Well-classified Examples are Underestimated in Classification with Deep Neural Networks

Guangxiang Zhao, Wenkai Yang, Xuancheng Ren, Lei Li, Yunfang Wu, Xu Sun.

Peking University

AAAI-22

Introduction

Have you ever trained deep classification models with Cross-Entropy (CE) loss, and been told to focus on hard examples but ignore the easy ones?

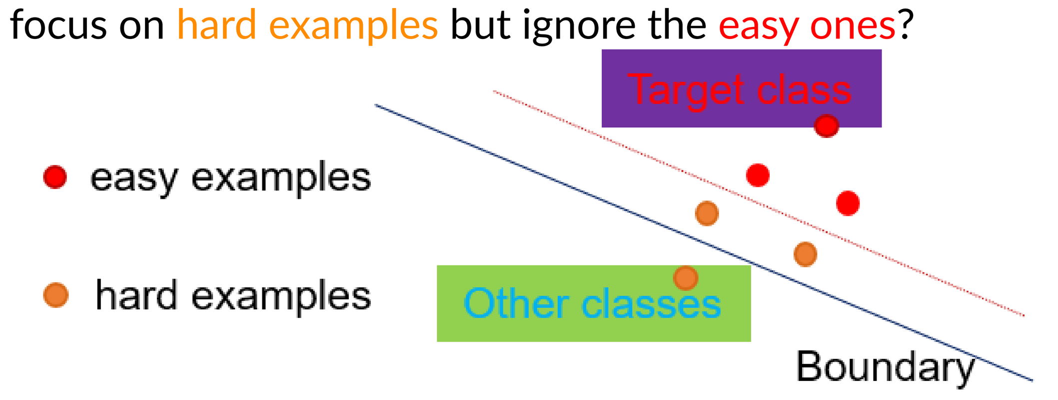


Figure 1. Easy examples are far from the decision boundary, all circles' label is the target class.

We find that common belief does not stand! Learning easy examples actually improves **margin**, **robustness** and **performance** on **graph**, **text**, and **image** classification tasks, without tuning *any* original hyperparameters.

Common practice in classification with deep neural networks

Cross Entropy loss and its variations dominate the training of deep classification models.

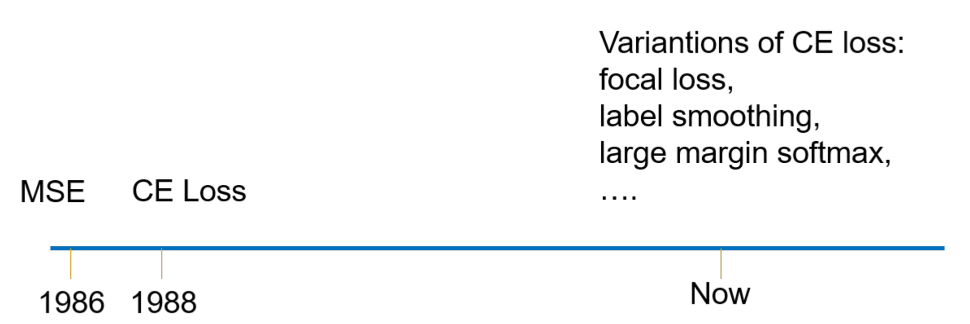


Figure 2. History of classification with deep neural networks.

What is the conventional wisdom?

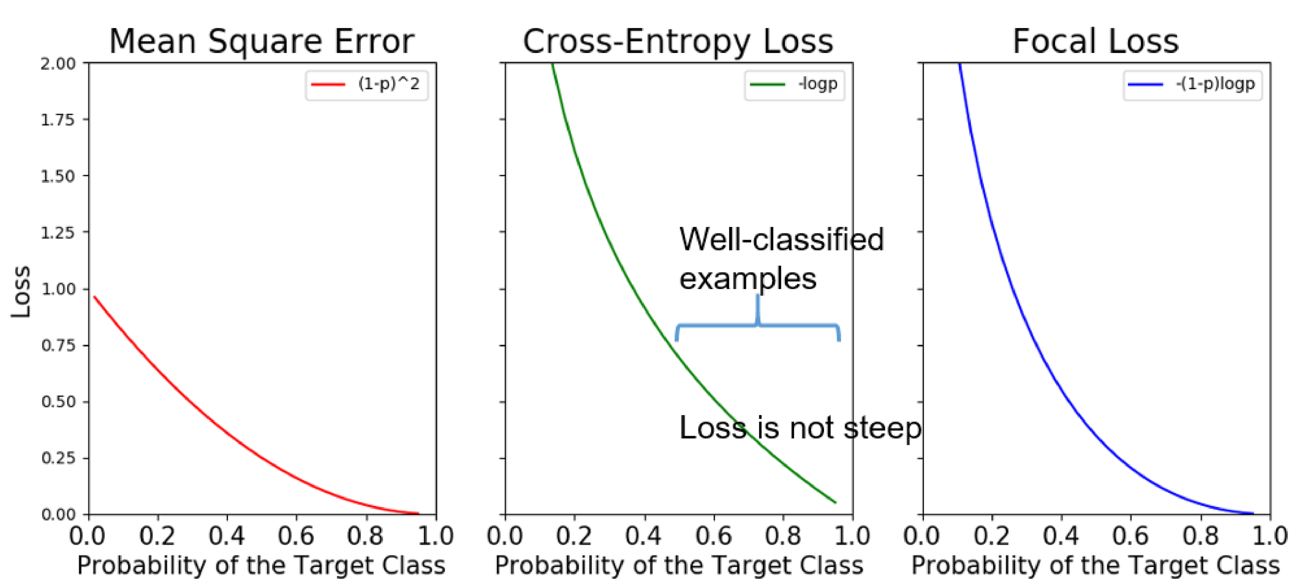
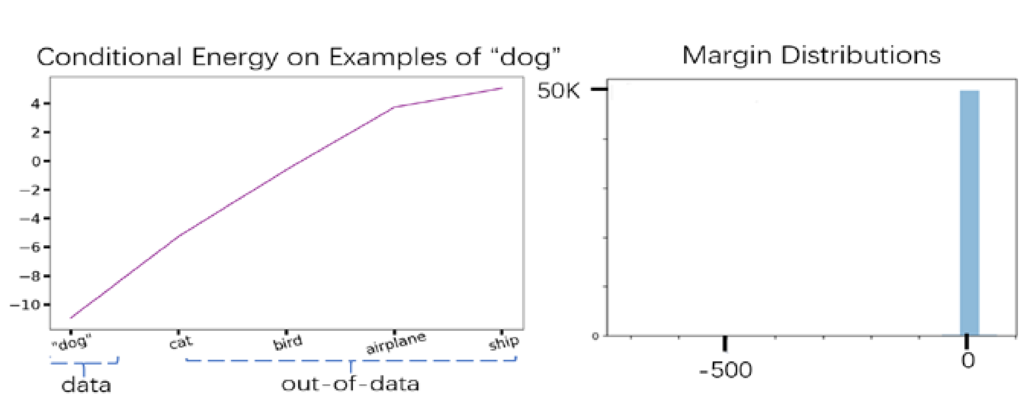


Figure 3. Well-classified examples receive less attention in optimization.

They all focus on bad-classified examples but ignore well-classified examples that are far from the decision boundary and have high probability regarding the target.

We doubt that common practice, two facts inspire us:

- Recent studies find that **down-weighting** the learning of examples with common classes hinders the representation learning.
- We observe that energy surface around data is not sharp, and margins are small.



How does CE Loss suffer from underestimating well-classified examples theoretically?

CE loss is to minimize negative log likelihood:

$$\mathcal{L}_{NLL} = -\log p_\theta(y \mid \boldsymbol{x}) = -\log p_\theta(\boldsymbol{x})[y]. \tag{1}$$

There are three issues:

1. Normalization function brings a **gradient** vanishing problem to CE loss and hinders the **representation learning** from easy examples.

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \boldsymbol{\theta}} = (p-1)\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]}{\partial \boldsymbol{\theta}}. \tag{2}$$

2.CE loss has power in the opposite direction for reducing the **energy** on the data manifold. $E_\theta(y \mid \boldsymbol{x}) = -f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]$.

$$\mathcal{L}_{NLL} = E_\theta(y \mid \boldsymbol{x}) + \log[\exp(-E_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})) + \sum_{y' \neq y} \exp(-E_{\boldsymbol{\theta}}(y' \mid \boldsymbol{x}))]. \tag{3}$$

3. CE loss is not effective in enlarging **margins**. When the prediction gets close to the target during training, $A = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})[y'] - f_{\boldsymbol{\theta}}(\boldsymbol{x})[y])$ gets close to 0, but the denominator has a constant 1, so the incentive to further enlarge the margin gets close to 0.

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \boldsymbol{\theta}} = \frac{\sum_{y' \neq y} A(\frac{\partial (f_{\boldsymbol{\theta}}(\boldsymbol{x})[y'] - f_{\boldsymbol{\theta}}(\boldsymbol{x})[y])}{\partial \boldsymbol{\theta}})}{1 + \sum_{y' \neq y} A}. \tag{4}$$

What can we gain from reviving the learning of well-classified examples theoretically?

1. We define **Encouraging Loss (EL)** that revives the learning of well-classified examples by **rewarding correct predictions**.

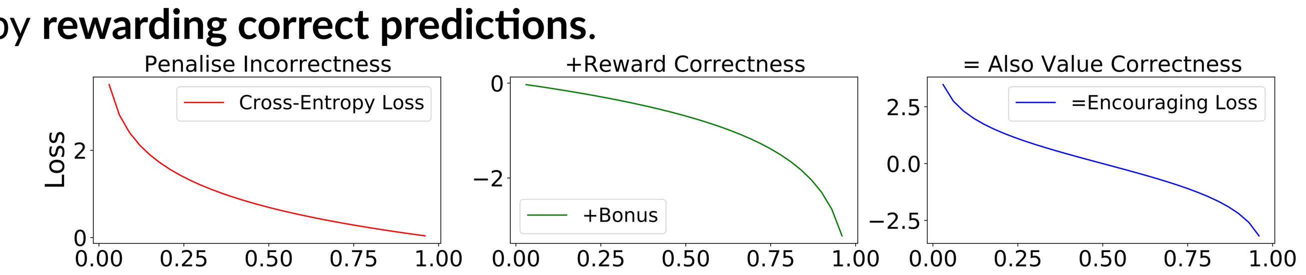


Figure 4. Illustration of the counterexample: the encouraging loss.

2. Enhancing the learning of well-classified examples by EL solve the above three issues.

$$\mathcal{L}_{NLL} = -\log p_\theta(y \mid \boldsymbol{x}) = -\log p_\theta(\boldsymbol{x})[y]. \qquad \mathcal{L}_{EL} = -\log p_{\boldsymbol{\theta}}(\boldsymbol{x})[y]) + \underline{log(1 - p_{\boldsymbol{\theta}}(\boldsymbol{x})[y])}$$

a. Gradient

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \boldsymbol{\theta}} = \underline{(p-1)}\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]}{\partial \boldsymbol{\theta}}. \qquad \frac{\partial \mathcal{L}_{EL}}{\partial \boldsymbol{\theta}} = \underline{-1} \cdot \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})[y]}{\partial \boldsymbol{\theta}}.$$

a. The first term of EL is always 1.

b. Energy

$$\mathcal{L}_{NLL} = E_\theta(y \mid \boldsymbol{x}) + log[\exp(-E_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})) + \sum_{y' \neq y} \exp(-E_{\boldsymbol{\theta}}(y' \mid \boldsymbol{x}))].$$

$$\mathcal{L}_{EL} = E_\theta(y \mid \boldsymbol{x}) - log[\sum_{y' \neq y} \exp(-E_{\boldsymbol{\theta}}(y' \mid \boldsymbol{x}))].$$

b. EL has no barrier in the second term.

c. Margin

$$\frac{\partial \mathcal{L}_{NLL}}{\partial \boldsymbol{\theta}} = \frac{\sum_{y' \neq y} A(\frac{\partial (f_{\boldsymbol{\theta}}(\boldsymbol{x})[y'] - f_{\boldsymbol{\theta}}(\boldsymbol{x})[y])}{\partial \boldsymbol{\theta}})}{\underline{1} + \sum_{y' \neq y} A}. \qquad \frac{\partial \mathcal{L}_{EL}}{\partial \boldsymbol{\theta}} = \frac{\sum_{y' \neq y} A(\frac{\partial (f_{\boldsymbol{\theta}}(\boldsymbol{x})[y'] - f_{\boldsymbol{\theta}}(\boldsymbol{x})[y])}{\partial \boldsymbol{\theta}})}{\sum_{y' \neq y} A}.$$

c. 1 is removed from denominator.

3. We also design conservative bonuses (please refer to the paper), which partly solve these issues and get considerate performance improvement.

Video

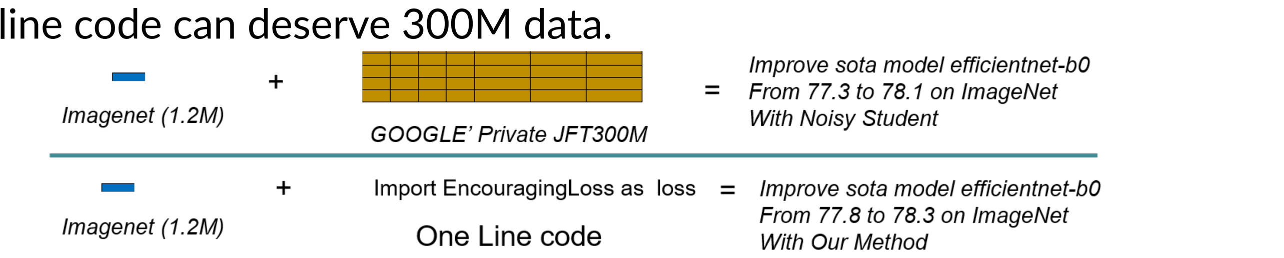
Code

Practical effect of learning well-classified examples

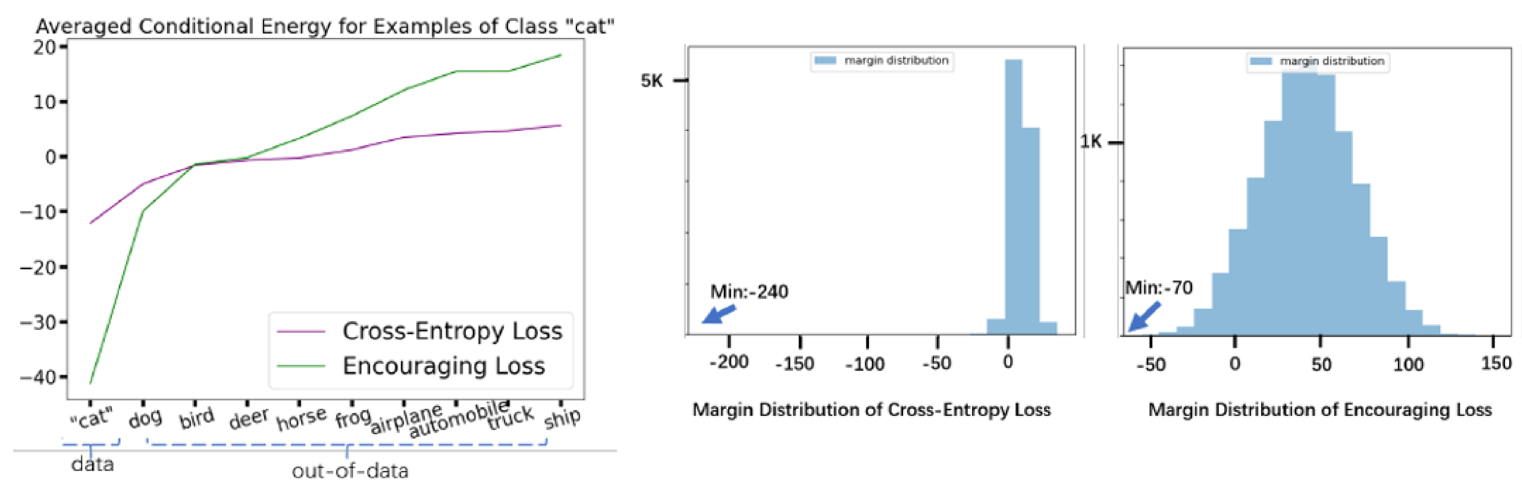
1. It improves performance across datasets (image classification datasets MNIST, CIFAR 10/100, ImageNet, graph classification datasets Proteins, NCI1, machine translation directions De-En, Fr-En) and models (ResNet, EfficientNet, Graph convolution, Transformer), some are from conservative bonus.

Setting	MNIST	C10-r50	C10-eb0	C100-r50	C100-eb0	Img-r50
CE	99.42±0.06	92.34±0.70	93.21±0.40	74.39±0.70	76.22±0.37	75.49±0.28
EL	99.56±0.05	92.97±0.42	94.24±0.17	75.80±0.09	77.21±0.26	76.43±0.15
Setting	**Img-eb0**	**Proteins**	**NCI1**	**De-En**	**Fr-En**	
CE	77.80±0.15	72.32±0.12	70.53±0.28	35.09±0.07	37.10±0.06	
EL	78.28±0.13	72.76±0.18	71.04±0.38	35.50±0.11	37.73±0.17	

Even one line code can deserve 300M data.



2. We can empirically address the issues: we reduce conditional energy E(y|x) by **4x** and move the majority of margin **from 0 to 50**.



3. We can deal with real scenarios since we deal with the three issues.

3.a We empirically verify that the traditional re-weighting at the sample level (CE loss down-weights the importance of well-classified samples) is also harmful to representation learning.

Decoupling Reps&Cls: Do not down-weight common classes in representation learning

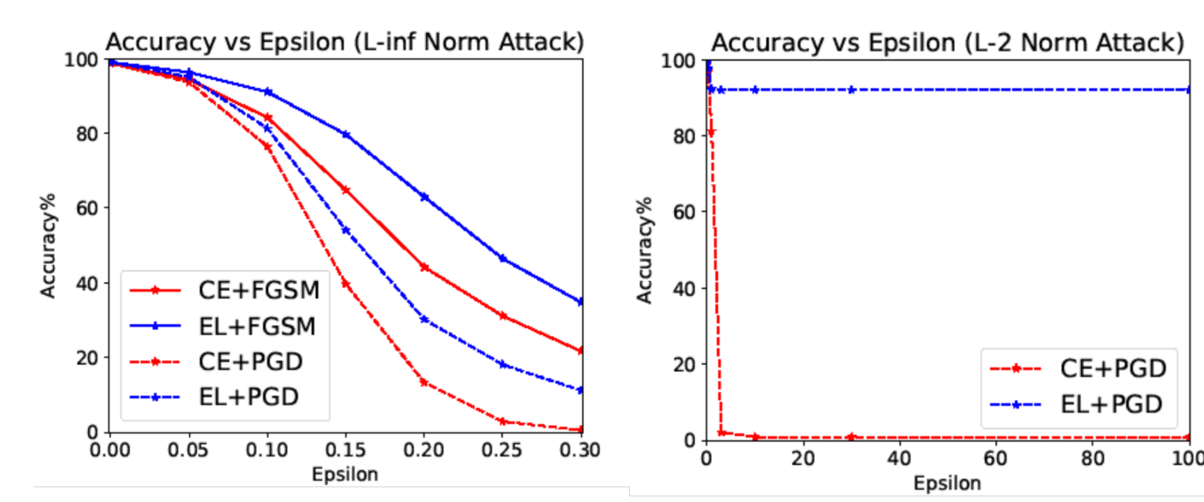
+Normal Bonus: Nor does it down-weight the learning from easy examples.

Method	iNaturalList2018			
	Overall	Many	Medium	Few
CE loss	64.3	74.1	65.9	59.8
+ Conservative Bonus (LE=0.5)	65.3	74.3	66.6	61.2
+ Normal Bonus	65.8	74.4	66.6	62.4
+ Aggressive Bonus	66.3 (+2.0)	**75.1** (+1.0)	67.4(+1.6)	62.6(+2.8)
Decoupling Reps&Cls (CRT)	64.9	71.4	65.9	61.9
+ Normal Bonus	66.8(+1.9)	71.7(+0.3)	67.6(+1.7)	64.6(+2.7)
Deferred Re-weighting	68.1	71.0	68.3	67.1
+ Normal Bonus	**70.3**(+2.2)	69.0(-2.0)	**70.1**(+1.8)	**70.9**(+3.8)

3.b The discriminative energy distribution help OOD detection.

Setting	MNIST vs. F. MNIST		Img vs.iNautralist2018	
Metric	AUROC↑	FPR95↓	AUROC↑	FPR95↓
CE	95.68	20.92	76.54	75.40
EL	98.04	8.69	78.41	71.86

3.c The large margin by learning easy examples improves adversarial robustness.



4. This surprise is applicable to MSE and variations of CE loss (refer to the paper).