Adversarial Training with Complementary Labels: On the Benefit of Gradually Informative Attacks

Jianan Zhou*, Jianing Zhu*, Jingfeng Zhang, Tongliang Liu, Gang Niu, Bo Han, Masashi Sugiyama

TL;DR: Is it possible to equip machine learning models with adversarial robustness when all the labels given for training are wrong (i.e., complementary labels)?
Affirmative! In this paper, we conduct adversarial training under a promising setting of weakly supervised learning.

Adversarial training with imperfect supervision is significant but receives limited attention.

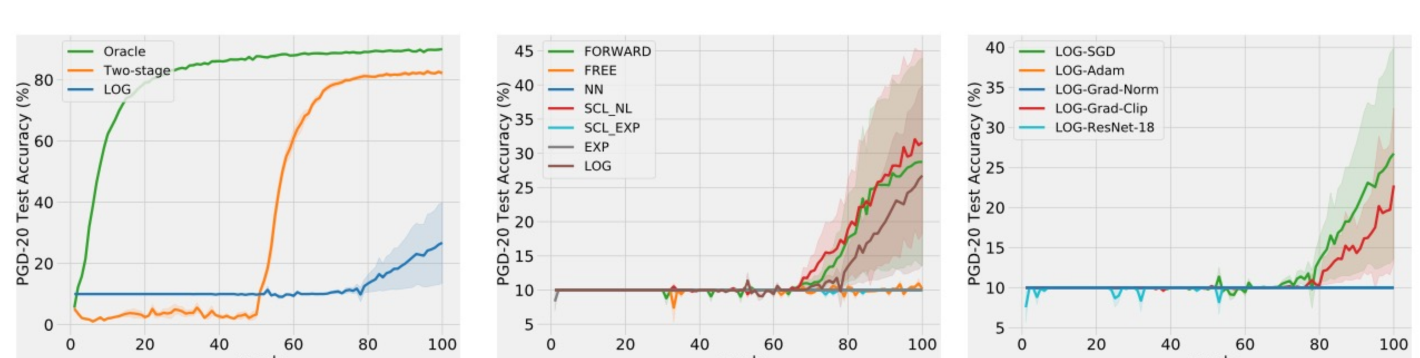
Motivation: Deep neural networks (DNNs) are vulnerable to adversarial examples. Adversarial Training (AT) is one of the most effective methods to equip DNNs with adversarial robustness against crafted perturbations. In this paper, we consider a brand new yet challenging setting (i.e., AT with Complementary Labels – a wrong class) motivated by:

- Most of the previous work focused on AT with perfect supervision, while imperfect supervision is much more common in real-life scenarios.
- Learning with complementary labels (CLs) is a promising setting of weakly supervised learning. It illustrates the possibility of training an ordinary classifier even when all the labels given for training are wrong.
- We believe studying AT with CLs could benefit both communities.

An illustration of complementary labels

Empirical Observations

The straightforward replacement of the ordinary loss with a complementary loss (either an unbiased or biased risk estimator of the ordinary risk) in the min-max formulation of AT results in consistent experimental failure:



(a) Failure of direct combination (b) Trials of Complementary Losses (c) Trials of Training Tricks

From theoretical and empirical perspectives, when using complementary losses as the objectives of adversarial optimization, we identify the underlying challenges as:

a) Intractable adversarial optimization with CLs;
b) Low-quality constructed adversarial examples.

Preliminaries

In complementary learning, through the general backward correction, an unbiased risk estimator (URE) is derived as follows:

Proposition 1. *The ordinary risk can be expressed in terms of CLs as follows,*

$$R(g;\ell) = E_{(x,y)\sim\mathcal{D}}[\ell(y,g(x))] = E_{(x,\bar{y})\sim\bar{\mathcal{D}}}[\bar{\ell}(\bar{y},g(x))] = \bar{R}(g;\bar{\ell}),$$

when $\bar{\ell}$ is rewritten as

$$\bar{\ell}(\bar{y},g(x)) = e_{\bar{y}}^{T}(Q^{-1})\ell(g(x)),$$

With the uniform assumption, $\bar{\ell}$ can be further rewritten as

$$\bar{\ell}(\bar{y},g(x)) = -(K-1)\ell(\bar{y},g(x)) + \sum_{j=1}^{K}\ell(j,g(x)).$$

With this expression, we can obtain an URE of the ordinary risk only from CLs.

Notations:
y – ordinary label
$\bar{y}$ – complementary label
ℓ – ordinary loss
$\bar{\ell}$ – complementary loss
Q – transition matrix
K – number of classes
$\ell(g(x)) = [l(1,g(x)),\dots l(K,g(x))]$

Theoretical Analysis

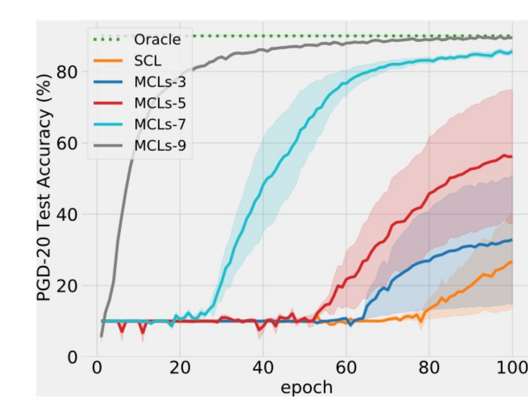
Theoretically, the adversarial optimization of the complementary risk with the general backward corrected loss is statistically consistent with that of the ordinary risk with the ordinary loss. However, the inconsistency between their empirical risks exists due to the unavailability of enough CLs in practice (e.g., only one CL for each data sample).

Proposition 2. *For the general backward correction, conducting AT on the complementary risk is equivalent to that on the ordinary risk. However, this is not the case on their empirical risks:*

$$\min_{\theta} E_{x\sim p(X)} \max_{\tilde{x}\in\mathcal{B}_{\epsilon}[x]} E_{y\sim p(Y|X=x)}[\ell(y,g(\tilde{x}))] = \min_{\theta} E_{x\sim p(X)} \max_{\tilde{x}\in\mathcal{B}_{\epsilon}[x]} E_{\bar{y}\sim p(\bar{Y}|X=x)}[\bar{\ell}(\bar{y},g(\tilde{x}))],$$

$$\min_{\theta}\frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{x}_i\in\mathcal{B}_{\epsilon}[x_i]}[\ell(y_i,g(\tilde{x}_i))] \neq \min_{\theta}\frac{1}{n}\sum_{i=1}^{n}\max_{\tilde{x}_i\in\mathcal{B}_{\epsilon}[x_i]}[\bar{\ell}(\bar{y}_i,g(\tilde{x}_i))].$$

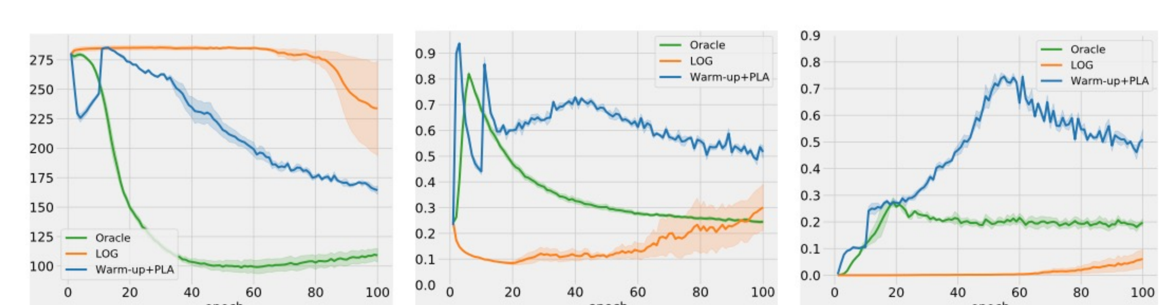
In summary, the solution to inner maximization of the complementary risk is equivalent to that of the ordinary risk if and only if maximizing the weighted loss over all candidate CLs (i.e., $E_{\bar{y}\sim p(\bar{Y}|X=x)}[\bar{\ell}(\bar{y},g(\tilde{x}))]$). Moreover, from the perspective of AT, it is hard to generate high-quality $\tilde{x}$ if only maximizing $\bar{\ell}$, since it can't guarantee the minimization of $p_{\theta}(y|\tilde{x})$ (based on $\bar{\ell}$'s formulation).



Empirical justification for Proposition 2

Empirical Analysis

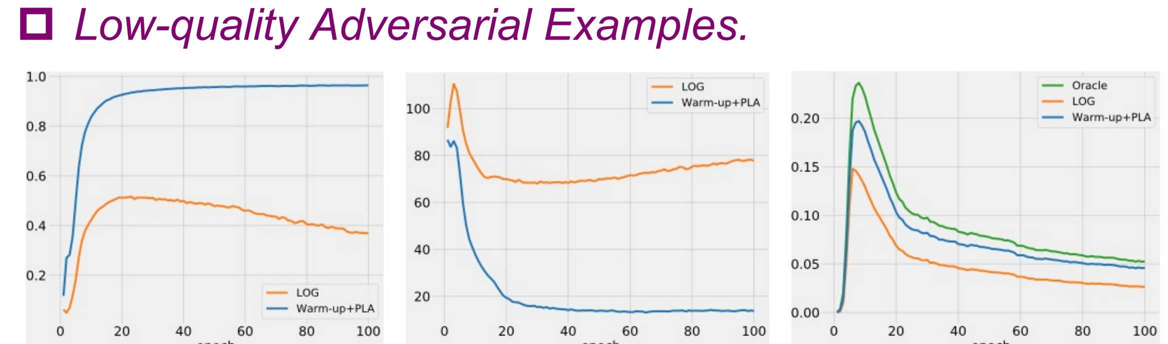
- *Intractable Adversarial Optimization.*



(a) Gradient Direction Variance (b) (Last Layer) Gradient l_2-Norm (c) (First Layer) Gradient l_2-Norm

Empirically, when directly using $\bar{\ell}$ (e.g., LOG) as the objective of AT, we observe 1) a huge variance in gradient directions; 2) gradient vanishing occurs at the early stage of optimization.

- *Low-quality Adversarial Examples.*



(d) Cosine Similarity (e) l_1-Distance (f) $p_{\theta}(y|x) - p_{\theta}(y|\tilde{x})$

Empirically, when directly using $\bar{\ell}$ (e.g., LOG) as the objective of AT, we observe it fails to generate high-quality $\tilde{x}$ (e.g., $p_{\theta}(y|x) - p_{\theta}(y|\tilde{x})$ is low).

Note that for Figure (a)-(c), we adversarially train the model using several loss functions separately. While for Figure (d)-(f), we generate $\tilde{x}$ using various loss functions, with the same optimization model (i.e., the oracle, which is trained using AT with ordinary labels).

Methodology

We propose a unified framework accordingly to deal with the challenges. Specifically, it uses the strategy of gradually information attacks, including:

- *Warm-up Attack.*

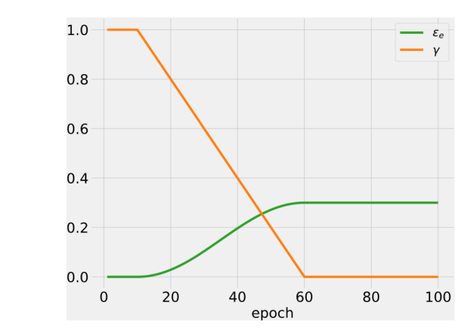
To mitigate the difficulty of adversarial optimization with CLs, we propose to gradually increase the adversarial budget (e.g., the radius of epsilon ball ϵ).

- *Pseudo-Label Attack.*

To improve the quality of constructed adversarial examples $\tilde{x}$, we propose to incorporate the progressively informative model predictions into a corrected complementary loss: $\bar{\ell}(x,\bar{y};\theta) = -(K-1)\log(\gamma\sum_{j\neq\bar{y}}p_{\theta}(j|x) + (1-\gamma)p_{\theta}(\hat{y}|x)),\ \hat{y} = \arg\max_{j\neq\bar{y}} p_c(j|x)$

Scheduler:

$$\mathbb{T}(a,e,E_s) = \begin{cases} \frac{a}{2}\cdot(1-\cos(\min(\frac{e}{E_s},1)\cdot\pi)) & \text{Warm-up Attack,} \\ a\cdot(1-\min(\frac{e}{E_s},1)) & \text{Pseudo-Label Attack,} \end{cases}$$



Algorithm 1 AT with CLs Using Gradually Informative Attacks

Input: number of epochs: E, number of batches: B, number of class: K, radius of epsilon ball: ϵ, number of attack steps: k, step size: α, hyperparameter for exponential moving average of cashed model predictions: β; hyperparameter for label weights: γ;
Output: robust model θ_R;

```
p_c ← 1/(K−1) (1 − e_ȳ)^T for all data
for e = 1, ..., E do
    β, γ ← Update hyperparameters
    ε_e, α_e, k_e ← Adversarial budget scheduler                      ▷ Warm-up Attack
    for b = 1, ..., B do
        x, ȳ ← Sample a batch of data
        p_θ(x) ← Softmax(θ(x)), p_c(x) ← βp_c(x) + (1 − β)p_θ(x), p_c(ȳ|x) ← 0
        ŷ ← arg max_{j≠ȳ} p_c(j|x)
        x̃ ← Approximate the solutions of max ℓ̄(x, ȳ; θ) (Eq. 10) by PGD(ε_e, α_e, k_e)
        L = −(K − 1) log(γ Σ_{j≠ȳ} p_θ(j|x̃) + (1 − γ)p_θ(ŷ|x̃))          ▷ Pseudo-Label Attack
        θ ← SGD(θ, ∇_θ L)
    end for
    Evaluate(ε, α, k)
end for
```

Experiments

Setting:

- For complementary learning, we mainly consider the setting of the single complementary label (SCL) with the uniform assumption, where labels other than the true label are chosen as a CL with the same probability, and each data sample only has one CL.
- For adversarial training, we mainly consider the standard adversarial training, which is formulated as a min-max optimization problem.

Table 1: Means (standard deviations) of natural and adversarial test accuracy.

Dataset	Method	Natural	PGD	CW	AA
MNIST	Oracle	99.46(±0.04)	98.14(±0.04)	97.45(±0.08)	92.53(±0.23)
	Two-stage	99.07(±0.02)	97.44(±0.23)	96.72(±0.21)	92.06(±0.45)
	FORWARD [39]	97.22(±1.13)	93.76(±2.38)	92.13(±2.87)	85.41(±3.69)
	FREE [18]	48.94(±26.04)	38.02(±22.21)	32.81(±19.63)	22.68(±13.91)
	NN [18]	68.48(±40.40)	66.80(±39.22)	66.25(±38.84)	60.65(±38.04)
	SCL_NL [7]	93.09(±7.35)	87.07(±12.40)	84.59(±14.51)	75.98(±18.29)
	SCL_EXP [7]	14.88(±4.99)	14.34(±4.23)	13.58(±3.16)	10.47(±1.25)
	EXP [13]	10.99(±0.50)	10.99(±0.50)	10.99(±0.50)	10.99(±0.50)
	LOG [13]	97.16(±0.64)	93.38(±1.25)	91.67(±1.39)	84.88(±2.09)
	Warm-up+PLA	**99.22(±0.02)**	**97.73(±0.06)**	**97.11(±0.07)**	**92.37(±0.19)**
Kuzushiji	Oracle	95.94(±0.15)	90.01(±0.43)	88.06(±0.96)	70.63(±0.48)
	Two-stage	89.75(±0.42)	82.91(±1.01)	80.21(±1.27)	64.57(±1.79)
	FORWARD	35.48(±27.96)	29.84(±25.55)	28.09(±24.37)	22.01(±18.98)
	FREE	16.17(±1.77)	12.01(±0.55)	9.33(±1.50)	4.08(±1.60)
	NN	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	8.87(±1.60)
	SCL_NL	40.83(±24.03)	32.82(±22.88)	29.93(±22.53)	20.86(±19.74)
	SCL_EXP	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	8.21(±2.54)
	EXP	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)
	LOG	32.66(±25.50)	26.87(±22.66)	24.90(±21.20)	18.78(±16.95)
	Warm-up+PLA	**91.60(±0.49)**	**85.88(±0.48)**	**83.74(±0.35)**	**68.75(±0.68)**
CIFAR10	Oracle	78.10(±0.36)	47.35(±0.05)	45.66(±0.22)	43.47(±0.23)
	Two-stage	64.98(±2.70)	42.48(±0.92)	39.90(±0.67)	38.89(±0.46)
	FORWARD	15.41(±0.85)	13.58(±0.56)	13.03(±0.34)	12.94(±0.35)
	FREE	11.60(±0.30)	10.54(±0.23)	10.45(±0.20)	10.32(±0.26)
	NN	11.79(±0.61)	11.13(±0.56)	11.14(±0.56)	11.10(±0.53)
	SCL_NL	14.45(±1.07)	13.13(±1.03)	12.84(±1.17)	12.79(±1.18)
	SCL_EXP	13.49(±1.76)	12.55(±1.39)	12.45(±1.35)	12.38(±1.31)
	EXP	10.97(±1.01)	10.56(±0.60)	10.38(±0.36)	10.37(±0.34)
	LOG	11.70(±1.06)	11.04(±0.72)	10.49(±0.59)	10.44(±0.55)
	Warm-up+PLA	**65.88(±1.51)**	**43.29(±0.70)**	**41.30(±0.59)**	**40.28(±0.39)**
SVHN	Oracle	91.95(±0.11)	53.89(±0.11)	50.59(±0.23)	47.05(±0.22)
	Two-stage	**90.62(±0.27)**	53.92(±0.33)	50.86(±0.28)	47.31(±0.11)
	FORWARD	19.59(±0.00)	19.59(±0.00)	19.59(±0.00)	19.68(±0.00)
	FREE	19.59(±0.00)	19.59(±0.00)	19.59(±0.00)	19.68(±0.00)
	NN	19.59(±0.00)	19.59(±0.00)	19.59(±0.00)	19.68(±0.00)
	SCL_NL	19.59(±0.00)	19.59(±0.00)	19.58(±0.00)	19.67(±0.00)
	SCL_EXP	19.59(±0.00)	19.59(±0.00)	19.58(±0.00)	19.68(±0.00)
	EXP	19.60(±0.01)	19.60(±0.02)	19.56(±0.03)	19.65(±0.04)
	LOG	19.59(±0.00)	19.60(±0.02)	19.59(±0.00)	19.68(±0.00)
	Warm-up+PLA	90.50(±0.16)	**54.58(±0.10)**	**51.04(±0.04)**	**47.47(±0.13)**

Extensive experiments validate the effectiveness of our method. It could even achieve a similar performance to the oracle (i.e., AT with ordinary labels). The ablation study demonstrates the necessity of both components in the proposed framework. We report Nat., PGD20, CW30, and AA within three runs.

Table 3: Means (standard deviations) of natural and adversarial test accuracy on Kuzushiji.

Method	Natural	PGD	CW	AA
FORWARD [39]	35.48(±27.96)	29.84(±25.55)	28.09(±24.37)	22.01(±18.98)
+Warm-up	**91.39(±0.60)**	**82.95(±0.47)**	**79.69(±0.52)**	**61.66(±0.86)**
FREE [18]	16.17(±1.77)	12.01(±0.55)	9.33(±1.50)	4.08(±1.60)
+Warm-up	**83.19(±1.39)**	**74.05(±1.30)**	**70.48(±1.17)**	**57.39(±0.80)**
NN [18]	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	8.87(±1.60)
+Warm-up	**86.43(±0.67)**	**77.85(±0.98)**	**74.55(±1.03)**	**59.84(±0.73)**
SCL_NL [7]	40.83(±24.03)	32.82(±22.88)	29.93(±22.53)	20.86(±19.74)
+Warm-up	**91.92(±0.29)**	82.93(±0.58)	79.77(±0.96)	62.27(±0.68)
+PLA	39.64(±32.18)	32.18(±30.82)	28.37(±30.97)	21.58(±23.01)
+Warm-up+PLA	91.74(±0.85)	**86.09(±1.01)**	**83.77(±1.03)**	**67.87(±0.94)**
SCL_EXP [7]	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	8.21(±2.54)
+Warm-up	**89.09(±0.58)**	80.94(±0.70)	78.12(±0.74)	61.17(±1.17)
+PLA	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)
+Warm-up+PLA	87.58(±2.48)	**82.17(±2.45)**	**80.25(±2.41)**	**66.36(±2.02)**
EXP [13]	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)
+Warm-up	**89.18(±0.21)**	**80.80(±0.21)**	**77.30(±0.35)**	60.90(±0.25)
+PLA	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)	10.00(±0.00)
+Warm-up+PLA	84.63(±0.94)	79.37(±1.27)	77.23(±1.42)	**65.04(±1.33)**
LOG [13]	32.66(±25.50)	26.87(±22.66)	24.90(±21.20)	18.78(±16.95)
+Warm-up	91.31(±0.53)	82.77(±0.33)	80.31(±0.25)	62.23(±0.42)
+PLA	57.78(±33.80)	53.10(±30.49)	51.11(±29.08)	39.65(±21.14)
+Warm-up+PLA	**91.60(±0.49)**	**85.88(±0.48)**	**83.74(±0.35)**	**68.75(±0.68)**

Welcome to check out our paper and source code for more details and information!

Paper Code