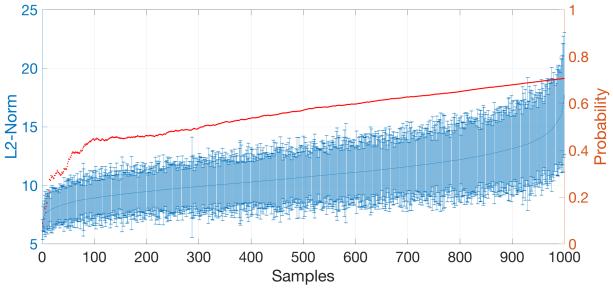
Probability Guided Maxout

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo
Media Integration and Communication Center (MICC), University of Florence

Motivation

Fact: Deep CNNs tend to overfit the training data. Regularization is essential to prevent excessive co-adaptation of hidden units.



Observation 1: low entropy output distributions are highly correlated with the L2-norm of the descriptor (penultimate layer).

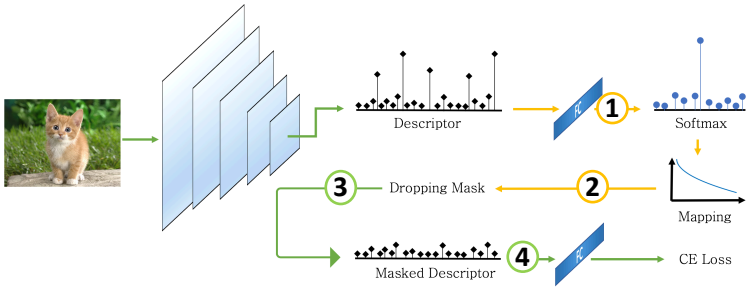
Observation 2: high L2-norm descriptors are characterized by highly-valued spikes.

Idea: regularize the training process by penalizing overconfident output distributions.

Solution: drop-out a fraction of **most active** neurons (correlated with low-entropy distributions), **proportionally** to the predicted probability of the actual class.

Code available:
https://github.com/clferrari/probability-guided-maxout

Method



1) Forward pass to estimate the actual class probability.

$$\hat{\mathbf{y}} = \frac{\exp\left(\mathcal{F}(\mathbf{x})\right)}{\sum_{j=1}^{C} \exp\left(\mathcal{F}(\mathbf{x}_j)\right)} \qquad P_{gt} = \mathbf{y} \cdot \hat{\mathbf{y}}$$

2) Estimate percentage ρ of units to drop based on P_{gt}

$$\rho(x) = \gamma x; \quad x \in [0,1] \qquad \gamma = \frac{1}{2}$$

3) Build the dropout mask $M \in \{0,1\}^d$ by sorting values of the descriptor $f \in \mathbb{R}^d$ in descending order. Apply the permutation to M and drop-out a number $p = \mathrm{d}\rho$ units:

$$\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{M}$$

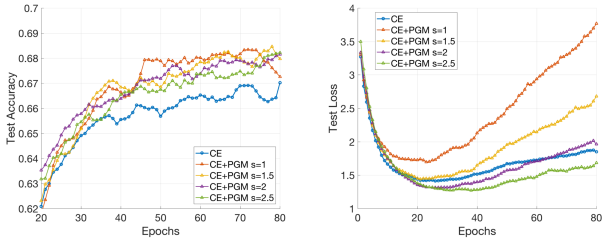
Scale the masked descriptor to maintain the expected output across train and inference with a learnable, per-sample scale factor $\mathbf{s} = \frac{\alpha}{(1-\rho)}$

4) Compute the standard CE loss

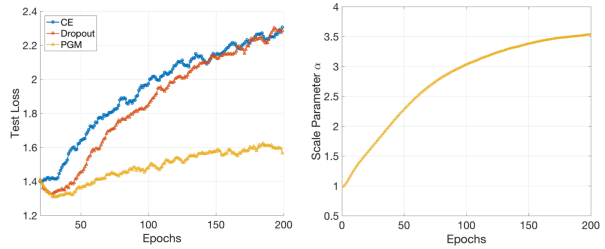
N.B. Steps 2-3 do not need gradient computation and are detached from the computational graph.

Results

Ablation: Fixed feature scaling



Finding the correct α can be complex. Let the network learn its value.



Test on benchmark datasets

Dataset	Method	Rank-1 Acc	Test Loss
CIFAR10	CE	99.69	0.006
CIFAR10	Dropout	99.71	**0.003**
CIFAR10	PGM	**99.73**	0.02
CIFAR100	CE	68.47	2.07
CIFAR100	Dropout	**69.65**	2.21
CIFAR100	PGM	69.18	**1.52**
Caltech256	CE	62.21	2.28
Caltech256	Dropout	61.61	2.30
Caltech256	PGM	**63.24**	**1.85**