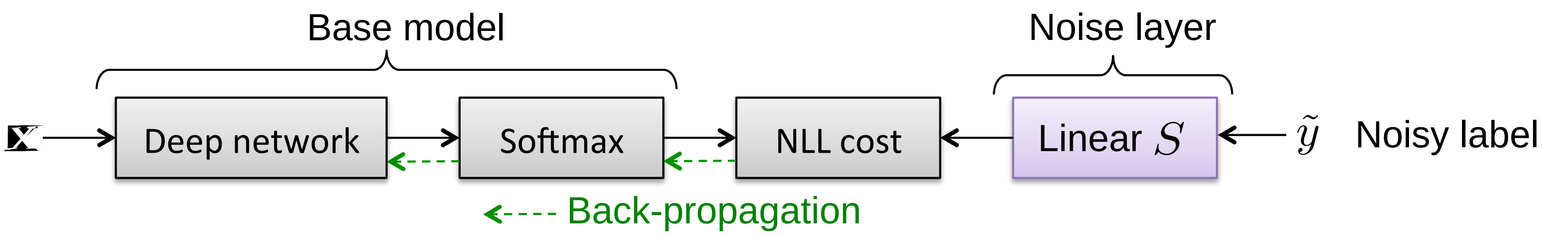
Proof. Let Q^* be the global solution. Then

$$tr(Q) = tr(Q^*C) = \sum_i \Big(\sum_j q^*_{ij} c_{ji}\Big) \leq \sum_i \Big(\sum_j q^*_{ii} c_{ji}\Big) = \sum_i q^*_{ii} \Big(\sum_j c_{ji}\Big) = \sum_i q^*_{ii} = tr(Q^*)$$

The equality will only hold true only when $C = I$. Therefore, $Q^* = \hat{Q}$. □

TOP-DOWN NOISE MODEL

- Modify cost function → unbiased classification (Natarajan et al, NIPS 2013)
- Same as changing noisy labels with matrix S
- S is inverse of Q (or at least SQ = Id + constant)
- Could not learn S using backprop (degenerate solution)



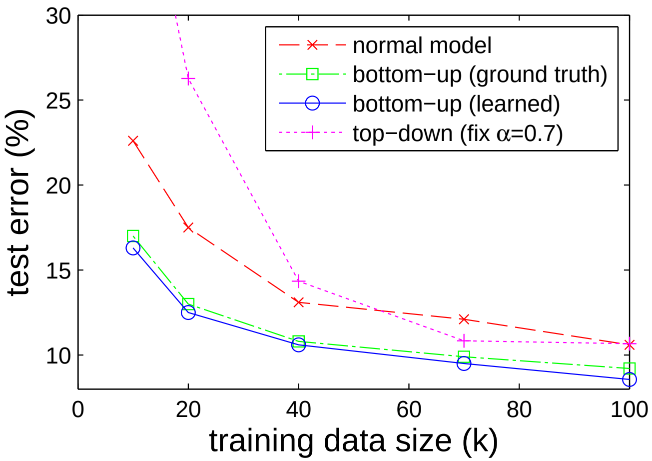
SIMPLE WEIGHTING TRICK

When there are clean and noisy training data, put less weight on the noisy labeled images

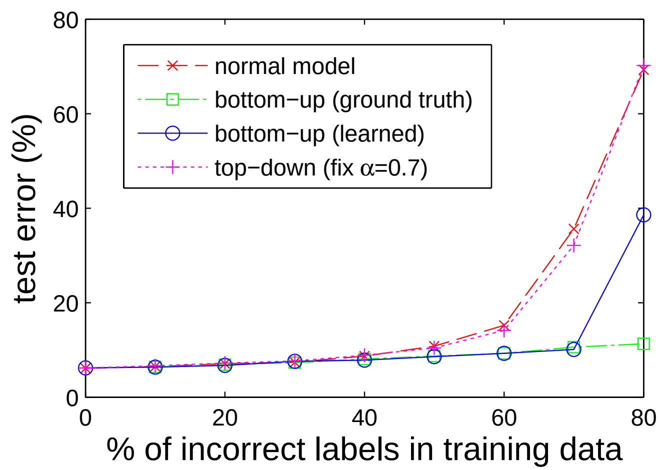
$$\mathcal{L}(\theta) = \frac{1}{N_c + N_n}\left(\sum_{n=1}^{N_c} \log p(y = y_n | \mathbf{x}_n, \theta) + \gamma \sum_{n=1}^{N_n} \log p(\tilde{y} = \tilde{y}_n | \tilde{\mathbf{x}}_n, \theta)\right)$$

EXPERIMENTS: DELIBERATE LABEL NOISE

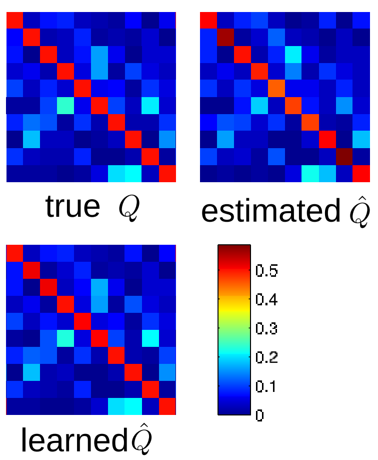
- **Street-view house number dataset (SVHN)**
- **Deliberately add noise to training data**
  - Randomly change labels with fixed probability (not uniform)
- **Base model with three convolutional layers (18% CIFAR10)**



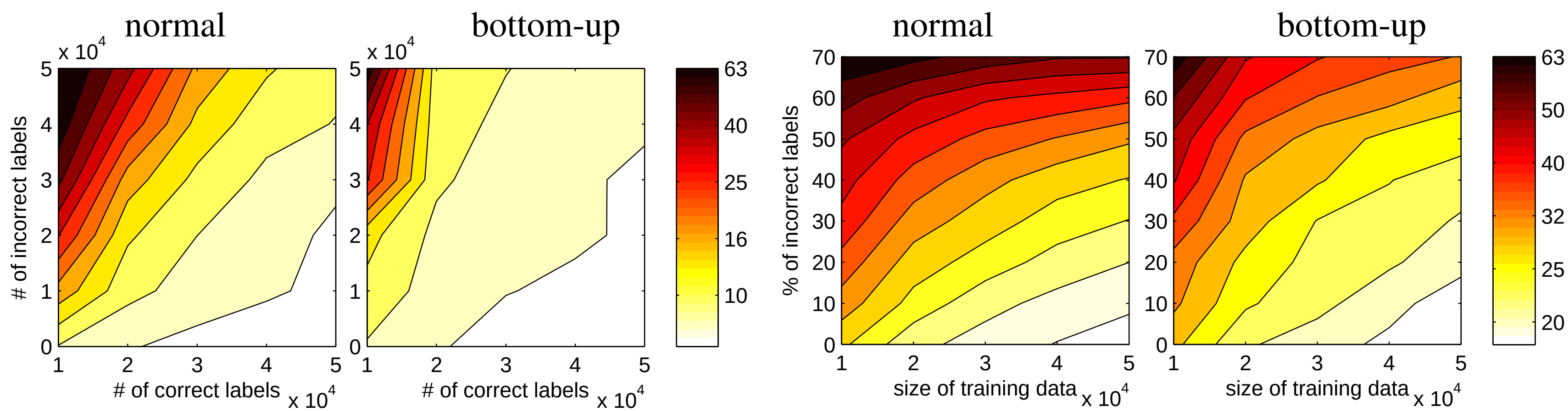
(a)



(b)



(c)



(a) SVHN (b) CIFAR-10

Cifar-10 clean + noisy

Training data 50k = clean 20k + 30k noisy

Clean 20k = 10k (for training) + 10k (for measuring confusion)

The final model only trained on noisy 30k

Model	normal	true Q	learned $\hat{Q}$	estimated $\hat{Q}$
Test error (50% noise)	38%	28%	30%	29%
Test error (70% noise)	60%	35%	40%	35%

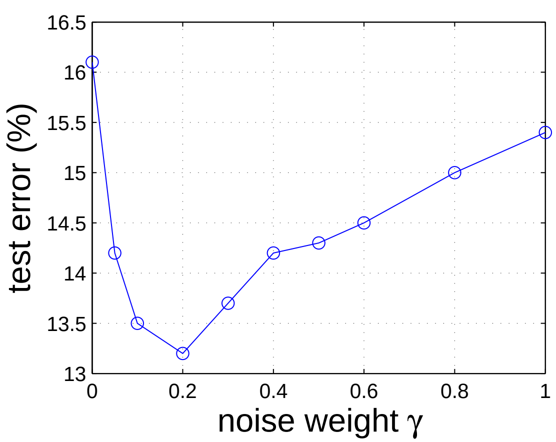
EXPERIMENTS: REAL NOISE CIFAR10 + TINY IMAGES

Train data = clean 50k (from CIFAR10) + noisy 150k (Tiny images/negatives)
Most of the noise is outside noise (not in the 10 categories)

airplane

cat

horse



Model	Extra data	Noisy weight γ	Test error
Conv. net	-	-	16.1%
Conv. net	150k noisy	1	15.4%
Conv. net		0.2	13.2%
Bottom-up		0.2	13.2%
Top-down		0.4	12.5%
Conv. net	150k random	0.2	13.8%

Random images with uniform label acts as regularization

EXPERIMENTS: REAL NOISE IMAGENET + WEB SEARCH IMAGES

Scrapped noisy labeled images from Internet image search using ImageNet keywords

Train data = clean 1.2M (from ImageNet) + noisy 1.4M (from search)

Given noisy label (search keyword)	True label	P(true|given)	note
jaguar (animal)	sport car	0.64	car manufacturer
black swan (animal)	mask	0.27	movie
plane	airliner	0.41	
impala (animal)	convertible	0.47	car name
computer keyboard	space bar	0.33	
maillot (swimsuit)	jersey	0.42	soccer t-shirt
bullfrog	tailed frog	0.61	
Shetland sheepdog	collie	0.74	

Model	Extra data	Noisy weight γ	Top 5 val. error
Krizhevsky et al. [8]	-	-	18.2%
Krizhevsky et al. [8]	15M full ImageNet	-	16.6%
Conv. net	-	-	18.0%
Conv. net	1.4M noisy images from Internet	1	18.1%
Conv. net		0.1	16.7%
Bottom-up (learned)		0.1	16.5%
Bottom-up (estimated)		0.2	16.4%

THANK YOU