Universum Prescription: Regularization using Unlabeled Data

Xiang Zhang, Yann LeCun
Courant Institute of Mathematical Sciences, New York University

Simple Idea: Prescribe "Junk" Labels

NYU

- **For a classification problem, if you have unlabeled data that satisfy**
  - **Universum assumption**: the possibility that the unlabeled data belongs to a supervised class is ignorable
- **Then, you can prescribe "none-of-the-above" labels to these unlabeled data and reduce the gap between training and testing errors.**
- **Too simple therefore cannot be a new idea**
  - But we have practical and theoretical results to show its effectiveness.

Is the Universum assumption practical?

NYU

- **We think so**
  - By **problem design**: we only care about a small number of classes among all possible classes.
    - CIFAR-10/100 (~60,000) vs TinyImages (~80,000,000)
  - By **data acquisition**: sometimes irrelevant data can be acquired either as a by-product or just because it is easy.
    - ImageNet 1000-class Classification Task (~1.28 million)
    - ImageNet 20,000-class 2011 Release (~20 million). 12 million after removal of data that belong to or is a sub-ordinate class of the 1000 ones.

But How to Prescribe "Junk" Labels?

NYU

- **Let's assume we are using negative-log-likelihood loss**
- **Uniform Prescription**

$$L(h, x, y) = -\sum_{i=1}^{k} Q[Y = i|x] \log \Pr[Y = i|x, h],$$

- **Dustbin Class**

$$L(h, x, k+1) = h_{k+1}(x) + \log \left[\sum_{i=1}^{k+1} \exp(-h_i(x)) \right]$$

- **Background Class**

$$L(h, x, k+1) = \tau + \log \left[\exp(-\tau) + \sum_{i=1}^{k} \exp(-h_i(x)) \right]$$

Experimental Results

Table: Result for universum prescription. The numbers are percentages. The three numbers in each tabular indicate training error, testing error and generalization gap. Bold numbers are the best ones for each case. CIFAR-100 F. and CIFAR-100 C. stand for fine-grained and coarse classification problems of CIFAR-100. STL-10 Tiny stands for using 80 million images as the unlabeled dataset. ImageNet-1 and ImageNet-5 are errors for top-1 and top-5 evaluation for ImageNet dataset.

DATASET	BASELINE			UNIFORM			DUSTBIN			BACKGROUND		
	Train	Test	Gap	Train	Test	Gap	Train	Test	Gap	Train	Test	Gap
CIFAR-10	**0.00**	7.02	7.02	0.72	7.59	6.87	0.07	**6.66**	**6.59**	1.35	8.38	7.03
CIFAR-100 F.	**0.09**	37.58	37.49	4.91	36.23	31.32	2.52	**32.84**	**30.32**	8.56	40.57	42.01
CIFAR-100 C.	**0.04**	22.74	22.70	0.67	23.42	22.45	0.40	**20.45**	**20.05**	3.73	24.97	21.24
STL-10	**0.00**	**31.16**	31.16	2.02	36.54	34.52	3.03	36.58	33.55	14.89	38.95	**24.06**
STL-10 Tiny	**0.00**	31.16	31.16	0.62	30.15	29.47	0.00	**27.96**	**27.96**	0.11	30.38	30.27
ImageNet-1	**10.19**	34.39	24.20	13.84	34.61	20.77	13.80	**33.67**	**19.87**	13.43	34.69	21.26
ImageNet-5	**1.62**	13.68	12.06	3.02	13.70	10.68	2.83	**13.35**	**10.52**	2.74	13.84	11.10

Why does it work?

NYU

Theorem in PAC-Learning Framework

Assume we have a joint problem where $p \leq 0.5$ and there are m random training samples from the joint distribution $(1-p)\mathbf{D} + p\mathbf{U}$. With probability at least $1-\delta$, the following holds

$$\mathfrak{R}_n(\mathcal{F}, \mathbf{D}) \leq \frac{2-p}{(1-p)\left(1-p-\sqrt{\frac{\log(1/\delta)}{2m}}\right)} \mathfrak{R}_m(\mathcal{F}, (1-p)\mathbf{D} + p\mathbf{U}), \tag{6}$$

where n is a random number indicating the number of supervised samples in the total joint samples, and m is large enough such that

$$1-p-\sqrt{\frac{\log(1/\delta)}{2m}} > 0. \tag{7}$$

More Analysis

$$\mathfrak{R}_n(\mathcal{F}, \mathbf{D}) \leq \frac{2-p}{(1-p)\left(1-p-\sqrt{\frac{\log(1/\delta)}{2m}}\right)} \mathfrak{R}_m(\mathcal{F}, (1-p)\mathbf{D} + p\mathbf{U}),$$

- **p: The probability of sampling from unlabeled data**
  - p < 0.5 in the theorem
  - The constant factor in the equation is monotonically increasing in [2, 6]
  - **Keep it small!**

The Experiments on p