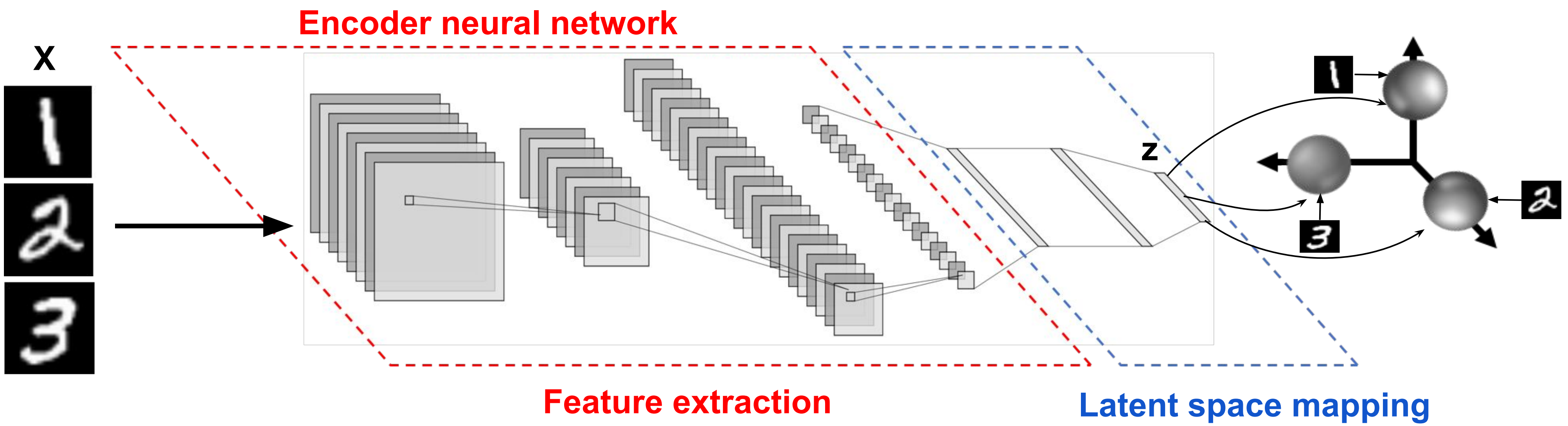
Beyond cross-entropy: learning highly separable feature distributions for robust and accurate classification

Arslan Ali, Andrea Migliorati, Tiziano Bianchi, Enrico Magli

Politecnico di Torino

GCCS - Motivation

- GCCS: Gaussian class-conditional simplex loss
- High Separability between classes
- Extract discriminative features from the input data
- Map the features to well behaved, target Gaussian distributions



GCCS - Loss

POLITECNICO DI TORINO
Image Processing and Learning group
IPL

- Assumption: network output tends to a Gaussian distribution
  - compute batch statistics for each class
  - minimize KL divergence between target and obtained distribution

Proposed loss

Authorized users loss

$$\mathcal{L}_i = \frac{1}{i}\left[\log\frac{|\Sigma_{Ti}|}{|\Sigma_{Oi}|} - D + \mathrm{tr}(\Sigma_{Ti}^{-1}\Sigma_{Oi}) + (\boldsymbol{\mu}_{Ti} - \boldsymbol{\mu}_{Oi})^\intercal \Sigma_{Ti}^{-1}(\boldsymbol{\mu}_{Ti} - \boldsymbol{\mu}_{Oi})\right]$$

$$\mathcal{K}_i = \left(\frac{x - \boldsymbol{\mu}_{Oi}}{\sigma_{Oi}}\right)^4$$

- target statistics
- batch statistics

Total loss: $\mathcal{L}^{\mathrm{GCCS}} = \sum_{i=1}^{D}\left[\mathcal{L}_i + \lambda(\mathcal{K}_i - 3)\right],$

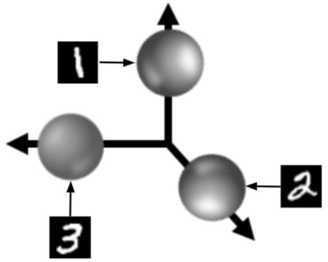
GCCS - Decision rule

- Partition the decision space into Voronoi regions
- Compute the distance from all the centers - choose minimum

$$\widehat{y} = \arg\max_{i} z_i ,$$

- Index of the predicted class for the test image

GCCS - Advantages



- Equidistant classes
- Uniformity of feature distributions - lack of short path
- Higher robustness
- Simple straightforward decision boundaries

Results: Maximum accuracy

Method	MNIST ResNet-18	FMNIST ResNet-18	SVHN ResNet-18	CIFAR-10 ResNet-18	CIFAR-10 Shake-Shake-96	CIFAR-100 Shake-Shake-112
GCCS - regular training	**99.58**	**92.69**	**94.20**	**82.97**	**96.19**	**76.53**
GCCS - fine tuning	**99.64**	**93.83**	**95.58**	**81.52**	**97.06**	**77.48**
No Defense - cross-entropy	99.35	91.91	94.12	78.59	95.78	76.30
Jacobian Reg. - regular training [60]	98.99	91.79	94.11	70.09	-	-
Jacobian Reg. - fine-tuning[60]	98.53	92.43	93.54	82.09	-	-
Input Gradient Reg. - regular training [53]	97.98	88.45	93.77	78.32	96.50	74.89
Input Gradient Reg. - fine-tuning [53]	99.11	92.55	93.17	76.15	96.90	75.68
Cross Lipschitz regular training [59]	96.78	92.54	91.42	80.10	-	-
Cross Lipschitz - fine-tuning [59]	98.77	92.41	93.50	79.39	-	-

Tab. 2 Maximum test accuracy obtained through *regular training* vs *fine-tuning* over different benchmark datasets with different competing techniques in the case in which no adversarial attack is performed.

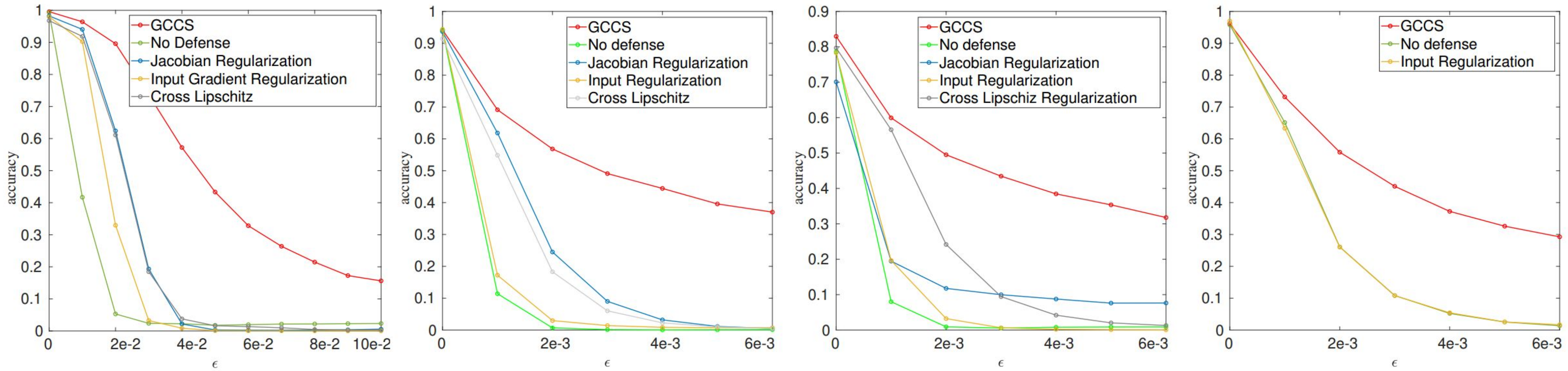
- GCCS yields high classification accuracy both for regular training and fine tuning

Adversarial attacks - Whitebox

- PGD
  - iterative FGSM
  - $x'_{t+1} = \mathrm{Proj}\{x'_t + \alpha \cdot \mathrm{sign}\left[\nabla_x J\left(\theta, x'_t, y\right)\right]\}$
- TGSM
  - descending the gradient towards target class
  - $x' = x - \epsilon \cdot \mathrm{sign}[\nabla_x J(\theta, x, y')]$
- JSMA
  - select the features to be altered to get desired output
  - $\nabla l(x) = \frac{\partial l(x)}{\partial x} = \left[\frac{\partial l_j(x)}{\partial x_\gamma}\right]_{\gamma \in 1,\ldots,M_{\mathrm{in}}, j \in 1,\ldots,M_{\mathrm{out}}}$

Robustness to targeted attacks

TGSM-5



(a) MNIST@ ResNet-18 (b) SVHN@ ResNet-18 (c) Cifar10@ ResNet-18 (d) Cifar10@ Shake-Shake

Test accuracy when applying the TGSM attack (5 steps) for (a) ([MNIST, ResNet-18]) ; (b) ([SVHN, ResNet-18]); (c) ([CIFAR-10, ResNet-18]) (d) ([CIFAR-10, Shake-Shake-96]), for different values of ϵ.

Conclusions

- Novel loss promoting class separability and robustness
- High classification accuracy
- High robustness against adversarial attacks

arslan.ali@polito.it