Learning with Limited Labeled Data

Hsuan-Tien Lin
林軒田

Dept. of Computer Science and Information Enginnering,
National Taiwan University
國立臺灣大學資訊工程學系

January 26, 2022
AI & Data Science Workshop

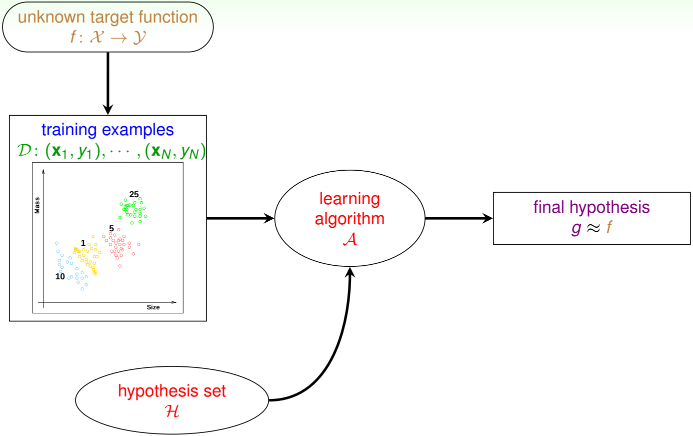
Outline

Learning with Limited Labeled Data

Learning from Label Proportions

Learning from Complementary Labels

Supervised Learning (Slide Modified from My ML Foundations MOOC)

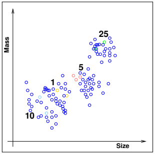


supervised learning: every input vector (picture) $\mathbf{x}_n$ with **its label (category)** y_n
—what if **limited labeled data**?

Semi-Supervised Learning

unknown target function
$f\colon \mathcal{X} \to \mathcal{Y}$

training examples
$\mathcal{D}\colon (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_M, y_M),$
$\mathbf{x}_{N+1}, \ldots, \mathbf{x}_N$



learning algorithm
$\mathcal{A}$

final hypothesis
$g \approx f$

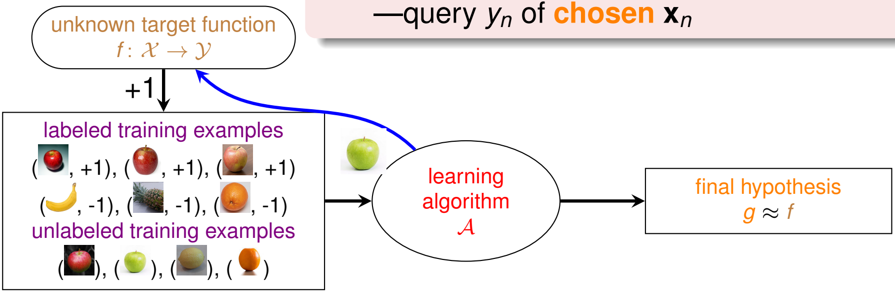
hypothesis set
$\mathcal{H}$

semi-supervised learning:
a few labeled examples
+ many unlabeled examples

Active Learning: Learning by 'Asking'

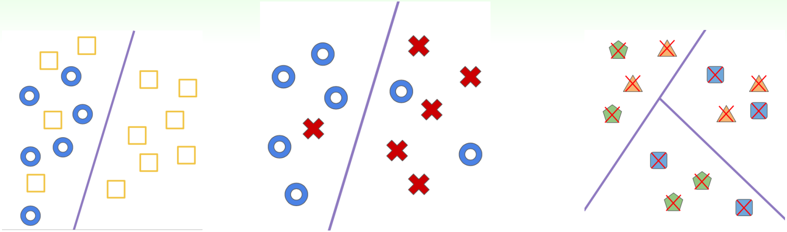
Protocol $\Leftrightarrow$ Learning Philosophy

- batch: 'duck feeding'
- **active: 'question asking'** (iteratively) —query y_n of **chosen** $\mathbf{x}_n$



active learning (on top of semi-supervised):
a few labeled examples + unlabeled pool
+ a few strategically-queried labels

Weakly-Supervised Learning: Learning without True Labels



(a) positive-unlabeled learning (b) learning with noisy labels (c) learning with complementary labels

- positive-unlabeled: some of true $y_n = +1$ revealed
- noisy: (cheaper) noisy label y'_n instead of true y_n
- complementary: 'not label' $\overline{y}_n$ instead of true y_n

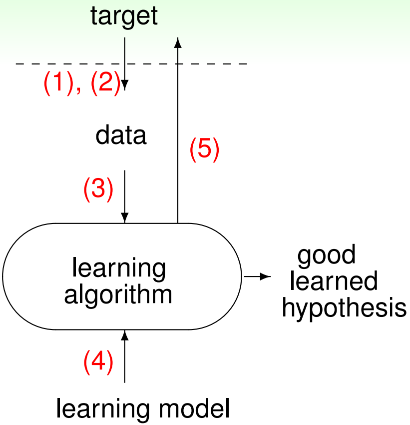
weakly-supervised:
a few (no) labeled examples
+ many 'related' and easier-to-get labels

Our Ongoing Research Quests

Learning from Limited Labeled Data (L^{3}D)

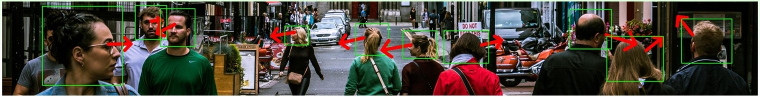
- in **supervised learning**
  - e.g. **uneven-margin augmentation** for imbalanced learning?
- in **interactive learning**
  - e.g. can **strategically** obtained labels push L^{3}D to the extreme?
- in **generative learning**
  - e.g. **development with cloned data** first, validate with limited labeled data later?
- in **weakly-supervised learning**
  - e.g. **sketch with weak labels** first, refine with limited labeled data later—or maybe **learn from many weak labels** only?

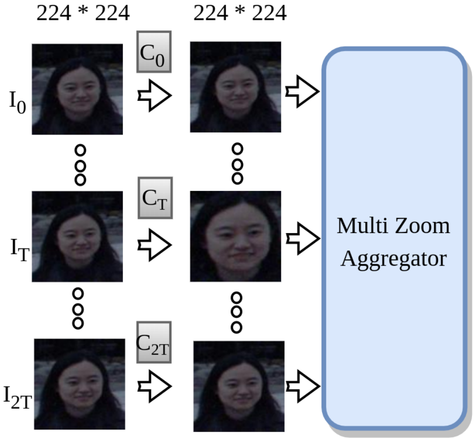
Some of Our Selected Work



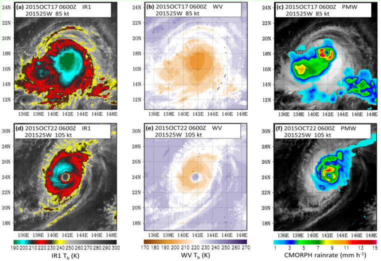
1. zero-shot learning (ICLR 2021): **no labeled data** but only descriptions for new classes
2. learning from complementary labels (ICML 2020): **cheaper weakly labeled data**
3. robust estimation (gaze: BMVC 2020, typhoon: KDD 2018): **domain-driven data augmentation**
4. robust generation (NeurIPS 2021): **math-driven objective augmentation**
5. active learning (EMNLP 2020): **a few actively labeled data**

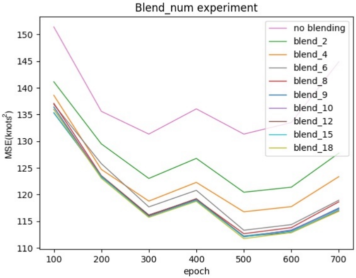
Quick Stories about Augmentation (1/3) (Ashesh, 2021)



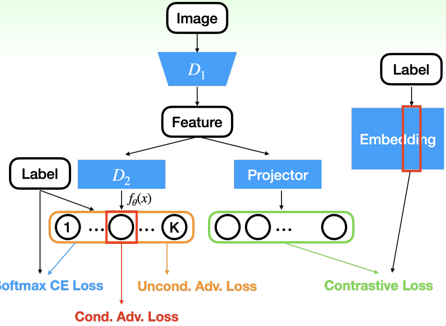


Quick Stories about Augmentation (2/3) (Chen, 2018)





Quick Stories about Augmentation (3/3) (Chen, 2021)



$$\begin{aligned}\log p(x,y) &= \overbrace{\boxed{\log p(x \mid y)}}^{\text{Cond. Distribution}} + \log p(y) \\ &= \underbrace{\boxed{\log p(y \mid x)}}_{\text{Classifier}} + \underbrace{\boxed{\log p(x)}}_{\text{Uncond. Distribution}}\end{aligned}$$

Outline

Learning with Limited Labeled Data

Learning from Label Proportions

Learning from Complementary Labels

Learning from Label Proportions

Training

bag	[a, o, s, k]
{ , , ,  }	$[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0]$
{ , , ,  }	$[\frac{1}{4}, \frac{1}{2}, 0, \frac{1}{4}]$

Test

?

motivations

- expensive labeling
- privacy issues

LLP: learn an instance-level classifier with
proportion labels

LLP Setting

input

Given M bags $B_1, \ldots, B_M$, where the m-th bag contains a set of instances $\mathcal{X}_m$ and a proportion label $\mathbf{p}_m$, defined by

$$\mathbf{p}_m = \frac{1}{|\mathcal{X}_m|} \sum_{n:\ \mathbf{x}_n \in \mathcal{X}_m} \mathbf{e}^{(y_n)}, \quad \bigcup_{m=1}^{M} \mathcal{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

output

learn a usual instance classifier $g_\theta : \mathbb{R}^D \to$ estimated probability

Our Sol.: LLP w/ Consistency Regularization (Tsai, 2020)

vanilla: bag-level proportion loss

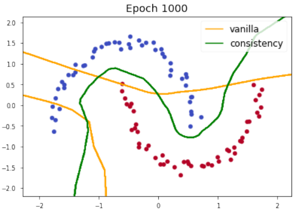
$$L_{prop} = KL(\mathbf{p} \| \hat{\mathbf{p}})$$

- 'distance' between target $\mathbf{p}$ and estimated $\hat{\mathbf{p}} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} g_\theta(\mathbf{x})$ small
- extension of **standard cross-entropy loss**

instance-level regularization

$$L_{cons} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} KL(g_\theta(\mathbf{x}) \| g_\theta(\hat{\mathbf{x}}))$$

- 'difference' between $\mathbf{x}$ and perturbed $\hat{\mathbf{x}}$ small
- mature technique for **semi-supervised learning**



LLP with consistency regularization:

$$L = L_{prop} + \alpha L_{cons}$$

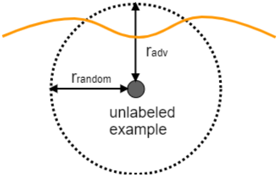
Consistency Loss by Virtual Adversarial Training

smoothness assumption

if $\mathbf{x}_i \approx \mathbf{x}_j$, then $y_i \approx y_j$

goal

encourage the classifier to produce consistent outputs for neighbors



Virtual Adversarial Training (Miyato, 2018)

generate a perturbed example $\hat{\mathbf{x}}$ that most likely causes the model to misclassify

$$\hat{\mathbf{x}} = \underset{\|\hat{\mathbf{x}}-\mathbf{x}\| \le r}{\operatorname{argmax}} \; KL(g_\theta(\mathbf{x}) \| g_\theta(\hat{\mathbf{x}}))$$

consistency loss w/ VAT:

$$L_{cons}(\theta) = KL(g_\theta(\mathbf{x}) \| g_\theta(\hat{\mathbf{x}}))$$

Experimental Results

		Bag Size				
Dataset	Method	16	32	64	128	256
SVHN	vanilla	95.28	95.20	94.41	88.93	12.64
	LLP-VAT	95.66	95.73	94.60	91.24	11.18
CIFAR10	vanilla	88.77	85.02	70.68	47.48	38.69
	LLP-VAT	89.30	85.41	72.49	50.78	41.62
CIFAR100	vanilla	58.58	48.09	20.66	5.82	2.82
	LLP-VAT	59.47	48.98	22.84	9.40	3.29

consistency regularization (VAT) helps!

Take-Home Message

- LLP: a typical **weakly-supervised** learning problem
- **consistency regularization** helps
  —can other regularization help?
- anyone using?
  - 50% **accuracy on 10 class for big bags**?!
  - **no real-world data yet**

Outline

Learning with Limited Labeled Data

Learning from Label Proportions

Learning from Complementary Labels

Fruit Labeling Task (Image from AICup in 2020)

→

hard: true label

- orange **?**
- mango **?**
- cherry
- banana

easy: complementary label

- orange
- mango
- cherry
- banana ✗

complementary: **less labeling cost/expertise** required

Comparison

Ordinary (Supervised) Learning

training: $\{(\mathbf{x}_n =$  $, y_n = \text{mango})\} \to$ **classifier**

Complementary Learning

training: $\{(\mathbf{x}_n =$  $, \bar{y}_n = \text{banana})\} \to$ **classifier**

testing goal: **classifier**() $\to$ cherry

ordinary versus complementary:
same goal via **different training data**

Learning with Complementary Labels Setup

Given

N examples (input $\mathbf{x}_n$, complementary label $\overline{y}_n$) $\in \mathcal{X} \times \{1, 2, \cdots K\}$ in data set $\mathcal{D}$ such that $\overline{y}_n \neq y_n$ for some hidden ordinary label $y_n \in \{1, 2, \cdots K\}$.

Goal

a multi-class classifier $g(\mathbf{x})$ that **closely predicts** (0/1 error) the ordinary label y associated with some **unseen** inputs x

LCL model design: connecting
complementary & ordinary

Unbiased Risk Estimation for LCL

Ordinary Learning

- empirical risk minimization (ERM) on training data

risk: $\mathbb{E}_{(\mathbf{x},y)}[\ell(y, g(\mathbf{x}))]$ **empirical risk:** $\mathbb{E}_{(\mathbf{x}_n,y_n)\in\mathcal{D}}[\ell(y_n, g(\mathbf{x}_n))]$

- loss ℓ: usually **surrogate** of 0/1 error

LCL (Ishida, 2019)

- rewrite the loss ℓ to $\bar{\ell}$, such that

unbiased risk estimator: $\mathbb{E}_{(\mathbf{x},\bar{y})}[\bar{\ell}(\bar{y}, g(\mathbf{x}))] = \mathbb{E}_{(\mathbf{x},y)}[\ell(y, g(\mathbf{x}))]$

under assumptions (e.g. uniform complementary labels)
- LCL by minimizing **URE**

URE: **pioneer models** for LCL

URE Overfits Easily

$$
\begin{aligned}
\ell &= -\log(\boldsymbol{p}(y \mid \mathbf{x})) \\
\bar{\ell} &= (K-1)\log(\boldsymbol{p}(\bar{y} \mid \mathbf{x})) - \sum_{k=1}^{K} \log(\boldsymbol{p}(k \mid \mathbf{x}))
\end{aligned}
$$

ordinary risk and URE very different

- $\ell > 0 \rightarrow$ ordinary risk non-negative
- small $\boldsymbol{p}(\bar{y} \mid \mathbf{x})$ (often) $\rightarrow$ possibly very negative $\bar{\ell}$
  empirical URE can be negative on **some observed** $\bar{y}$
- negative empirical URE **drags minimization** towards overfitting

how can we avoid negative empirical URE?

Proposed Framework (Chou, 2021)

Minimize Complementary 0/1

- our goal: minimize 0/1 loss instead of ℓ
- unbiased estimator of R_{01} is **simple**

$$\boldsymbol{\overline{R}_{\overline{01}}}\text{ :}\quad \mathbb{E}_{\overline{y}}[\overline{\ell}_{01}(\overline{y}, g(\mathbf{x}))] = \ell_{01}(y, g(\mathbf{x}))$$

- $\overline{\ell}_{01}$ as the complementary 0/1 loss:

$$\overline{\ell}_{01}(\overline{y}, g(\mathbf{x})) = [\![\overline{y} = g(\mathbf{x})]\!]$$

Surrogate Complementary Loss (SCL):
surrogate **after** complementary 0/1

Illustrative Difference between URE and SCL

$$
\begin{tikzcd}
\text{URE} & & \boxed{R_\ell} \arrow[r, "E"] & \boxed{\overline{R}_{\overline{\ell}}} \arrow[r] & \boxed{\hat{\overline{R}}_{\overline{\ell}}} \\
& \boxed{R_{01}} \arrow[ur, "A"] \arrow[dr, "E"'] & & & \\
\text{SCL} & & \boxed{\overline{R}_{\overline{01}}} \arrow[r, "A"'] & \boxed{\overline{R}_{\phi}} \arrow[r] & \boxed{\hat{\overline{R}}_{\phi}}
\end{tikzcd}
$$

URE: Ripple effect of errors

- Theoretical motivation (Ishida, 2017)
- Estimation step (E) amplifies approximation error (A) in $\overline{\ell}$

SCL: 'Directly' minimize complementary likelihood

- Non-negative loss ϕ
- Practically prevents ripple effect

Negative Risk Avoided

Unbiased Risk Estimator (URE)

URE loss $\overline{\ell}_{CE}$ from cross-entropy ℓ_{CE},

$$\overline{\ell}_{CE}(\overline{y}, g(\mathbf{x})) = \underbrace{(K-1)\log(\boldsymbol{p}(\overline{y} \mid \mathbf{x}))}_{\text{negative loss term}} - \sum_{j=1}^{K} \log(\boldsymbol{p}(j \mid \mathbf{x}))$$

can go negative.

Surrogate Complementary Loss (SCL)

a surrogate of $\overline{\ell}_{01}$ (Kim, 2019)

$$\phi_{\mathrm{NL}}(\overline{y}, g(\mathbf{x})) = -\log(1 - \boldsymbol{p}(\overline{y} \mid \mathbf{x}))$$

remains non-negative.

Classification Accuracy

Methods

1. Unbiased risk estimator (URE) (Ishida, 2019)
2. Surrogate complementary loss (SCL)

Table: URE and NN are based on $\bar{\ell}$ rewritten from cross-entropy loss, while SCL is based on exponential loss $\phi_{\text{EXP}}(\bar{y}, g(\mathbf{x})) = \exp(\boldsymbol{p}_{\bar{y}})$.

Data set + Model	URE	SCL
MNIST + Linear	0.850	**0.902**
MNIST + MLP	0.801	**0.925**
CIFAR10 + ResNet	0.109	**0.492**
CIFAR10 + DenseNet	0.291	**0.544**

Gradient Analysis

Gradient Direction of URE

- Very diverged directions on each $\overline{y}$ to maintain unbiasedness
- Low correlation to the target ℓ_{01}

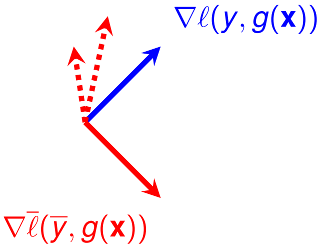


Figure: Illustration of URE

Gradient Direction of SCL

- Targets to minimum likelihood objective
- High correlation to the target $\overline{\ell}_{01}$

Gradient Estimation Error

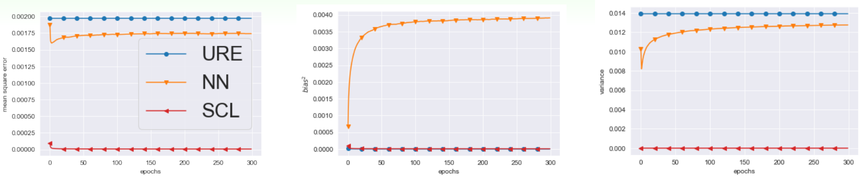
Bias-Variance Decomposition

$$\begin{aligned}\text{MSE} &= \mathbb{E}\left[(\boldsymbol{f}-\boldsymbol{c})^2\right] \\ &= \underbrace{\mathbb{E}\left[(\boldsymbol{f}-\boldsymbol{h})^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\boldsymbol{h}-\boldsymbol{c})^2\right]}_{\text{Variance}}\end{aligned}$$

Gradient Estimation

1. Ordinary gradient $\boldsymbol{f} = \nabla \ell(y, g(\mathbf{x}))$
2. Complementary gradient $\boldsymbol{c} = \nabla \bar{\ell}(\bar{y}, g(\mathbf{x}))$
3. Expected complementary gradient $\boldsymbol{h}$

Bias-Variance Tradeoff



(a) MSE (b) Bias^2 (c) Variance

Findings

- SCL reduces variance by introducing small bias (towards $\overline{y}$)

	Bias	Variance	MSE
URE	0	Big	Big
SCL	Small	Small	Small

Take-Home Message

- LCL: another popular **weakly-supervised** learning problem
- **surrogate on complementary** helps
  - avoid negative loss
  - lower gradient variance (with trade-off in bias)
- anyone using?
  - **uniform complementary generation unrealistic** (ongoing)
  - **need stronger theoretical guarantee** (ongoing)

Thank you! Questions?