UNIVERSITÀ
DEGLI STUDI
FIRENZE

MICC
Media Integration and Communication Center

Probability Guided Maxout

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo

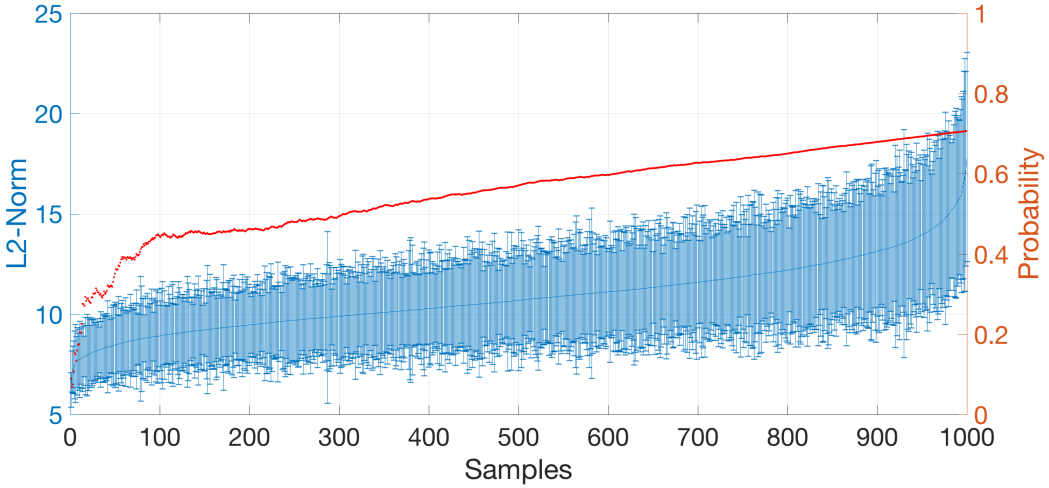
Media Integration and Communication Center (MICC)
University of Florence

Motivation

- Fact: Deep CNNs tend to overfit the training data, because of the excessive co-adaptation of hidden neurons. Regularization is essential to prevent such behavior.

- Observation 1: confident predictions (low entropy output distributions) are highly correlated with the L2-norm of the descriptor, that is the penultimate layer before the classification layer.

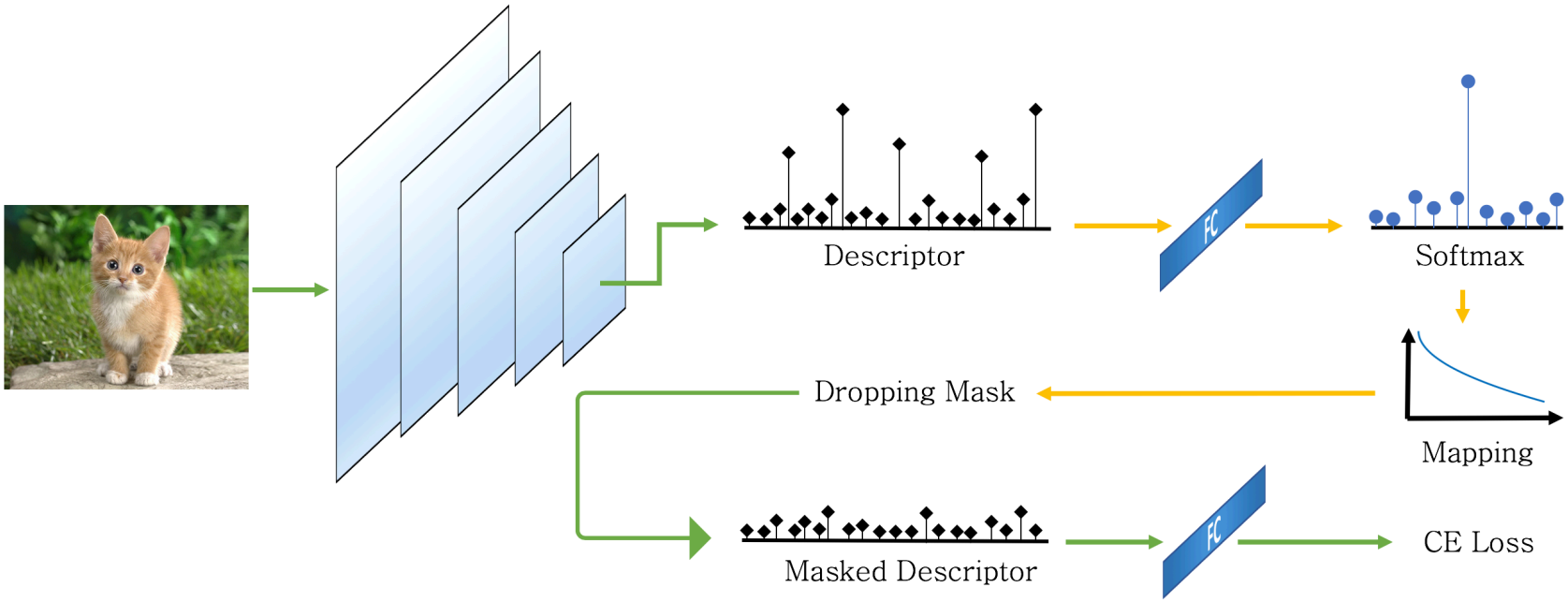
- Observation 2: high L2-norm descriptors are characterized by highly-valued spikes.



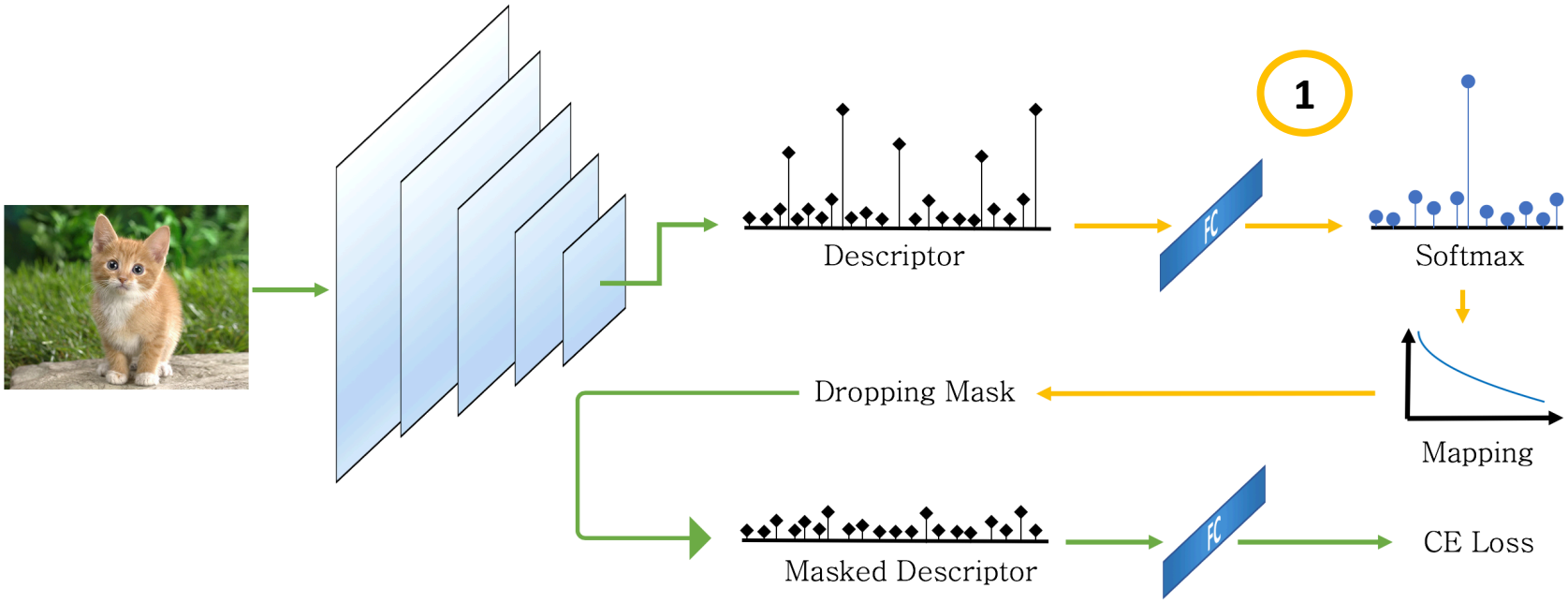
- Idea: regularize the training process by penalizing overconfident output distributions.

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

Probability Guided Maxout



Probability Guided Maxout



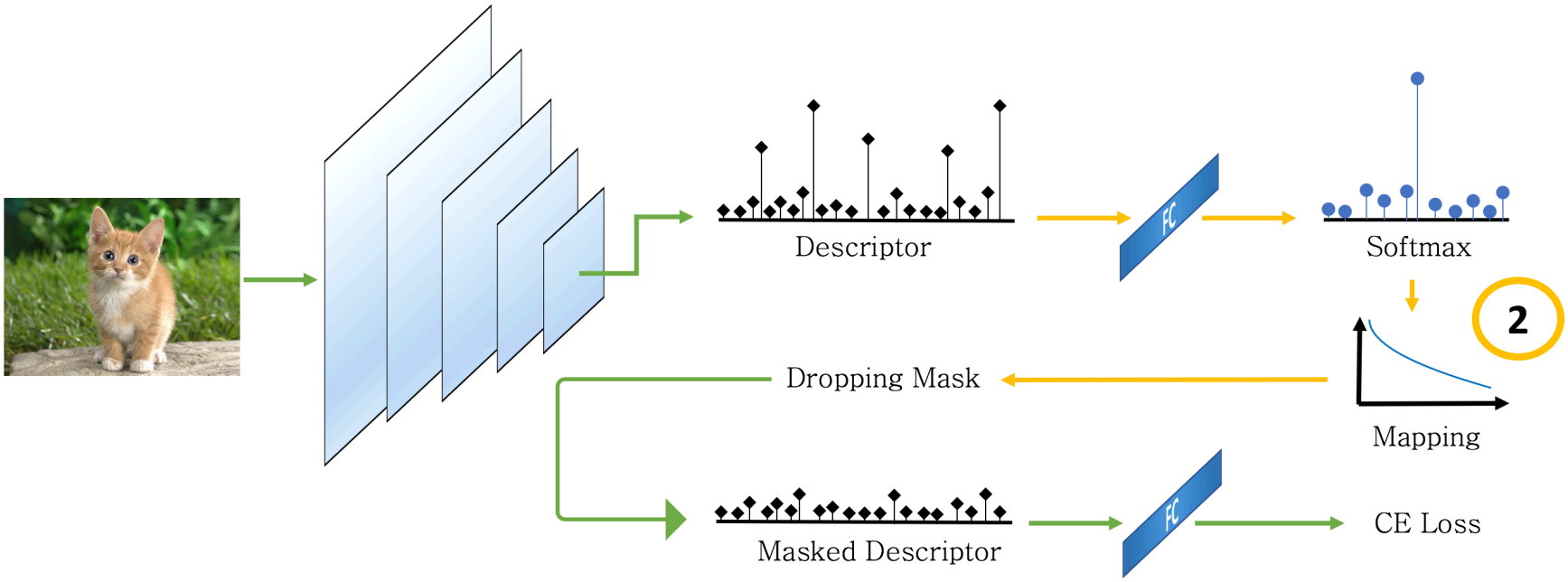
Forward pass to estimate the output probability distribution (standard cross-entropy)

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

Given the ground-truth label (one-hot encoded) $\mathbf{y}$, get the probability estimate of the ground-truth class.

$$P_{gt} = \mathbf{y} \cdot \hat{\mathbf{y}}$$

Probability Guided Maxout

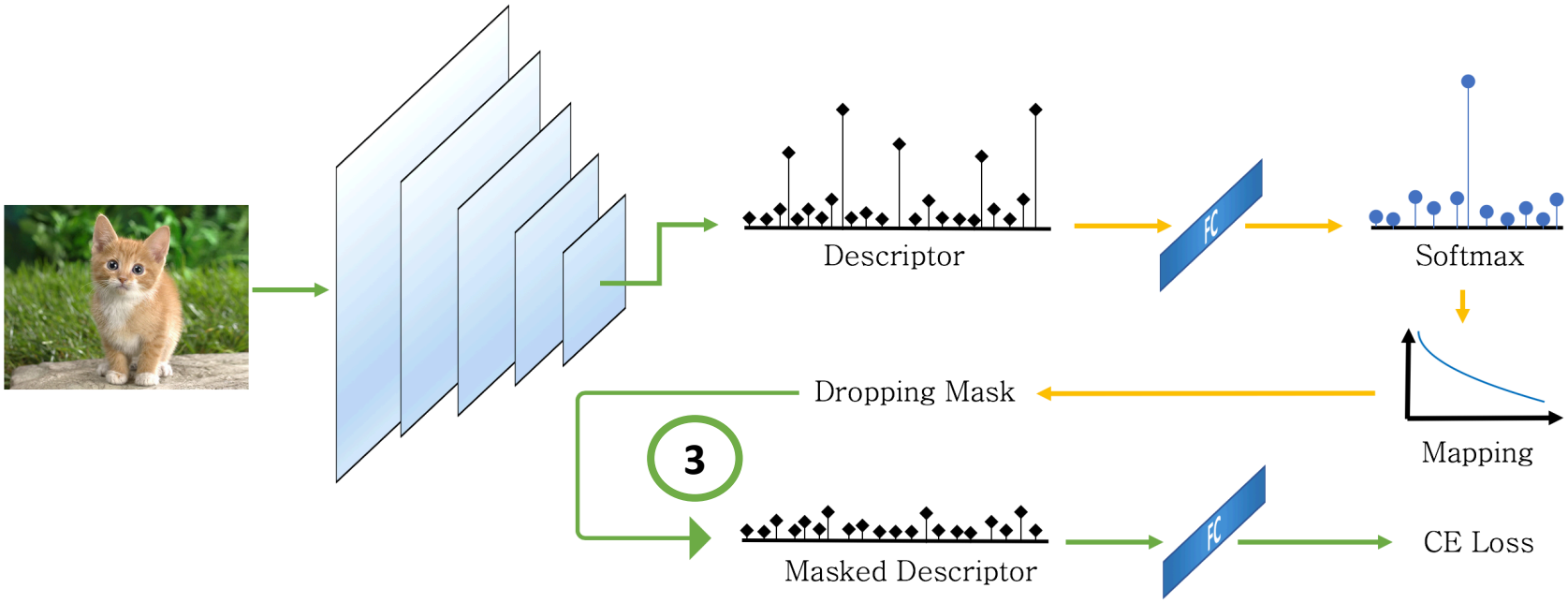


Estimate the percentage ρ of units to drop with respect to P_{gt}

1. At beginning of training, predictions are equally distributed over the C classes → $\mathrm{P}_{\mathrm{gt}} = 1/\mathrm{C}$
2. Confident predictions (large P_{gt}) should be penalized more;
3. Low values of P_{gt} indicate misclassifications; should be penalized less.

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

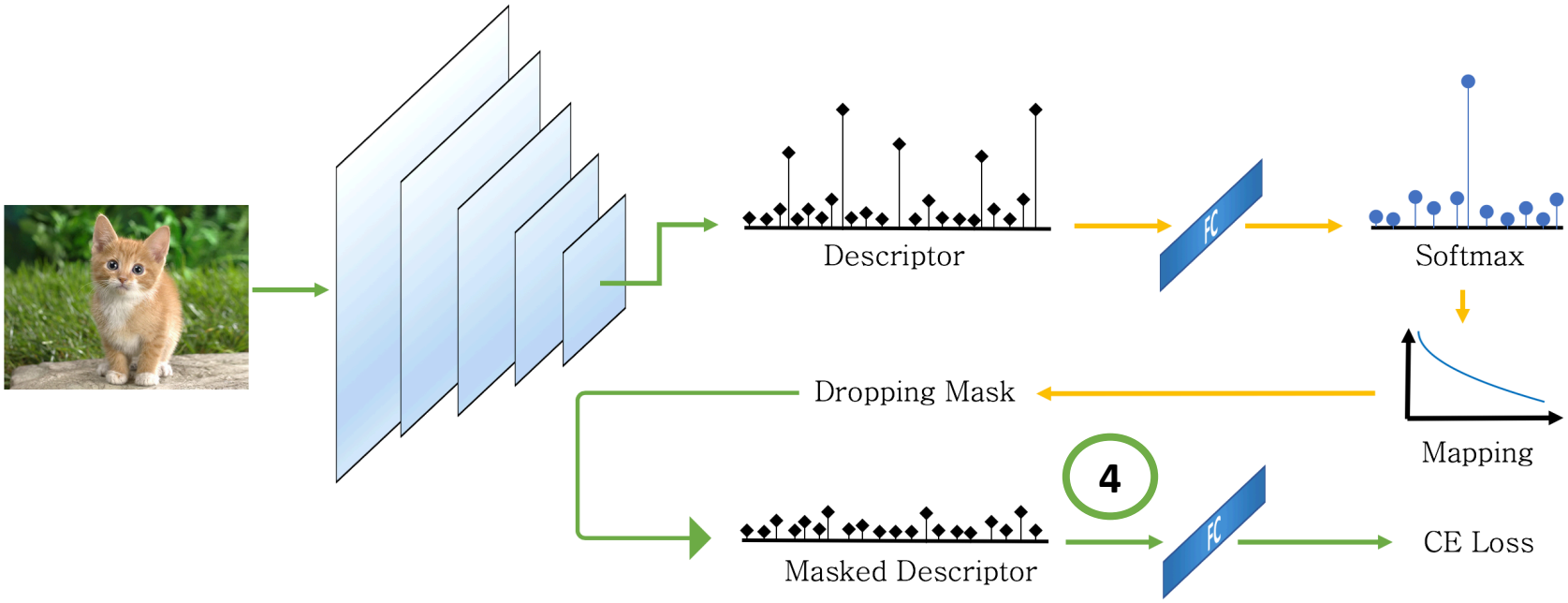
Probability Guided Maxout



Build the dropping mask and drop-out a percentage of highly active neurons.

1. Given a descriptor $f \in \mathbb{R}^d$ build a binary mask $M \in \{0,1\}^d$ with the first $p = \mathrm{d}\rho$ entries with value 0, and the last $d - p$ entries with value 1;
2. Sort values of f in descending order, and apply the sorting permutation to M;
3. Drop-out the least active neurons as $\hat{f} = f \cdot M$

Probability Guided Maxout



The final classification is performed using the masked descriptor.

NOTE:

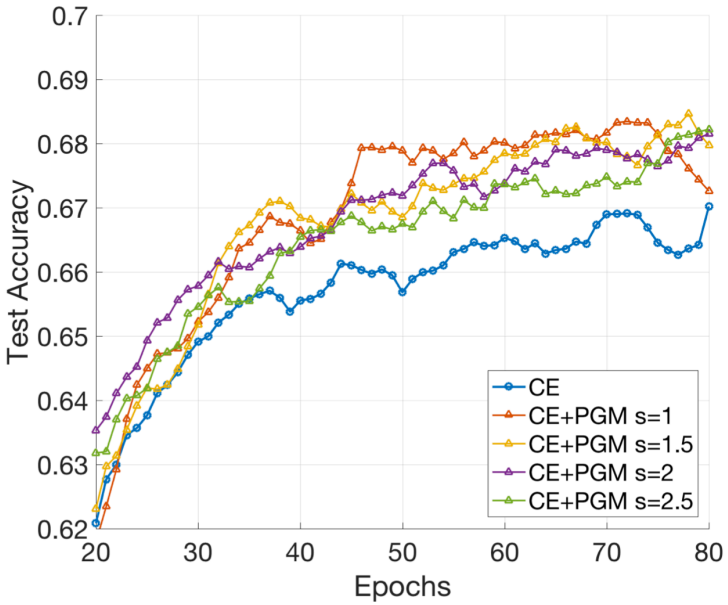
1. Estimating the dropping mask (yellow arrows) does not require gradient computation;

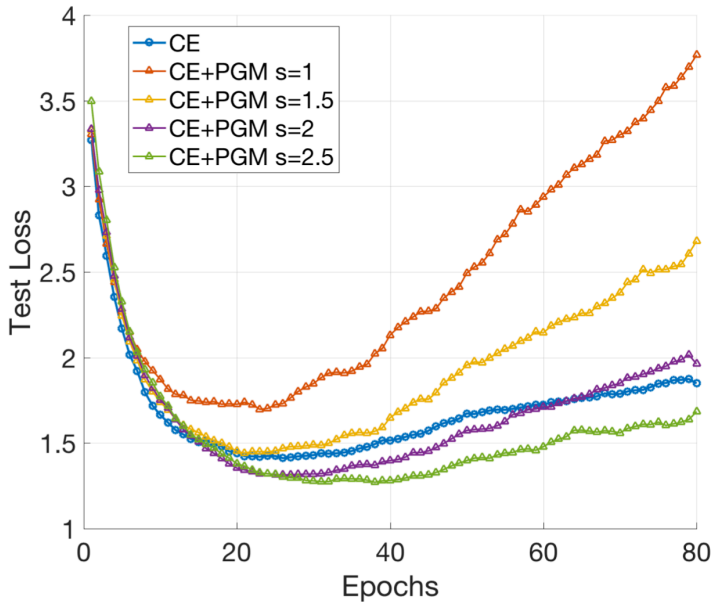
2. Each sample in a mini-batch has its own dropping mask. Similar to dropout, the gradient is averaged overall the samples, each one contributing w.r.t. non-zeroed parameters.

Maintaining the Expected Output

- To maintain the expected output consistent across training and inference, the *inverted dropout* scheme works by multiplying the masked descriptor for a scale factor $s = \frac{1}{(1-\rho)}$.
- It works well for random dropout (high and low activations are dropped with equal probability).
- In our case, we drop only the most active nodes, which can lead to an imbalance.
- So, we modify the scale factor as $s = \frac{\alpha}{(1-\rho)}$, with $\alpha \geq 1$
- Choosing the right α can be tricky, so we let the network learn the parameter α

CIFAR100

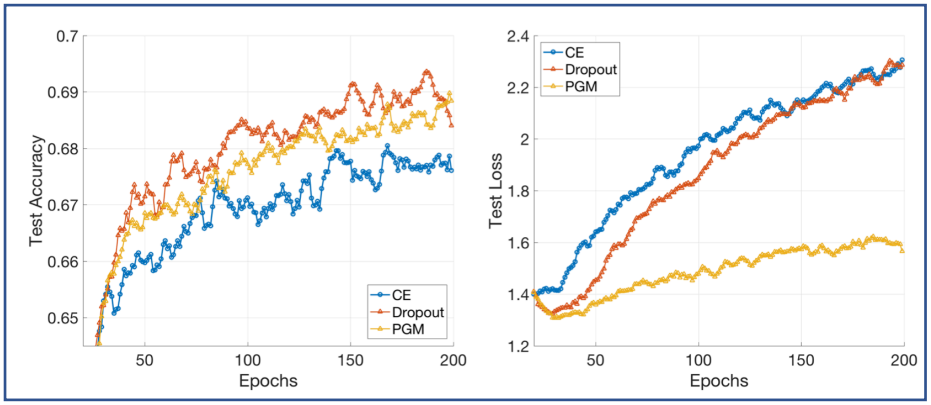


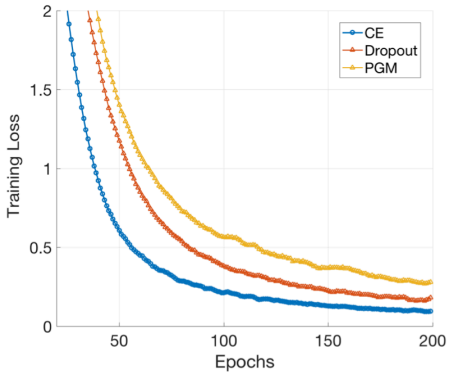


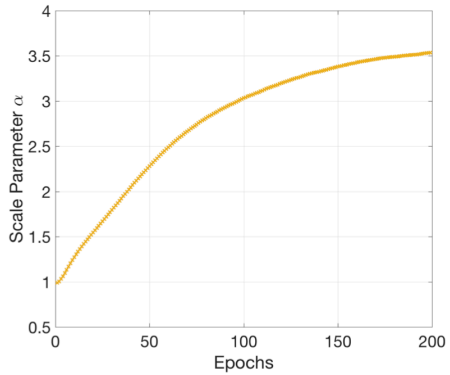
Experimental Results

- Experiments are conducted with a **ResNet-18** on **CIFAR10**, **CIFAR100**, and **Caltech256**

Effect of the learnable scale parameter (CIFAR100)



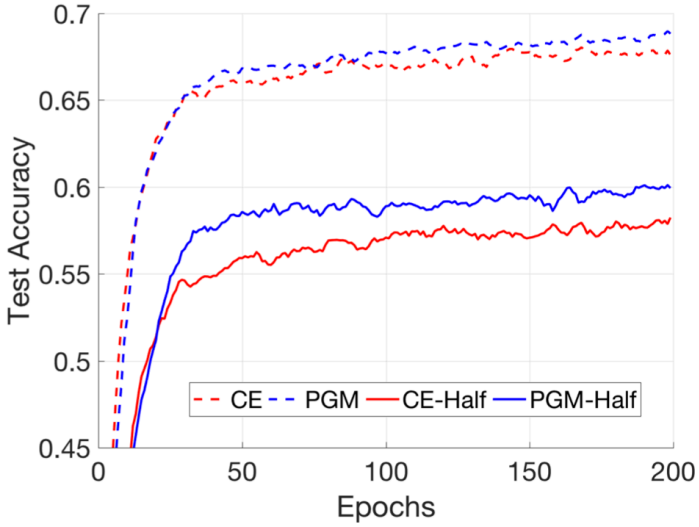




Results on Benchmark datasets

RESULTS ON THE BENCHMARK DATASETS

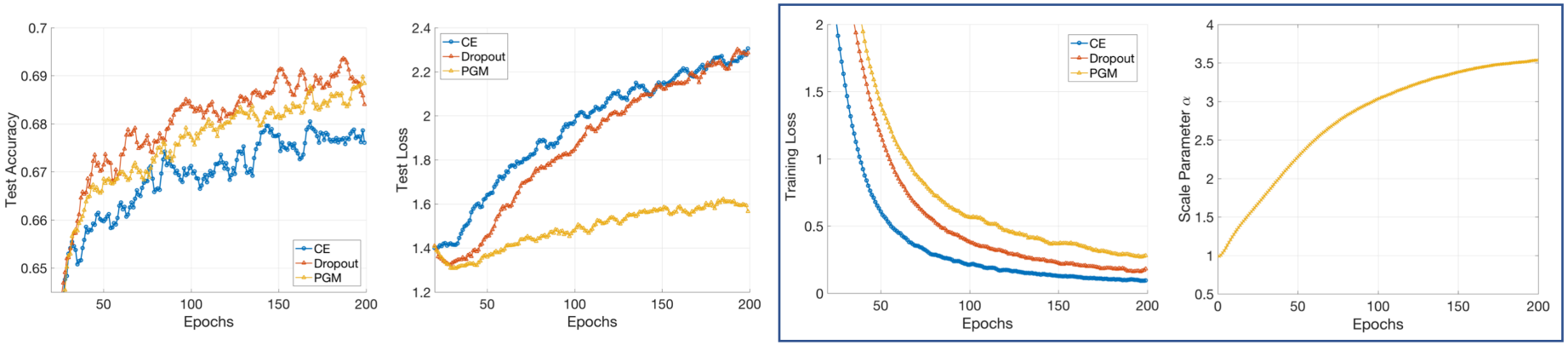
Dataset	Method	Rank-1 Acc	Test Loss
CIFAR10	CE	99.69	0.006
CIFAR10	Dropout	99.71	**0.003**
CIFAR10	PGM	**99.73**	0.02
CIFAR100	CE	68.47	2.07
CIFAR100	Dropout	**69.65**	2.21
CIFAR100	PGM	69.18	**1.52**
Caltech256	CE	62.21	2.28
Caltech256	Dropout	61.61	2.30
Caltech256	PGM	**63.24**	**1.85**



Experimental Results

- Experiments are conducted with a **ResNet-18** on **CIFAR10**, **CIFAR100**, and **Caltech256**

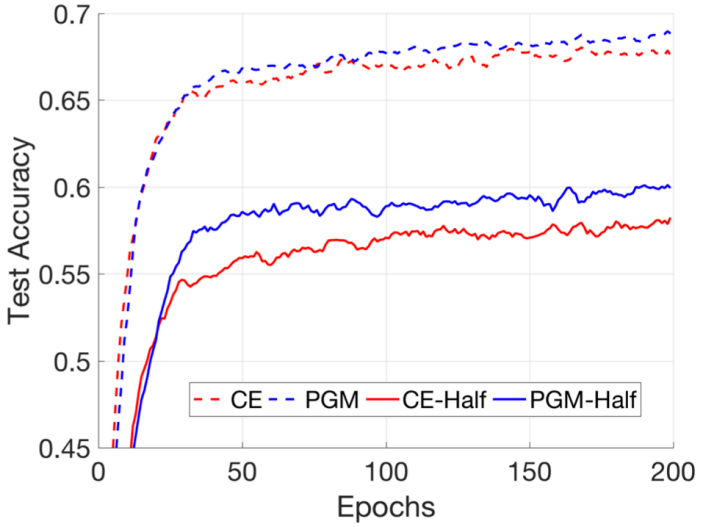
Effect of the learnable scale parameter (CIFAR100)



Results on Benchmark datasets

RESULTS ON THE BENCHMARK DATASETS

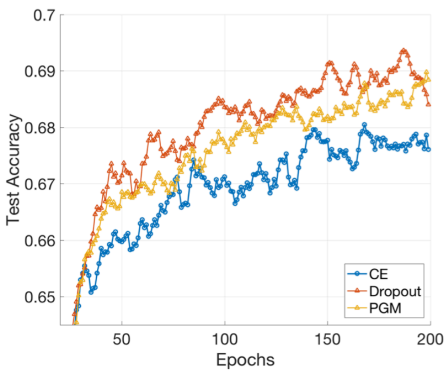
Dataset	Method	Rank-1 Acc	Test Loss
CIFAR10	CE	99.69	0.006
CIFAR10	Dropout	99.71	**0.003**
CIFAR10	PGM	**99.73**	0.02
CIFAR100	CE	68.47	2.07
CIFAR100	Dropout	**69.65**	2.21
CIFAR100	PGM	69.18	**1.52**
Caltech256	CE	62.21	2.28
Caltech256	Dropout	61.61	2.30
Caltech256	PGM	**63.24**	**1.85**

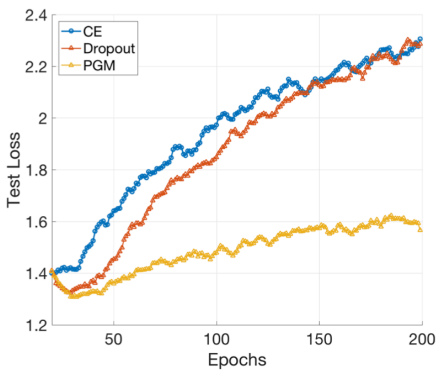


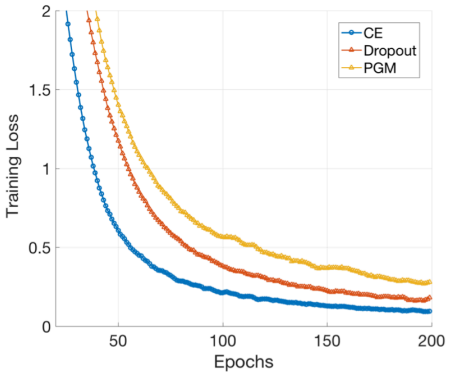
Experimental Results

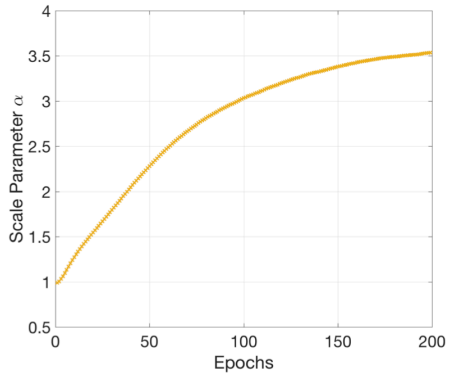
- Experiments are conducted with a **ResNet-18** on **CIFAR10**, **CIFAR100**, and **Caltech256**

Effect of the learnable scale parameter (CIFAR100)





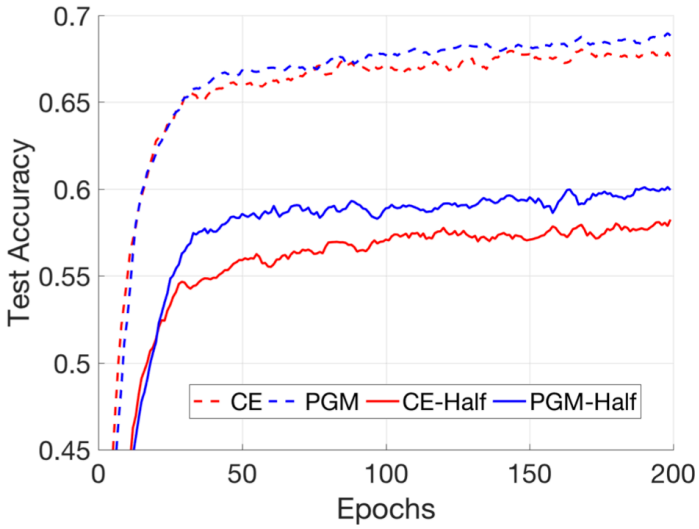




Results on Benchmark datasets

RESULTS ON THE BENCHMARK DATASETS

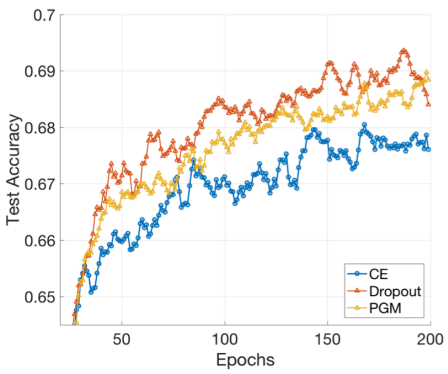
Dataset	Method	Rank-1 Acc	Test Loss
CIFAR10	CE	99.69	0.006
CIFAR10	Dropout	99.71	**0.003**
CIFAR10	PGM	**99.73**	0.02
CIFAR100	CE	68.47	2.07
CIFAR100	Dropout	**69.65**	2.21
CIFAR100	PGM	69.18	**1.52**
Caltech256	CE	62.21	2.28
Caltech256	Dropout	61.61	2.30
Caltech256	PGM	**63.24**	**1.85**

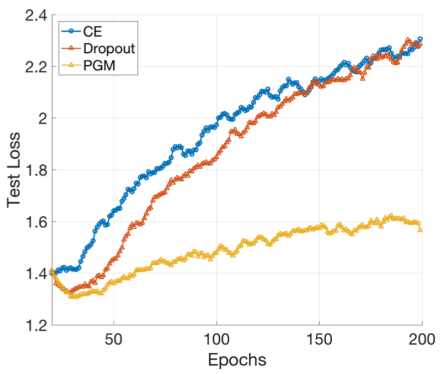


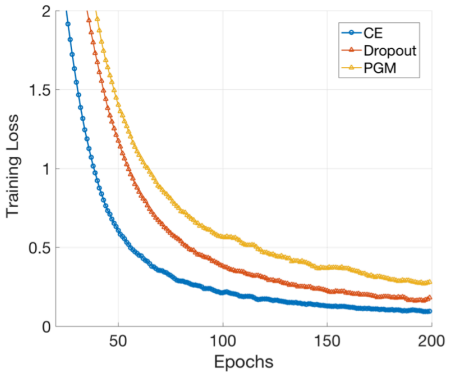
Experimental Results

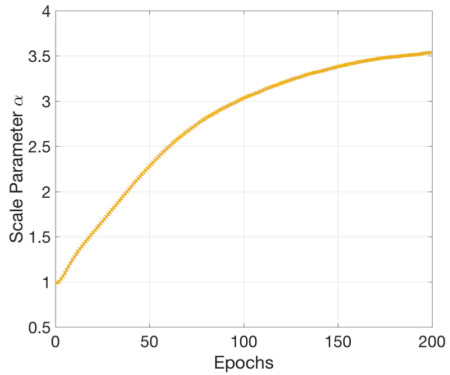
- Experiments are conducted with a **ResNet-18** on **CIFAR10**, **CIFAR100**, and **Caltech256**

Effect of the learnable scale parameter (CIFAR100)





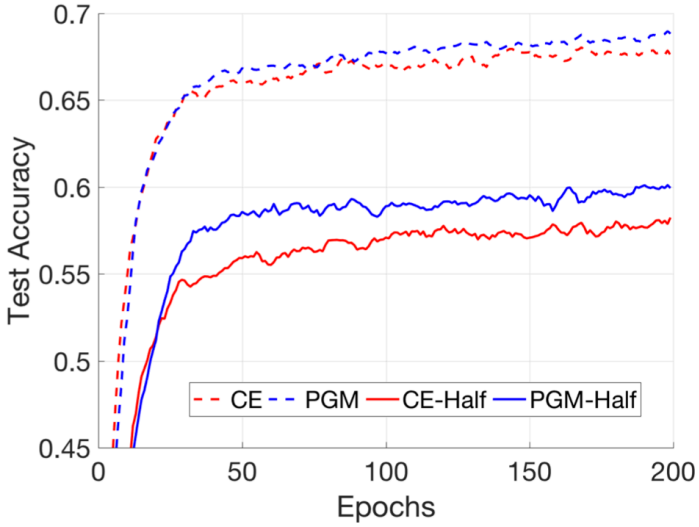




Results on Benchmark datasets

RESULTS ON THE BENCHMARK DATASETS

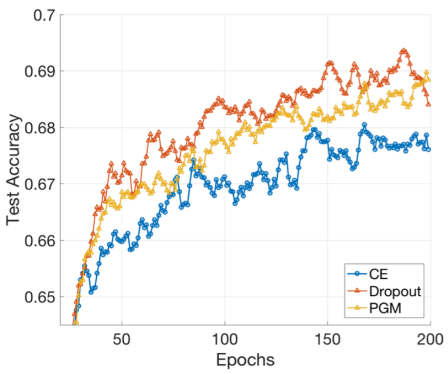
Dataset	Method	Rank-1 Acc	Test Loss
CIFAR10	CE	99.69	0.006
CIFAR10	Dropout	99.71	**0.003**
CIFAR10	PGM	**99.73**	0.02
CIFAR100	CE	68.47	2.07
CIFAR100	Dropout	**69.65**	2.21
CIFAR100	PGM	69.18	**1.52**
Caltech256	CE	62.21	2.28
Caltech256	Dropout	61.61	2.30
Caltech256	PGM	**63.24**	**1.85**

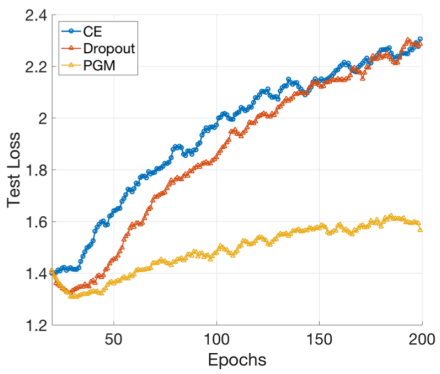


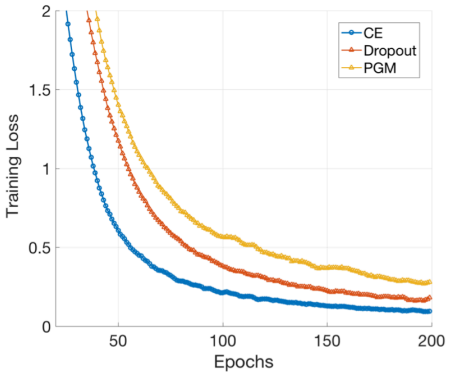
Experimental Results

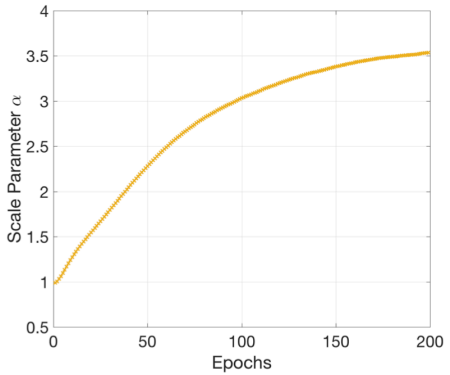
- Experiments are conducted with a **ResNet-18** on **CIFAR10**, **CIFAR100**, and **Caltech256**

Effect of the learnable scale parameter (CIFAR100)





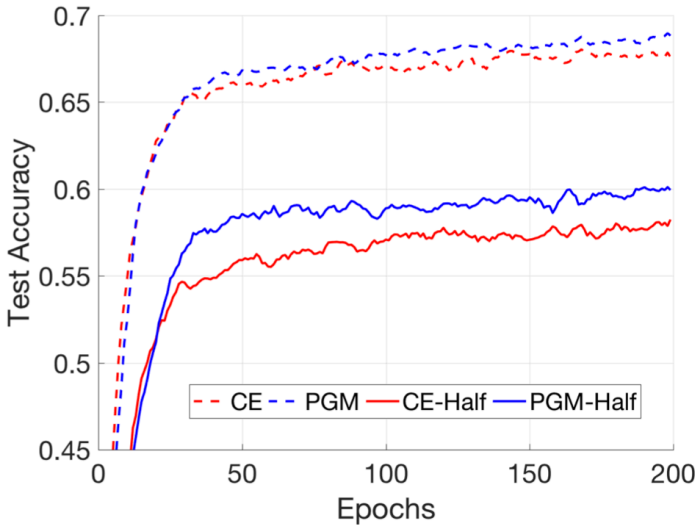




Results on Benchmark datasets

RESULTS ON THE 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**



UNIVERSITÀ
DEGLI STUDI
FIRENZE

MICC
Media Integration and Communication Center

Thank You!

Code available at:

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