IEEE WCCI 2020

IEEE World Congress on Computational Intelligence

Virtual Conference – July 19-24, 2020

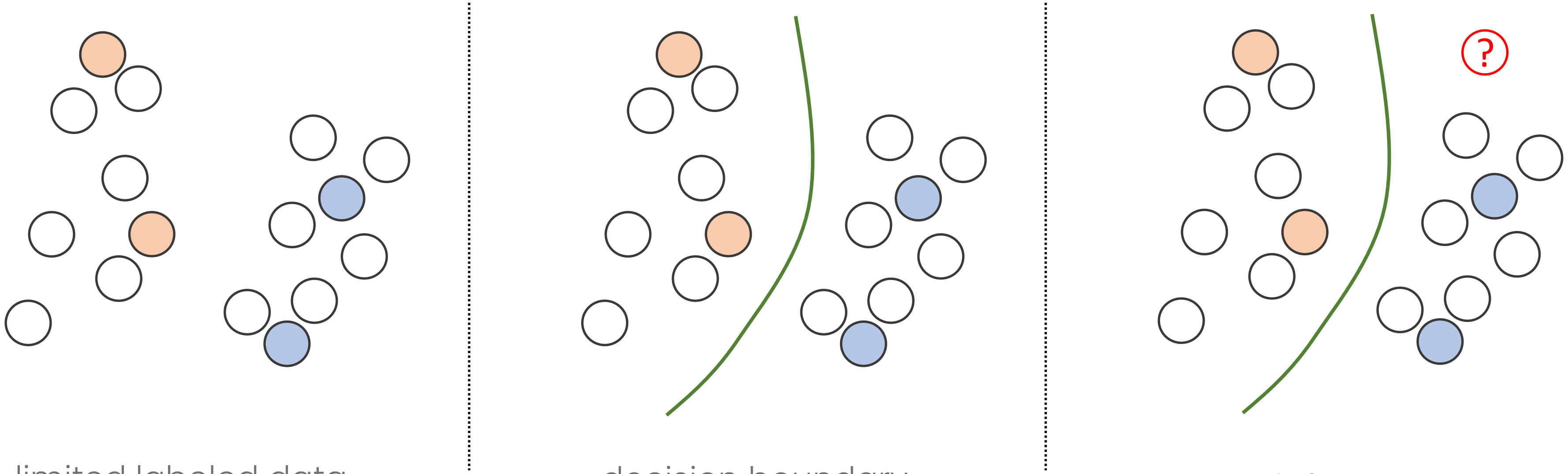
Mixing Up Real Samples and Adversarial Samples for Semi-Supervised Learning

Yun Ma*, Xudong Mao*, Yangbin Chen, Qing Li

The Hong Kong Polytechnic University 香港理工大學

CityU 香港城市大學 City University of Hong Kong

Semi-Supervised Learning (SSL)



limited labeled data + unlabeled data

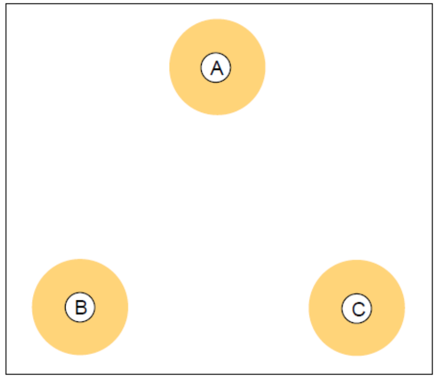
decision boundary learning

inference on query sample

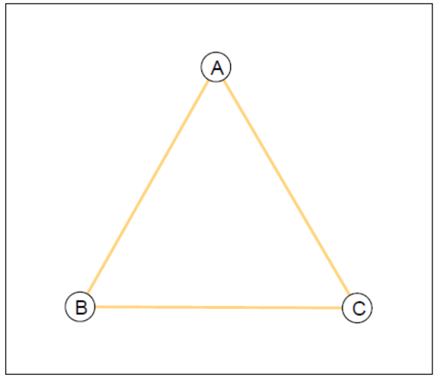
Cluster Assumption: the data distribution forms discrete clusters, and samples in the same cluster tend to share the same class label.

Consistency Regularization based SSL

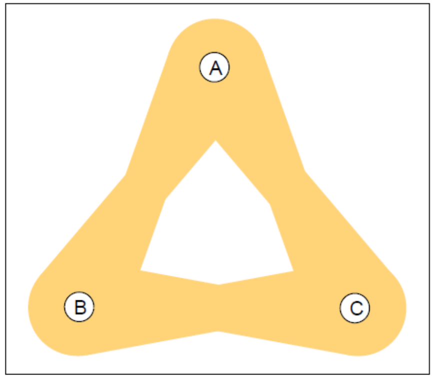
- Enforcing the model consistency between a sample x and its neighbor $\hat{x}$
- Existing Methods: how to find $\hat{x}$?
  - Local neighborhood approaches
  - In-between neighborhood approaches
- Our Motivation:
  - unifying the local neighborhood and the in-between neighborhood



local neighborhood



In-between neighborhood



our approach

Our Approach: AdvMixup

- AdvMixup: consider neighborhood formed by the samples lying along the paths between the real samples and adversarial samples.
  - Consistency Regularization

$$\hat{x}_{i,j} = \lambda x_i + (1-\lambda) x_j^{(adv)},$$
$$\hat{y}_{i,j} = \lambda f_t(x_i) + (1-\lambda) f_t(x_j),$$

$$\boxed{\mathcal{L}_{\text{reg}} = \mathbb{E}_{x_i \sim \mathcal{S}_u, x_j \sim \mathcal{S}_u} \left[D_{\mathcal{Y}}[f(\hat{x}_{i,j}), \hat{y}_{i,j}] \right]}$$

  - Adversarial Sample Generation

$$x_j^{(adv)} = x_j + r_j^{(adv)}$$

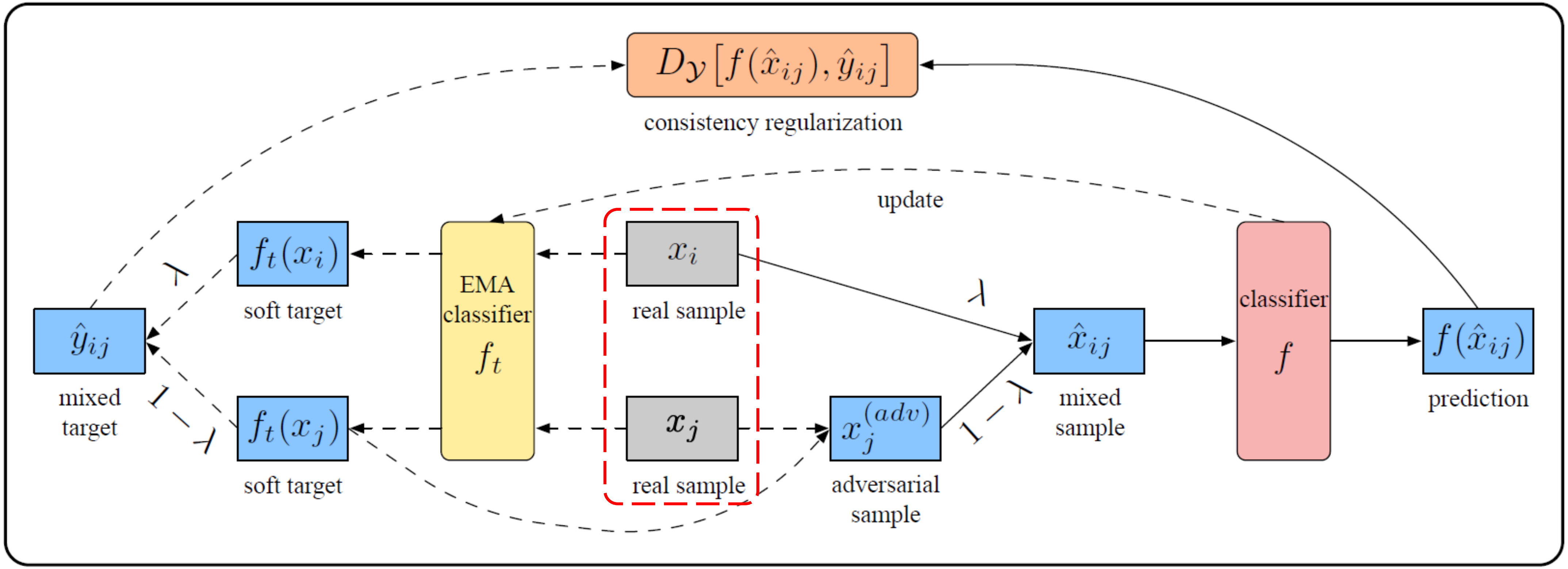
$$r_j^{(adv)} = \underset{\|r\|_2 \le \epsilon}{\arg\max}\, D_{\mathcal{Y}}\left[f_t(x_j), f(x_j + r) \right]$$ [Miyato et al. TPAMI 2018]

  - Loss Function

$$\boxed{\mathcal{L}_{\text{nll}}} + \beta \mathcal{L}_{\text{reg}}$$

$$\mathcal{L}_{\text{nll}} = \mathbb{E}_{(x_i, y_i) \sim \mathcal{S}_l} \left[-y_i^\top \ln f(x_i) \right]$$

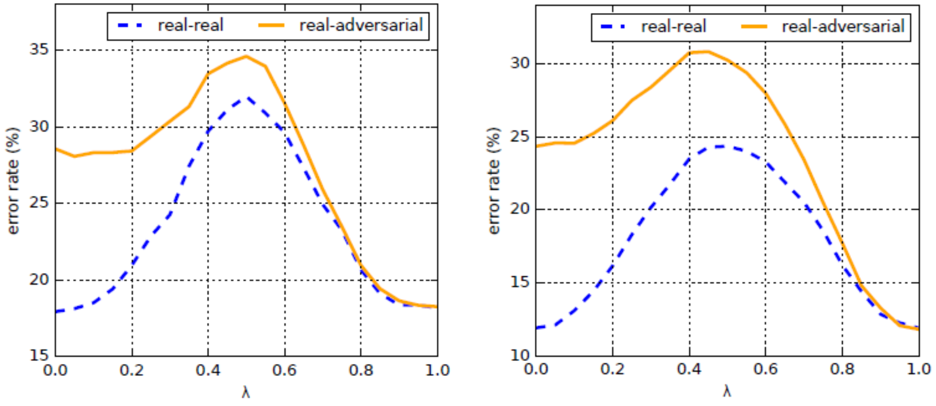
Our Approach: AdvMixup



Advantages of AdvMixup

Consistency regularization approaches are actually fixing the classifier's flaws which violate the cluster assumption.

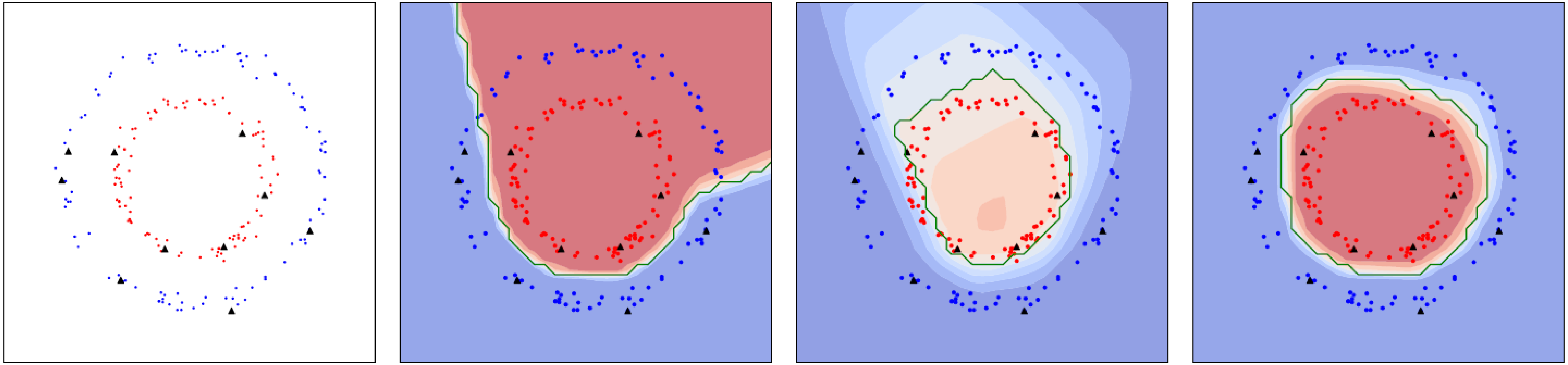
- Compared with local neighborhood approaches: we search the flaws in a more comprehensive area
- Compared with in-between neighborhood approaches: we search the flaws which violates the assumption more significantly



(a) CIFAR-10 (b) SVHN

Prediction error rates of the supervised model on the virtual samples along the real-real interpolation paths defined by the in-between neighborhood based ICT model and the real-adversarial interpolation paths defined by the proposed AdvMixup.

Case Study on Synthetic Data



synthetic data: Two Circles | local neighborhood approach: VAT [Miyato et al. TPAMI 2018] | In-between neighborhood approach: ICT [Verma et al. IJCAI 2019] | our approach: AdvMixup

The proposed AdvMixup can (1) successfully separate the two classes and (2) learn a decision boundary in low-density regions

Experiments on Benchmark Datasets

Method	Test error rates (%)		
	1000 labels	2000 labels	4000 labels
Supervised	39.95 ± 0.75	31.16 ± 0.66	21.75 ± 0.46
Π model [6]	31.65 ± 1.20	17.57 ± 0.44	12.36 ± 0.31
TempEns [6]	23.31 ± 1.01	15.64 ± 0.39	12.16 ± 0.24
MT [7]	21.55 ± 1.48	15.73 ± 0.31	12.31 ± 0.28
VAT [8]	–	–	11.36 ± 0.34
VAT+EntMin [8]	–	–	10.55 ± 0.05
VAdD [23]	–	–	11.32 ± 0.11
VAdD + VAT [23]	–	–	9.22 ± 0.10
TempEns+SNTG [15]	18.41 ± 0.52	13.64 ± 0.32	10.93 ± 0.14
VAT+EntMin+SNTG [15]	–	–	9.89 ± 0.34
CT-GAN [13]	–	–	9.98 ± 0.21
CVT [24]	–	–	10.11 ± 0.15
MT+ fast-SWA [25]	15.58 ± 0.12	11.02 ± 0.23	9.05 ± 0.21
ICT [9]	15.48 ± 0.78	9.26 ± 0.09	7.29 ± 0.02
AdvMixup	$\mathbf{9.67 \pm 0.08}$	$\mathbf{8.04 \pm 0.12}$	$\mathbf{7.13 \pm 0.08}$

Results on CIFAR-10 (consisting of 50000 training samples) with 1000, 2000, and 4000 labeled samples

Method	Test error rates (%)		
	250 labels	500 labels	1000 labels
Supervised	40.62 ± 0.95	22.93 ± 0.67	15.54 ± 0.61
Π model [6]	9.93 ± 1.15	6.65 ± 0.53	4.82 ± 0.17
TempEns [6]	12.62 ± 2.91	5.12 ± 0.13	4.42 ± 0.16
MT [7]	4.35 ± 0.50	4.18 ± 0.27	3.95 ± 0.19
VAT [8]	–	–	5.42 ± 0.22
VAT+EntMin [8]	–	–	3.86 ± 0.11
VAdD [23]	–	–	4.16 ± 0.08
VAdD + VAT [23]	–	–	3.55 ± 0.05
Π+SNTG [15]	5.07 ± 0.25	4.52 ± 0.30	3.82 ± 0.25
MT+SNTG [15]	4.29 ± 0.23	3.99 ± 0.24	3.86 ± 0.27
ICT [9]	4.78 ± 0.68	4.23 ± 0.15	3.89 ± 0.04
AdvMixup	$\mathbf{3.95 \pm 0.70}$	$\mathbf{3.37 \pm 0.09}$	$\mathbf{3.07 \pm 0.18}$

Results on SVHN (consisting of 73257 training samples) with 250, 500, and 1000 labeled samples

Robustness Analysis

- Attack the models with adversarial samples crafted with the Fast Gradient Method [Goodfellow et al. ICLR 2015]

Method	CIFAR-10 $\epsilon_w = 1.0$	CIFAR-10 $\epsilon_w = 2.0$	CIFAR-10 $\epsilon_w = 3.0$	CIFAR-10 $\epsilon_w = 5.0$	CIFAR-10 $\epsilon_w = 8.0$	SVHN $\epsilon_w = 0.1$	SVHN $\epsilon_w = 0.5$	SVHN $\epsilon_w = 1.0$	SVHN $\epsilon_w = 2.0$	SVHN $\epsilon_w = 3.0$
Supervised	58.50	77.73	86.73	94.2	96.91	19.81	51.71	69.94	82.28	86.46
ICT [9]	24.77	43.28	56.24	69.42	78.38	7.72	28.57	41.87	52.35	58.00
AdvMixup	**17.40**	**30.91**	**42.52**	**58.59**	**70.82**	**5.11**	**14.59**	**24.39**	**37.84**	**47.63**

white-box attacks

Method	CIFAR-10 $\epsilon_b = 1.0$	CIFAR-10 $\epsilon_b = 2.0$	CIFAR-10 $\epsilon_b = 3.0$	CIFAR-10 $\epsilon_b = 5.0$	CIFAR-10 $\epsilon_b = 8.0$	SVHN $\epsilon_b = 0.1$	SVHN $\epsilon_b = 0.5$	SVHN $\epsilon_b = 1.0$	SVHN $\epsilon_b = 2.0$	SVHN $\epsilon_b = 3.0$
Supervised	29.25	39.38	48.83	63.06	75.75	14.37	24.76	36.92	52.91	62.05
ICT [9]	9.78	12.68	16.03	24.85	37.83	4.19	8.17	15.59	30.43	41.29
AdvMixup	**8.62**	**10.17**	**12.34**	**17.34**	**25.77**	**3.47**	**6.62**	**12.31**	**24.92**	**35.39**

black-box attacks

The integration of local neighborhood with in-between neighborhood gives AdvMixup an edge in robustness against adversarial perturbations.

Conclusion

- We propose a new consistency regularization approach for SSL, AdvMixup, by enforcing the model to fit virtual data points on the interpolation paths between training samples and adversarial samples.
- By unifying the local neighborhood and in-between neighborhood, AdvMixup outperforms existing methods on both synthetic data and benchmark datasets. Moreover, AdvMixup achieves better robustness against both white-box and black-box attacks with adversarial samples.
- Limitation: computational overhead brought by the adversarial sample generation
- Future work: evaluate AdvMixup with different adversarial sample generation strategies, study the trade-off between model efficiency and classification performance.

IEEE WCCI 2020

IEEE World Congress on Computational Intelligence

Virtual Conference – July 19-24, 2020

Thank you !

SPONSORS:

IEEE Advancing Technology for Humanity

IEEE Computational Intelligence Society

IET

The International Neural Network Society (INNS)

EPS Evolutionary Programming Society