Participant Presentation by VISTA

Yuncheng Li
Snap Research

HAJIM
SCHOOL OF ENGINEERING & APPLIED SCIENCES
UNIVERSITY of ROCHESTER

Outline

- Learning from noisy labels with distillation
- Our webvision challenge submission

Learning from Noisy Labels with Distillation (ICCV2017)

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo and Li-Jia Li

HAJIM
SCHOOL OF ENGINEERING
& APPLIED SCIENCES
UNIVERSITY of ROCHESTER

YAHOO!
RESEARCH

Motivation

Learning from noisy labels

Cost: delay, dollars, manpower

Low cost

High cost

Unsupervised Learning

Learning From Noisy Labels

Semi Supervised Learning

Supervised Learning

YFCC100M Dataset

- Yahoo Flickr Creative Commons 100M
- 100,000,000 Flickr photos
- Pixels and metadata:
  - User tags, machine tags, username, title, description, geo tags, device, date
- Visual concept learning with YFCC100M

Types of label noise

Traditional assumption:
Random Classification Noise (RCN): Bird ⇒ Cat

In practice: text ambiguity

Willet: the bird

Willet: the name

Related Work

Bootstrap

Reed, et al. "Training deep neural networks on noisy labels with bootstrapping." *ICLR* 2014.

- Make prediction based on current model:

$$z_k := \mathbb{1}[k = \text{argmax } q_i, i = 1...L]$$

- Update with the modified labels:

$$\mathcal{L}_{hard}(\mathbf{q}, \mathbf{t}) = \sum_{k=1}^{L} [\beta t_k + (1 - \beta) z_k] \log(q_k)$$

Reweight Liu, et al. "Classification with noisy labels by importance reweighting." *IEEE TPAMI 2016*.

- Estimate noise level with a pretrained classifier $P_{D_\rho}(\hat{Y}|X)$

$$\rho_{-\hat{Y}} = \min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X).$$

- Estimate instance importance: (how likely it is a noise sample)

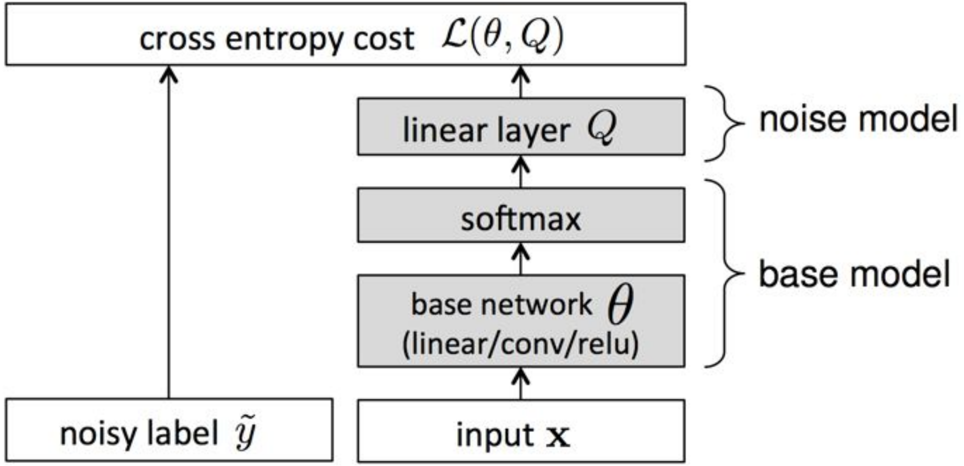
$$\beta(X, \hat{Y}) = \frac{P_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1 - \rho_{+1} - \rho_{-1}) P_{D_\rho}(\hat{Y}|X)};$$

- Retrain the model with the weighted loss

$$\beta(X, \hat{Y}) \ell(f(X), \hat{Y})$$

Noise layer Sukhbaatar, et al. "Learning from noisy labels with deep neural networks." ICLR 2014.

- Add a new layer on top of softmax to “absorb” noise



(a)

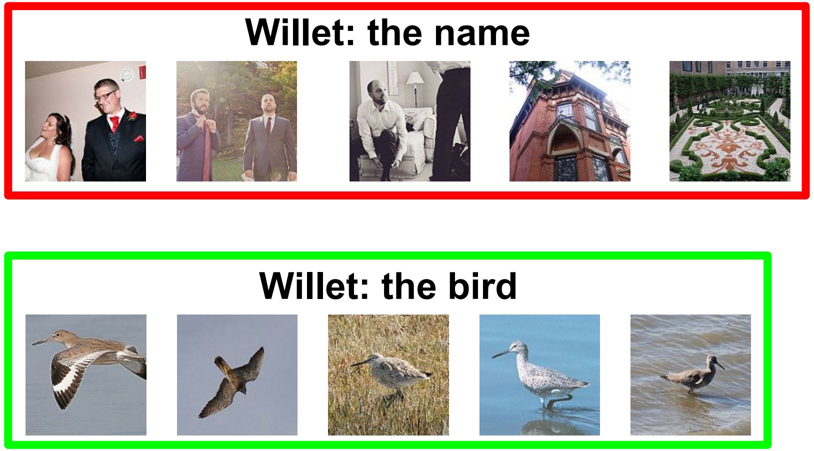
Label smooth Szegedy, et al. "Rethinking the inception architecture for computer vision." *ICLR* 2015

- Modify the label map with smoothed version:

Original label map	0	0	0	1	0	0	0	0	0	0
Smoothed label map	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	0.01

Just do it! Krause, Jonathan, et al. "The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition." *ECCV 2016*.

- Two kinds of noise
  - Cross domain noise
  - Cross category noise
- The cross domain images are not shown in the evaluation
  - This is true only for fine grained classification



Our work

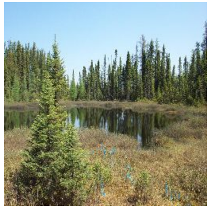
Contributions

1. Learning with knowledge graph to handle label noise
    a. Noise caused by the text ambiguity
    b. Generic visual classifier
2. A new dataset to benchmark learning from noisy labels
    a. Real-world label noise

Semantic knowledge graph

Family: Pinaceae


Fir


Larix_laricina


Spruce


Larch

Order: Hemiptera

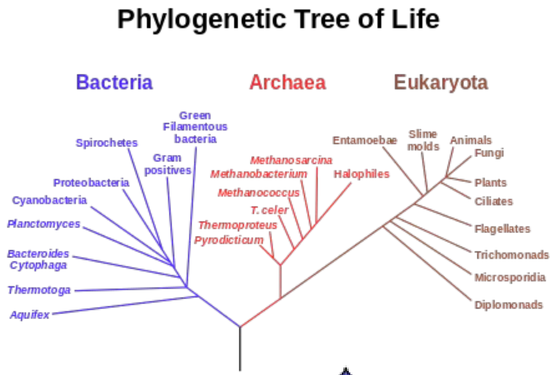

Leafhopper


Aphid

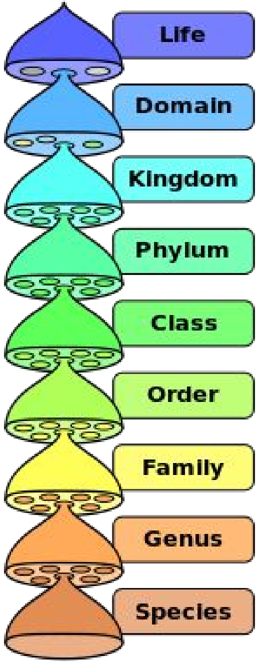

Cicada

Class: Bird

Hummingbird, Ostrich, Tanager, Ruff, Willet, Darter, …



Source: wikipedia



Visual knowledge graph


Bison


Gaur


Takin


Wildebeest


Zebu


Abalone


Clam


Mussel


Oyster


Scallop

A motivating example

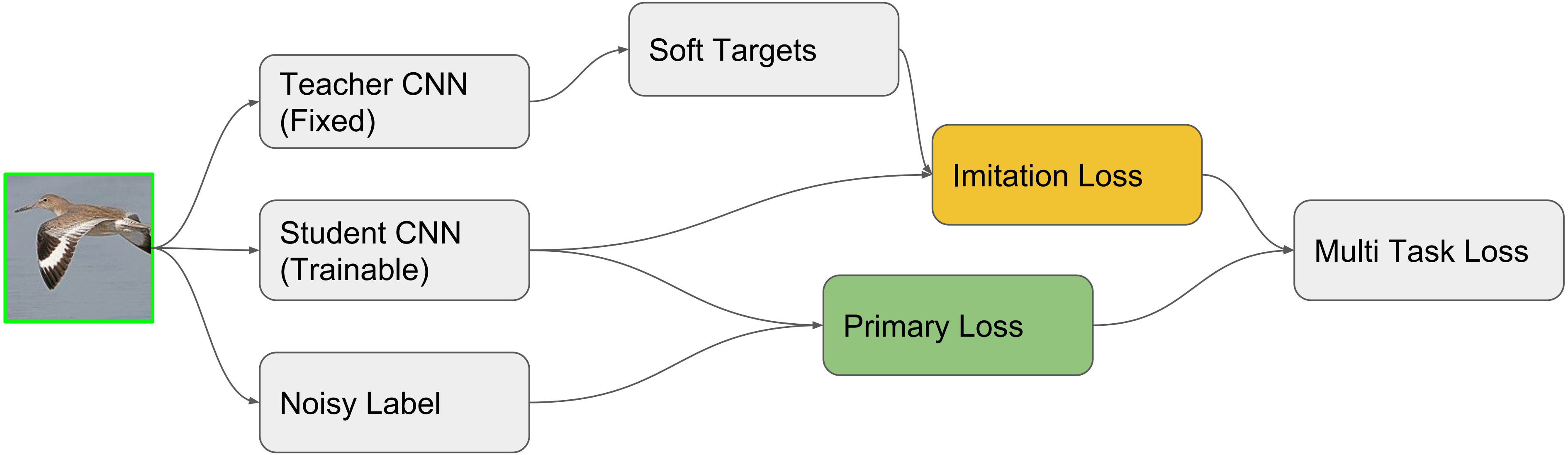
Willet

There is no way to get rid of the ambiguity by itself

Dunlin

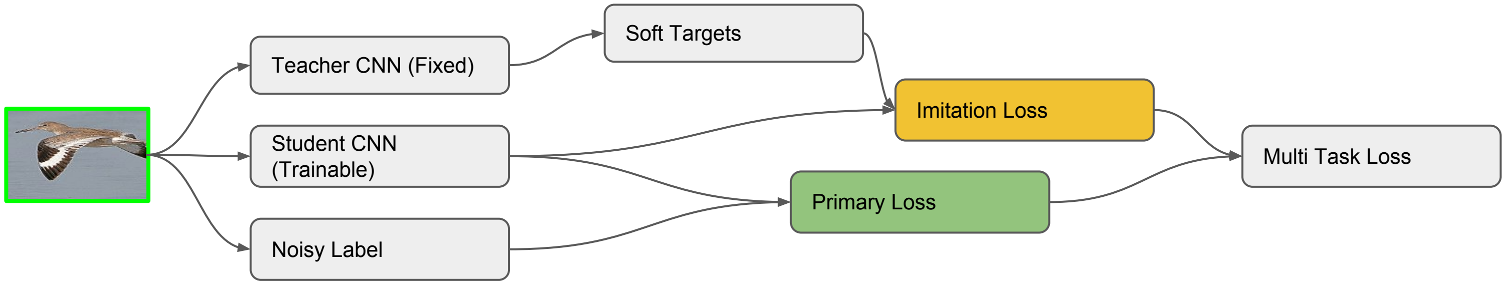
Greylag_goose

Distillation



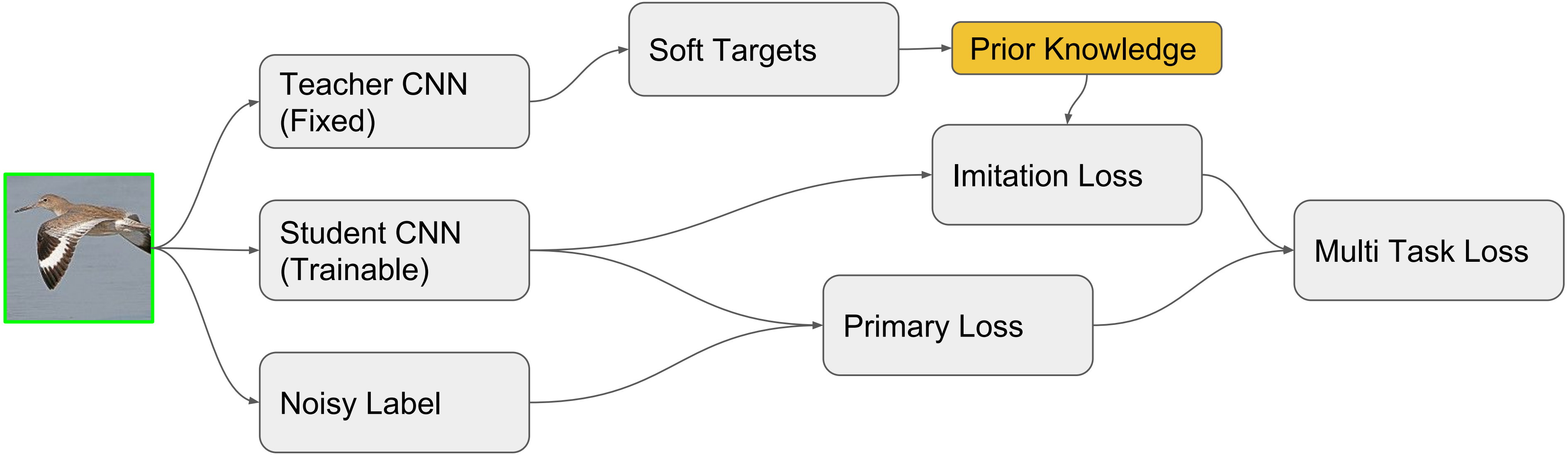
$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \Big[\underbrace{(1-\lambda)\ell(y_i, \sigma(f(x_i)))}_{\text{Primary Loss}} + \underbrace{\lambda \ell(s_i, \sigma(f(x_i)))}_{\text{Imitation Loss}} \Big],$$

Examples of Distillation



Teacher CNN	Student CNN	Reference
Expensive strong CNN ensemble	Deployable weak CNN	Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
Privileged features	Generic features	Lopez-Paz, David, et al. "Unifying distillation and privileged information."*arXiv preprint arXiv:1511.03643* (2015).
Small set of clean labels	Large set of noisy labels + Knowledge graph	Ours

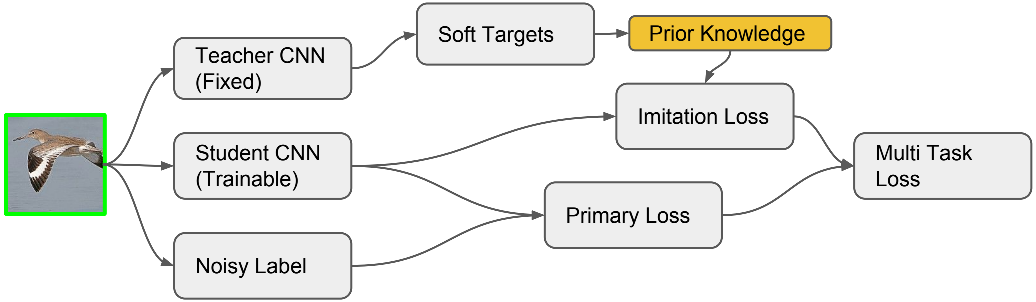
Guided Distillation

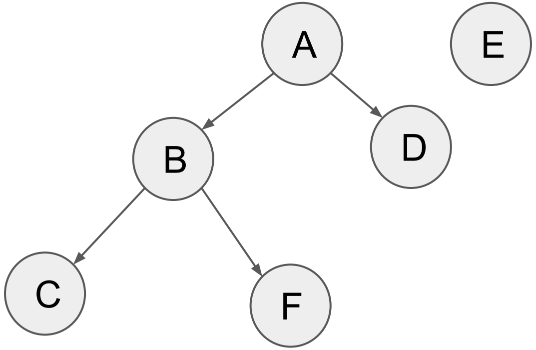


$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \Big[(1-\lambda)\ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))) \Big],$$

$$s = g(\tilde{s}, \Phi)$$

Knowledge Graph Guided Distillation



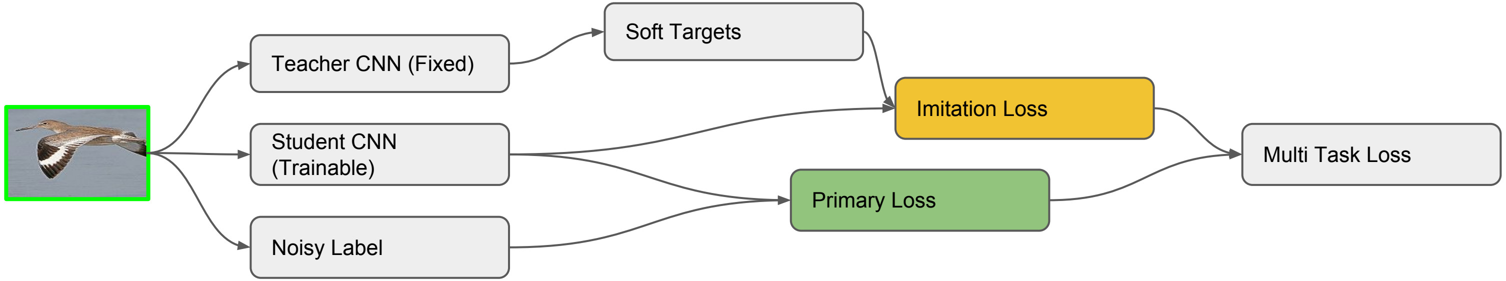


$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \Big[(1-\lambda)\ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))) \Big],$$

$$s \equiv \tilde{s} \times T$$

$$T(m,n) = \begin{cases} 1-\beta & m = n \\ \dfrac{\beta}{|\text{sibling}(n)|} & m \in \text{sibling}(n) \\ 0 & \text{otherwise} \end{cases}$$

Knowledge Distillation == Risk Hedging



$$\hat{y}_i^\lambda = \lambda y_i + (1 - \lambda) s_i$$

Proposition 1. *The optimal risk associated with $\hat{y}^\lambda$ is smaller than both risks with y and s,* i.e.

$$\min_\lambda R_{\hat{y}^\lambda} < \min\{R_y, R_s\}, \tag{7}$$

where y is the unreliable label on $\mathcal{D}$, and s is the soft label output from $f_{\mathcal{D}_c}$. By setting $\lambda = \frac{R_s}{R_s + R_y}$, $R_{\hat{y}^\lambda}$ reaches its minimum,

$$\min_\lambda R_{\hat{y}^\lambda} = \frac{R_y R_s}{R_s + R_y}. \tag{8}$$

Guided Knowledge Distillation: Collaborative Ensembling

Assumption: correlation between classifiers connected on knowledge graph.

$$\mathbf{S} \equiv \hat{\mathbf{S}} \times \mathbf{T}$$

	Ŝ(Willet)	Ŝ(Dunlin)	Ŝ(Wader)	Ŝ(Ruff)	S(Willet, 0.5)
Willet: the name	0.9	0.0	0.0	0.0	0.45
Willet: the bird	0.9	0.6	0.8	0.7	0.80

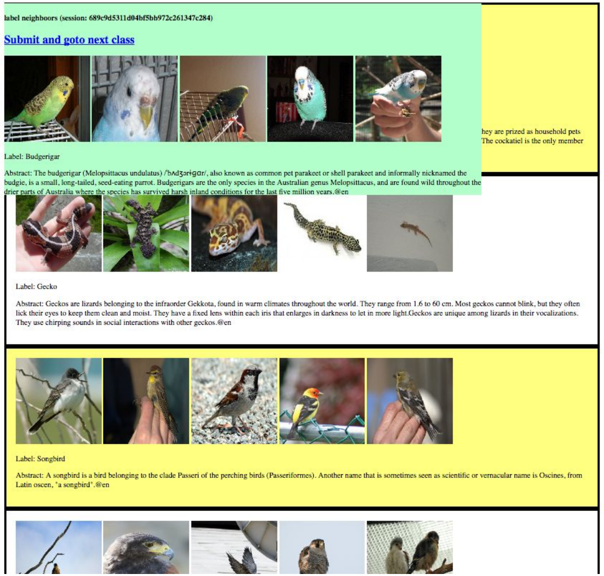
S(Willet, β) = β Willet + (1-β)/3 * (Dunlin + Wader + Ruff)

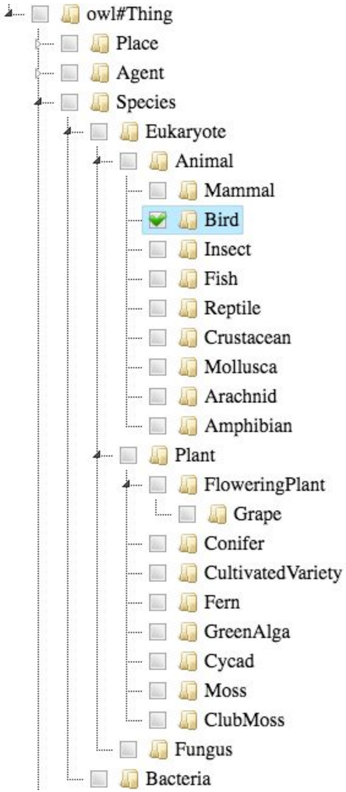
Willet: the name

Willet: the bird

Knowledge Graph

- Semantic knowledge graph
  - Wikipedia/DBPedia/Yago/Probase
- Visual knowledge graph
  - Manual labeling (linear scalable)

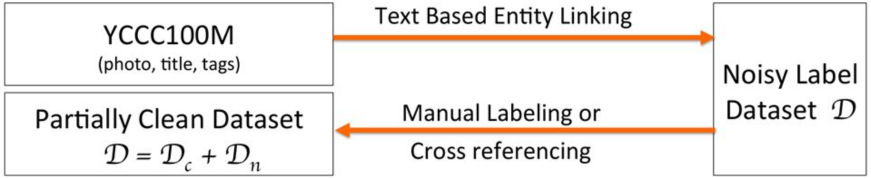




Experiments

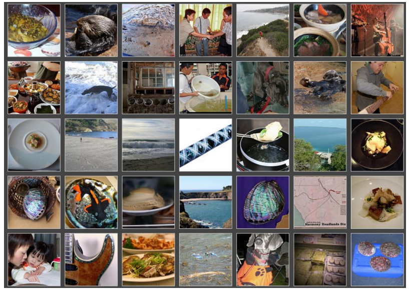
Data collection

- Data collection pipeline



- Difference from WebVision
  - Images are collected in bottom-up fashion
  - Noisy images are kept in test set (background images)
  - mAP, instead of top-K, is used as metric

Abalone



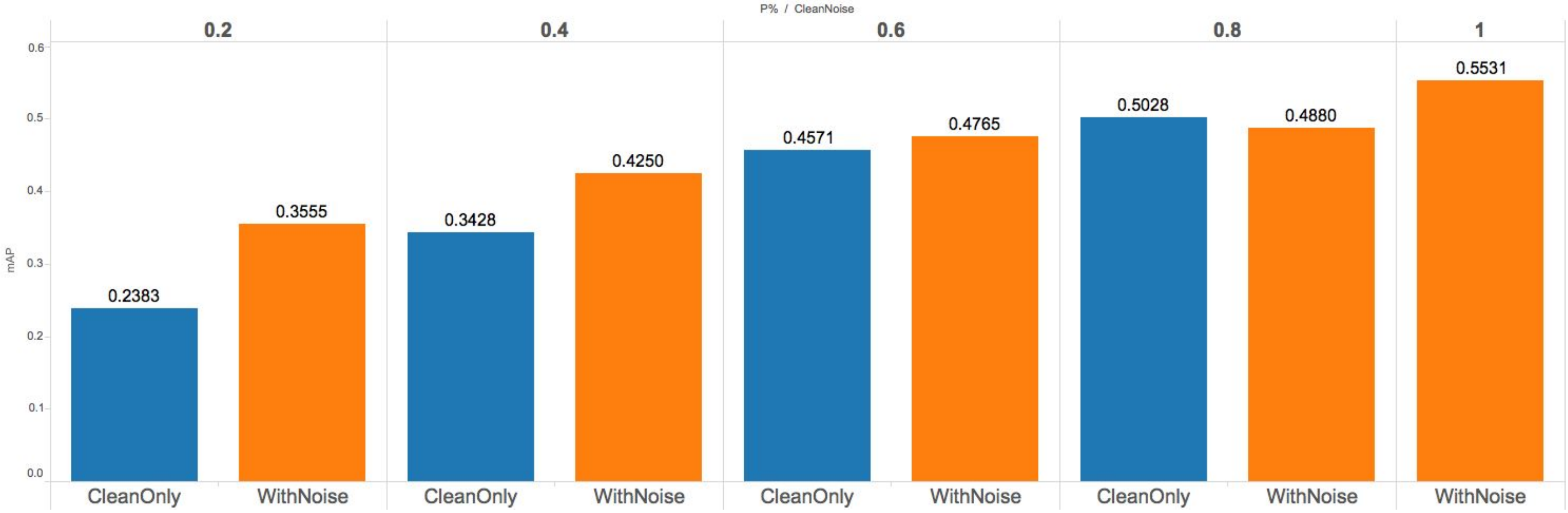
Aphid

Epiphyllum

Datasets Statistics

Name	#Categories	#Train	#Dev	#Test
Sports	238	86K	18K	52K
Species-Y	219	50K	10K	28K
Species-I	219	93K	14K	40K
Artifacts	323	112K	16K	48K

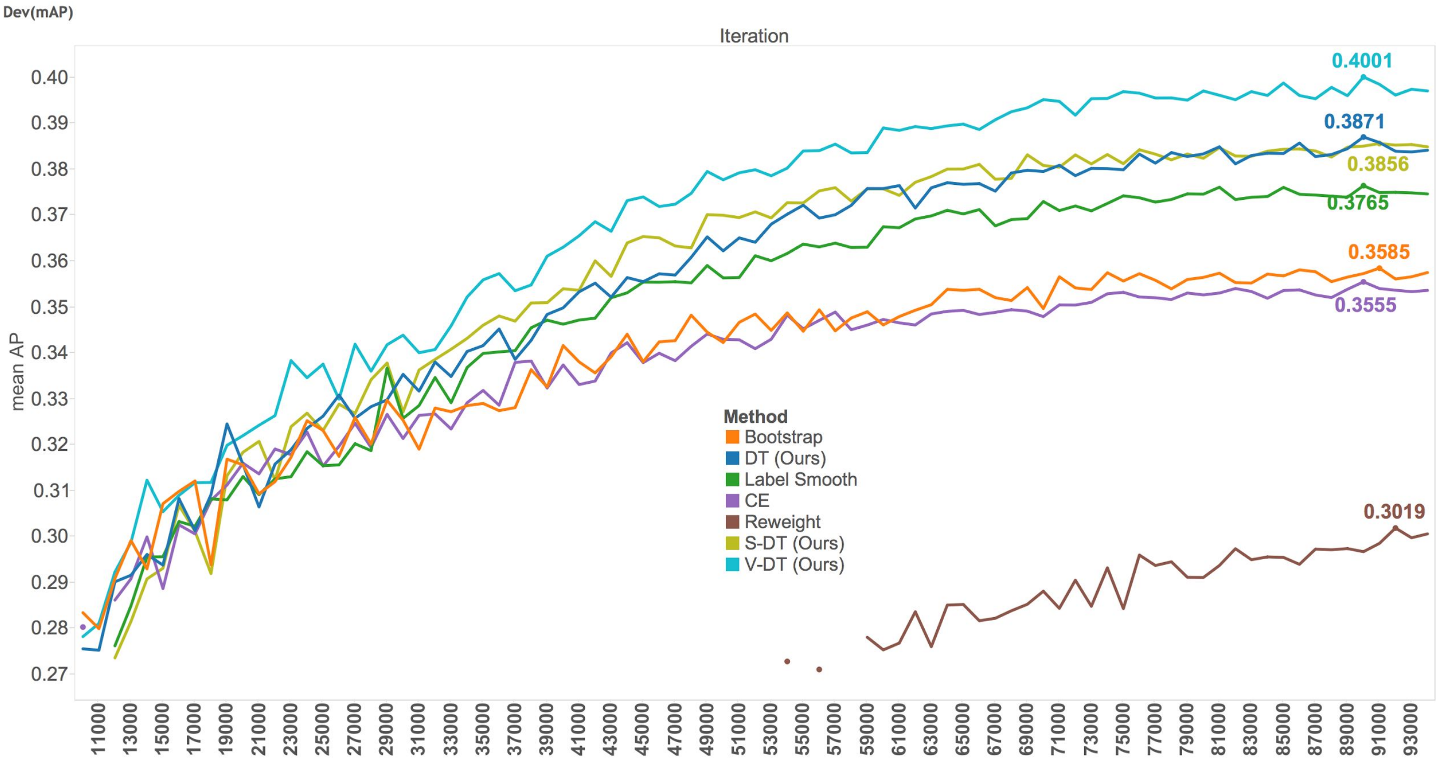
Benefits of the additional noise labels (Cross Entropy)



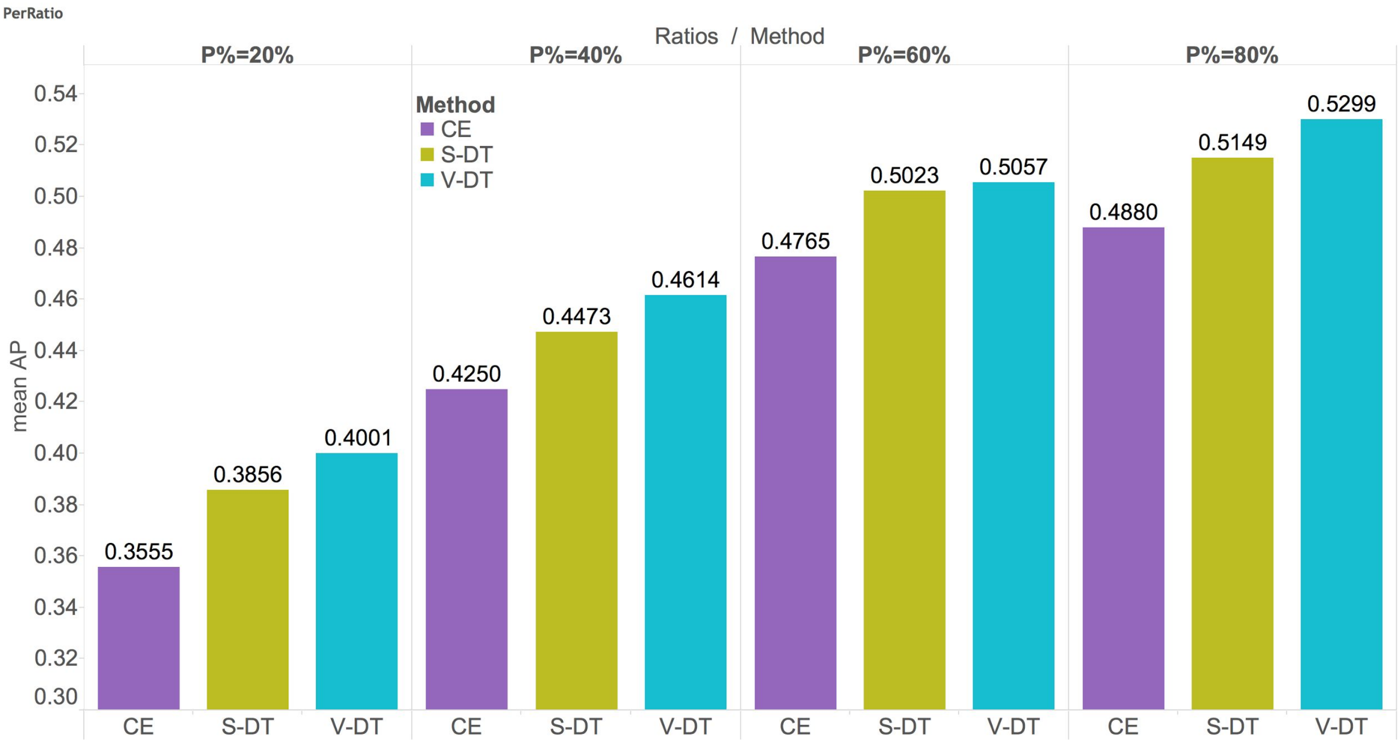
Compared benchmarking methods

- Cross Entropy (CE), Krause, et al. ECCV 2016
- Bootstrap, Reed, et al. ICLR 2014
- Label Smooth, Szegedy, et al. ICLR 2015
- Reweight, Liu, et al. TPAMI 2016
- Ours
  - Distillation (DT)
  - Semantic Knowledge Guided Distillation (S-DT)
  - Visual Knowledge Guided Distillation (V-DT)

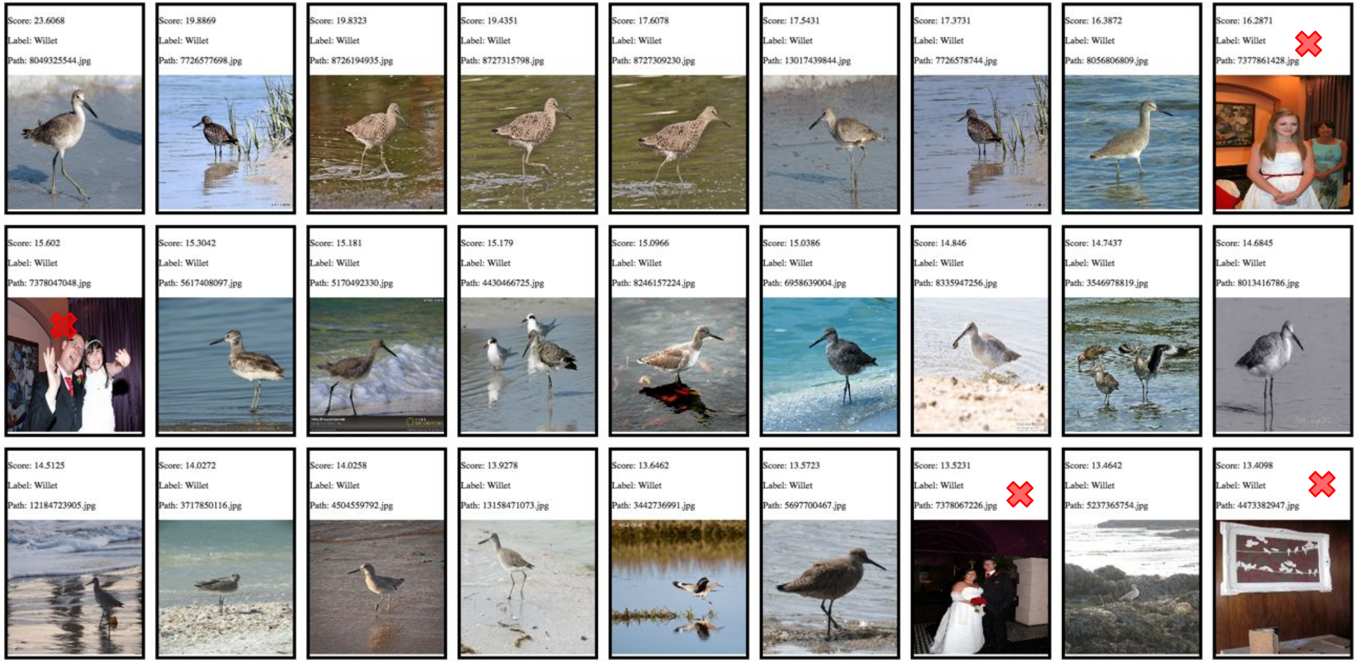
Dev set learning curve (P%=20%)



Results with different ratios of clean data

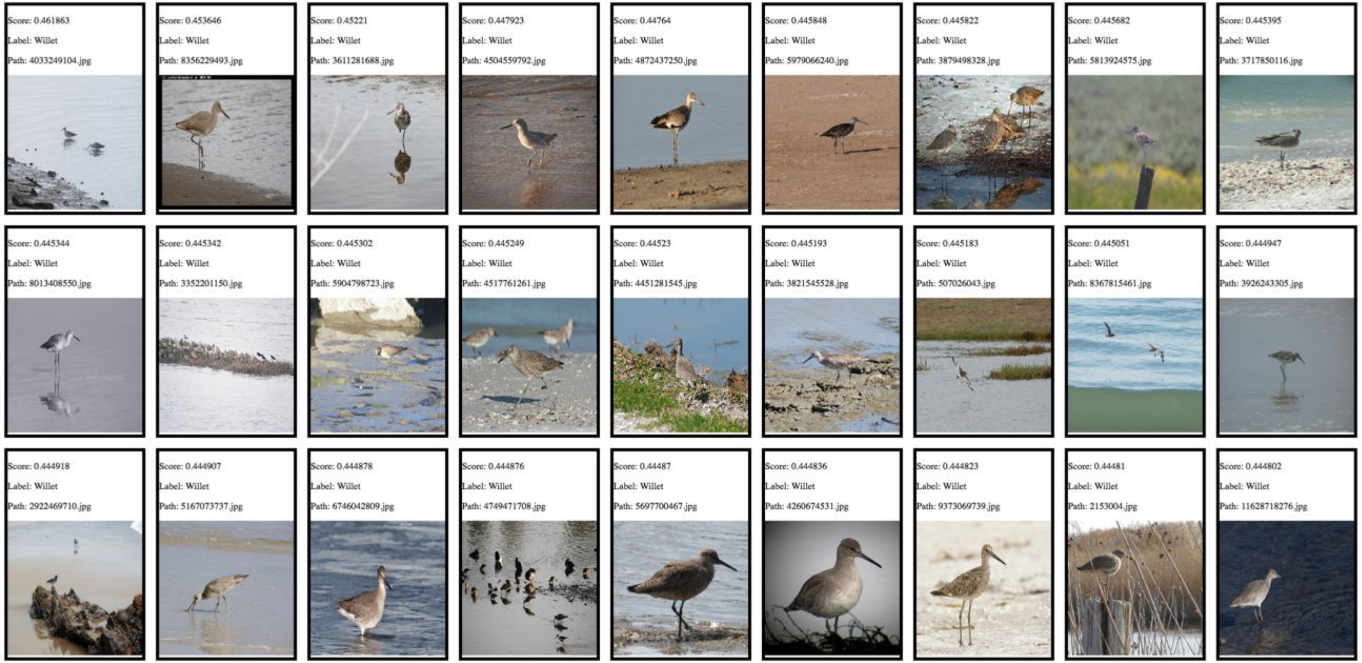


Rank images with baseline model output



The CNN itself is hard to get rid of the "Willet, the name" concept

Rank images with guided distillation



With the aid of the knowledge graph, the concept “Willet, the name” is removed

More results

	Sports	Species-Y	Species-I	Artifacts
Baseline-Clean	44.0	18.1	22.0	19.2
Baseline-Noisy [10]	50.7	23.7	38.5	22.0
Baseline-Ensemble	52.2	25.1	39.1	**26.9**
Bootstrap [16]	50.6	23.6	38.8	23.4
Label Smooth [19]	51.9	25.1	41.4	22.9
Finetune	50.8	22.2	37.5	19.7
Noise Layer [18]	50.8	23.7	38.5	22.0
Importance Re-weighting [12]	50.8	23.7	41.6	24.8
Distillation (Eqn. (4))	53.5	**26.1**	41.6	26.0
Semantic Guided Distillation (Eqn. (13))	**53.7**	25.2	**42.3**	26.0
Upper Bound	54.1	27.4	-	-

Future work

- Learn visual knowledge from data
- Knowledge graph for larger-scale datasets

VISTA Team Submission

Yuncheng Li, Jianchao Yang
Snap Research

Learning rate scheduling

Decay 0.94 at every 2 epochs, stop when validation accuracy converge

- Scratch: base-lr=0.01 ⇒ 0.81 (top-5)
- Finetune: base-lr=0.001 ⇒ 0.85
- Finetune: base-lr=0.0001 ⇒ 0.89
- Finetune: base-lr=0.00001 ⇒ 0.89

“Negative results”

- Bootstrap
- Label smooth
- Subset bootstrap
- Co-training
  - Iterate between text and image classifier training

More details

- Inception-v3
- 4 GPUs
- Tensorflow https://github.com/tensorflow/models/tree/master/slim

Take home message

- Carefully learning rate scheduling is more important
  - “Label noise is fine, just makes the learning slower”
- Conjecture: Using larger model to absorb noise works just fine.