Deep Gamblers: Learning to Abstain with Portfolio Theory

Liu Ziyin Liu (UTokyo), Zhikang T. Wang(UTokyo), Paul Liang(CMU), Ruslan salakhutdinov (CMU), Louis-Philippe Morency (CMU), Masahito Ueda (UTokyo),

Classification and the Inadequacy of nll loss

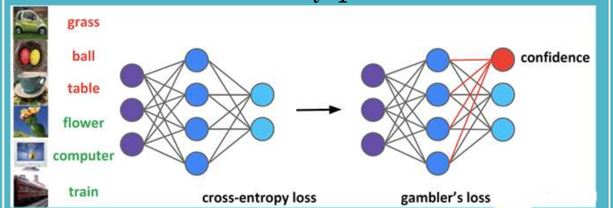
Want to find: $\theta = \arg\max_{\theta} \Pr(Y|\theta)$

In practice, minimize *negative log loss* (nll loss): $\min_{\theta} -\log p(Y|\theta)$




The proposed method: the gambler's loss

$$\max E \log(S) = \max \sum_{i=1}^{m} p_i \log(o_i b_i + b_0)$$



Intuition: Prediction as Horse Race

Horse Race with Reservation

- m horses
- Betting strategy: $\sum_{i=1}^{M} b_i \;\rightarrow\; \sum_{i=0}^{m} b_i$
- Chance of winning: p_i
- Payoff if we bet on the winning horse: o_i
- Return after winning: $S = o_i b_i \;\rightarrow\; o_i b_i + b_0$

Objective: maximize doubling rate:

$$\max W = \max E \log(S) = \max \sum_{i=1}^{m} p_i \log(o_i b_i + b_0)$$

Classification Problem = Betting problem with Reservation with $o = 1, b_0 = 0$

Classification Problem ≤ Betting problem with Reservation

Toy Example: Identifying Disconfident Images..

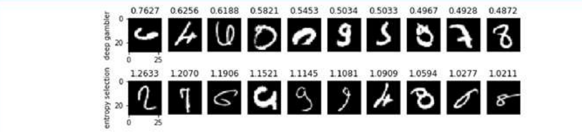
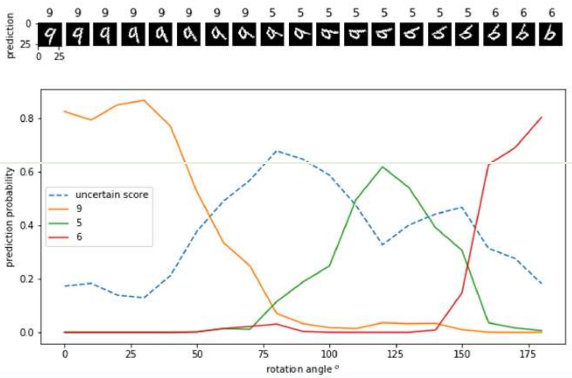
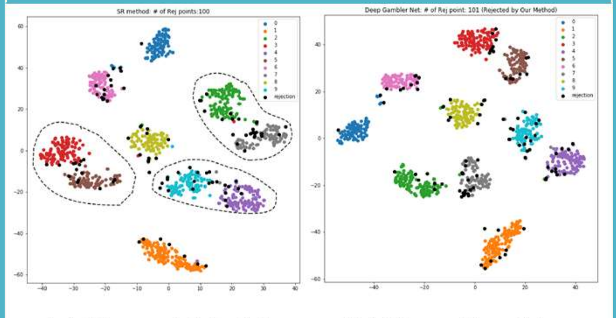


Figure 1: Top-10 rejected images in the MNIST testing set found by two methods. The number above image is the predicted uncertainty score (ours) or the entropy of the prediction (baseline). For the top-2 images, our method chooses images that are hard to recognize, while that of the baseline can be identified unambiguously by human.

Toy Example: Image Rotation..



The Learned Representation is Better Separable:

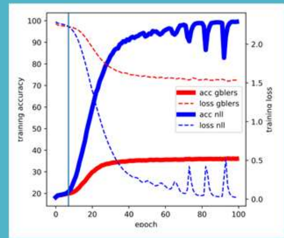


(a) Normal Model (b) Deep Gambler

Surprising Benefit:

-Training with gambler's loss reduces overfit

-Improved performance when noisy label is present



SOTA Performance...

Coverage	Ours (Best Single Model)	Ours (Best per coverage)	SR	BD	SN
1.00	$^{o=2.6}$3.24 ± 0.09	–	3.21	3.21	3.21
0.95	$^{o=2.6}$**1.36 ± 0.02**	$^{o=2.6}$**1.36 ± 0.02**	1.39	1.40	1.40
0.90	$^{o=2.6}$**0.76 ± 0.05**	$^{o=2.6}$**0.76 ± 0.05**	0.89	0.90	0.82 ± 0.01
0.85	$^{o=2.6}$**0.57 ± 0.07**	$^{o=3.6}$0.66 ± 0.01	0.70	0.71	**0.60 ± 0.01**
0.80	$^{o=2.6}$**0.51 ± 0.05**	$^{o=3.6}$**0.53 ± 0.04**	0.61	0.61	**0.53 ± 0.01**

Table 3: SVHN. The number is error percentage on the covered dataset; the lower the better. We see that our method achieved competitive results across all coverages. It is the SOTA method at coverage (0.85, 1.00).

Coverage	Ours (Single Best Model)	Ours (Best per Coverage)	SR	BD	SN
1.00	$^{o=2.0}$2.93 ± 0.17	–	3.58	3.58	3.58
0.95	$^{o=2.0}$1.23 ± 0.12	$^{o=1.4}$**0.88 ± 0.38**	1.91	1.92	1.62
0.90	$^{o=2.0}$**0.59 ± 0.13**	$^{o=2.0}$**0.59 ± 0.13**	1.10	1.10	0.93
0.85	$^{o=2.0}$0.47 ± 0.10	$^{o=1.2}$**0.24 ± 0.10**	0.82	0.78	0.56
0.80	$^{o=2.0}$**0.46 ± 0.08**	$^{o=2.0}$**0.46 ± 0.08**	0.68	0.55	**0.35 ± 0.09**

Table 5: Cats vs. Dogs. The number is error percentage on the covered dataset; the lower the better. This dataset is a binary classification, and the input images have larger resolution.

東京大学 THE UNIVERSITY OF TOKYO

Institute for Physics of Intelligence

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900