Repetitive Reprediction Deep Decipher for Semi-Supervised Learning

Guo-Hua Wang, Jianxin Wu

National Key Laboratory for Novel Software Technology, Nanjing University, China

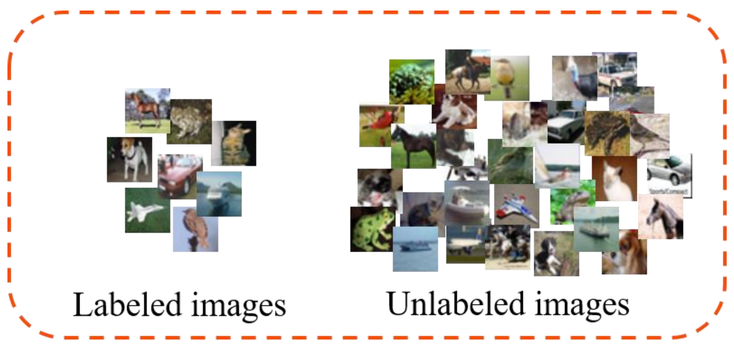
wangguohua@lamda.nju.edu.cn, wujx2001@nju.edu.cn

AAAI-20, New York

Problem

Semi-Supervised Deep Learning

- Semi-supervised learning
- Deep learning
- Image classification



Full training dataset

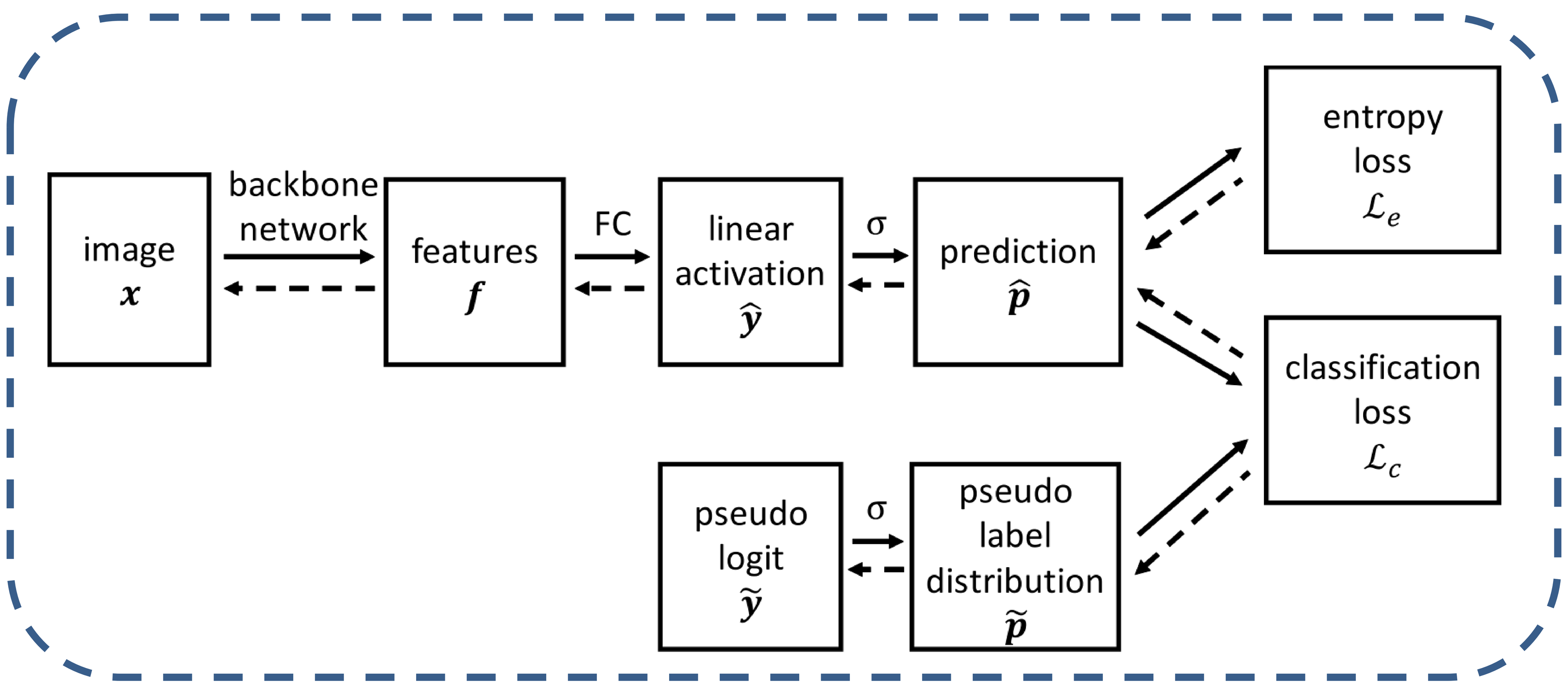
- Major issues
  - How to train deep network with the help of unlabeled data?
  - ***Why predictions are good candidates for pseudo-labels***?
  - Why pseudo-labels will become uncertain (flat) during training?

Contributions

- We propose ***deep decipher*** (D2), a deep learning framework that deciphers the relationship between network predictions and pseudo-labels. D2 updates pseudo-labels ***by back-propagation***.
- Within D2, we prove that ***pseudo-labels are exponentially transformed from the predictions***.
- We prove that pseudo-labels will become flat during the optimization. To mitigate this problem, we propose a simple but effective remedy, ***repetitive reprediction*** (R2).

The R2-D2 Method

Deep Decipher



- Loss function $\mathcal{L} = \alpha\mathcal{L}_c + \beta\mathcal{L}_e$

$$= \alpha\sum_{j=1}^{N}\hat{p}_j\left[\log(\hat{p}_j) - \log(\tilde{p}_j)\right] - \beta\sum_{j=1}^{N}\hat{p}_j\log(\hat{p}_j)$$

- ***An exponential link between pseudo-labels and predictions***

Theorem 1 *Suppose D2 is trained by SGD with the loss function $\mathcal{L} = \alpha\mathcal{L}_c + \beta\mathcal{L}_e$. Let $\hat{\mathbf{p}}$ denote the prediction by the network for one example and $\hat{p}_n$ is the largest value in $\hat{\mathbf{p}}$. After the optimization algorithm converges, we have $\tilde{p}_n \to \exp(-\frac{\mathcal{L}}{\alpha})\,(\hat{p}_n)^{1-\frac{\beta}{\alpha}}$.*

- Pseudo-labels will become flat during the optimization

Theorem 2 *Suppose D2 is trained by SGD with the loss function $\mathcal{L} = \alpha\mathcal{L}_c + \beta\mathcal{L}_e$. If $\tilde{p}_n = \exp(-\frac{\mathcal{L}}{\alpha})\,(\hat{p}_n)^{1-\frac{\beta}{\alpha}}$, we must have $\tilde{p}_n \le \hat{p}_n$.*

- An equality constraint bias

$\sum_{i=1}^{N}\tilde{y}_i$ will not change during D2 training.

Repetitive Reprediction

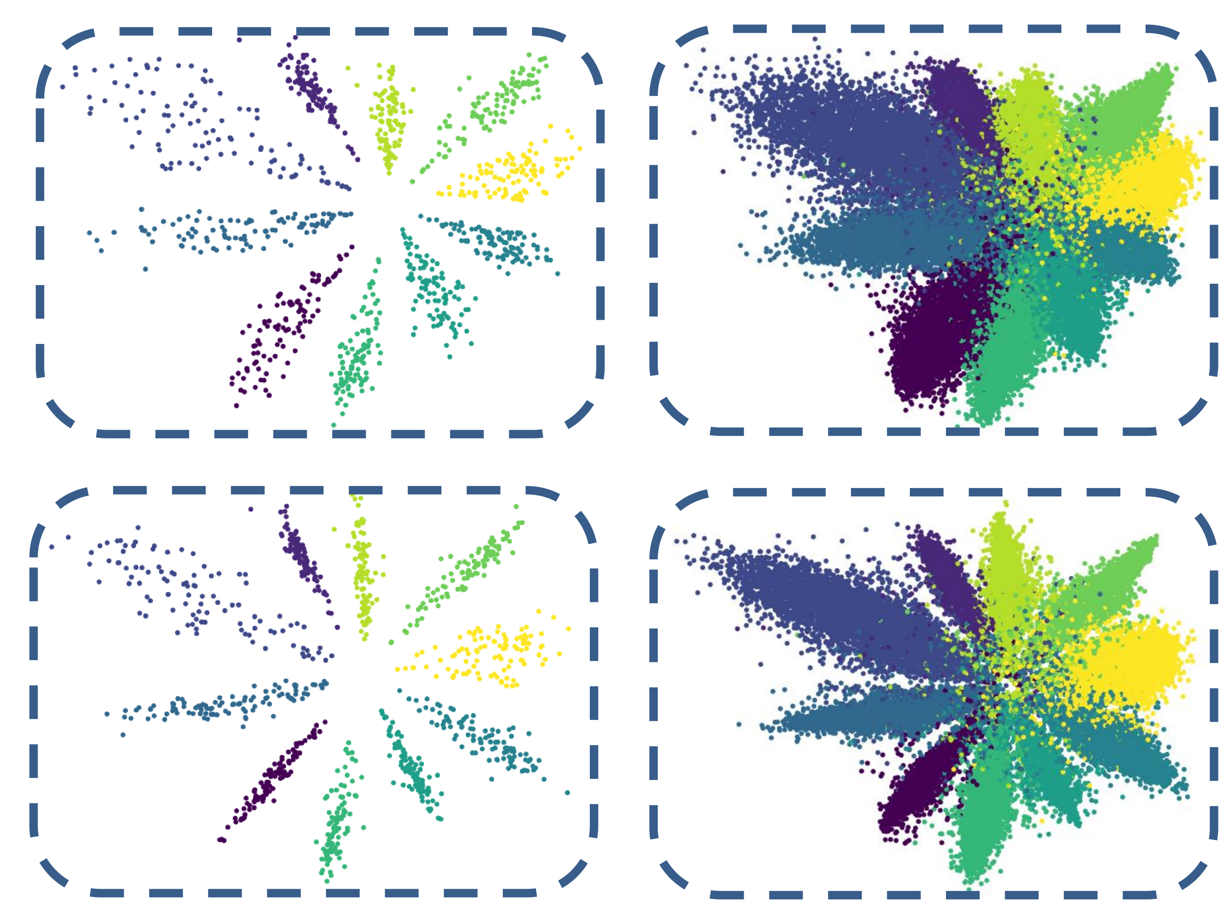
- Using the prediction to re-initialize the pseudo-labels several times during training D2.
- Benefits of R2:
  - Make pseudo-labels sharper and more accurate.
  - Reduce the impact of equality constraint bias.

An Illustrative Example

MNIST & LeNet with 2D-FC

Train with only labeled images and CE loss

Train D2 with all images



feature distribution of labeled images

feature distribution of labeled images + unlabeled images

The overall R2-D2 algorithm

- 1st stage

  Use only labeled images to train the backbone network with cross entropy loss.
- 2nd stage

  Predict pseudo-labels for unlabeled images.

  Use D2 to train the network and optimize pseudo-labels together.
- repeat 2nd stage several times
- 3rd stage

  Finetune the network by pseudo-labels.

Experiments

ImageNet

Method	Backbone	#Param	Top-1	Top-5
100% Supervised	ResNet-18	11.6M	30.43	10.76
10% Supervised	ResNet-18	11.6M	52.23	27.54
Stochastic Transformations	AlexNet	61.1M	-	39.84
VAE with 10% Supervised	Customized	30.6M	51.59	35.24
Mean Teacher	ResNet-18	11.6M	49.07	23.59
Dual-View Deep Co-Training	ResNet-18	11.6M	46.50	22.73
R2-D2	ResNet-18	11.6M	**41.55**	**19.52**

CIFAR-10

Method	Backbone	Error rates (%)
100% Supervised	Shake-Shake	2.86
Only 4000 labeled images	Shake-Shake	14.90 ± 0.28
Mean Teacher	ConvLarge	12.31 ± 0.28
Temporal Ensembling	ConvLarge	12.16 ± 0.24
VAT+EntMin	ConvLarge	10.55 ± 0.05
DCT with 8 Views	ConvLarge	8.35 ± 0.06
Mean Teacher	Shake-Shake	6.28 ± 0.15
HybridNet	Shake-Shake	6.09
R2-D2	Shake-Shake	$\mathbf{5.72 \pm 0.06}$

CIFAR-100

Method	Backbone	Error rates (%)
100% Supervised	ConvLarge	26.42 ± 0.17
Using 10000 labeled images only	ConvLarge	38.36 ± 0.27
Temporal Ensembling	ConvLarge	38.65 ± 0.51
LP	ConvLarge	38.43 ± 1.88
Mean Teacher	ConvLarge	36.08 ± 0.51
LP + Mean Teacher	ConvLarge	35.92 ± 0.47
DCT	ConvLarge	34.63 ± 0.14
R2-D2	ConvLarge	$\mathbf{32.87 \pm 0.51}$

Ablation studies

	a	b	c	d	e
The 2nd stage	✓	✓	✓	✓	✓
Repeat the 2nd stage		✓	✓	✓	✓
Reprediction			✓		✓
Reducing LR				✓	✓
Error rates (%)	6.71	6.37	6.23	5.94	5.78

- $\alpha = 0.1, \beta = 0.3$ in all experiments
- ImageNet

  labeled: 128,000 unlabeled: 14,069,122
- CIFAR-100

  labeled: 10,000 unlabeled: 40,000
- CIFAR-10

  labeled: 4,000 unlabeled: 46,000
- Code

  https://github.com/DoctorKey/R2D2.pytorch

Learning And Mining from DatA

http://lamda.nju.edu.cn