WCP: Worst-Case Perturbations for Semi-supervised Deep Learning

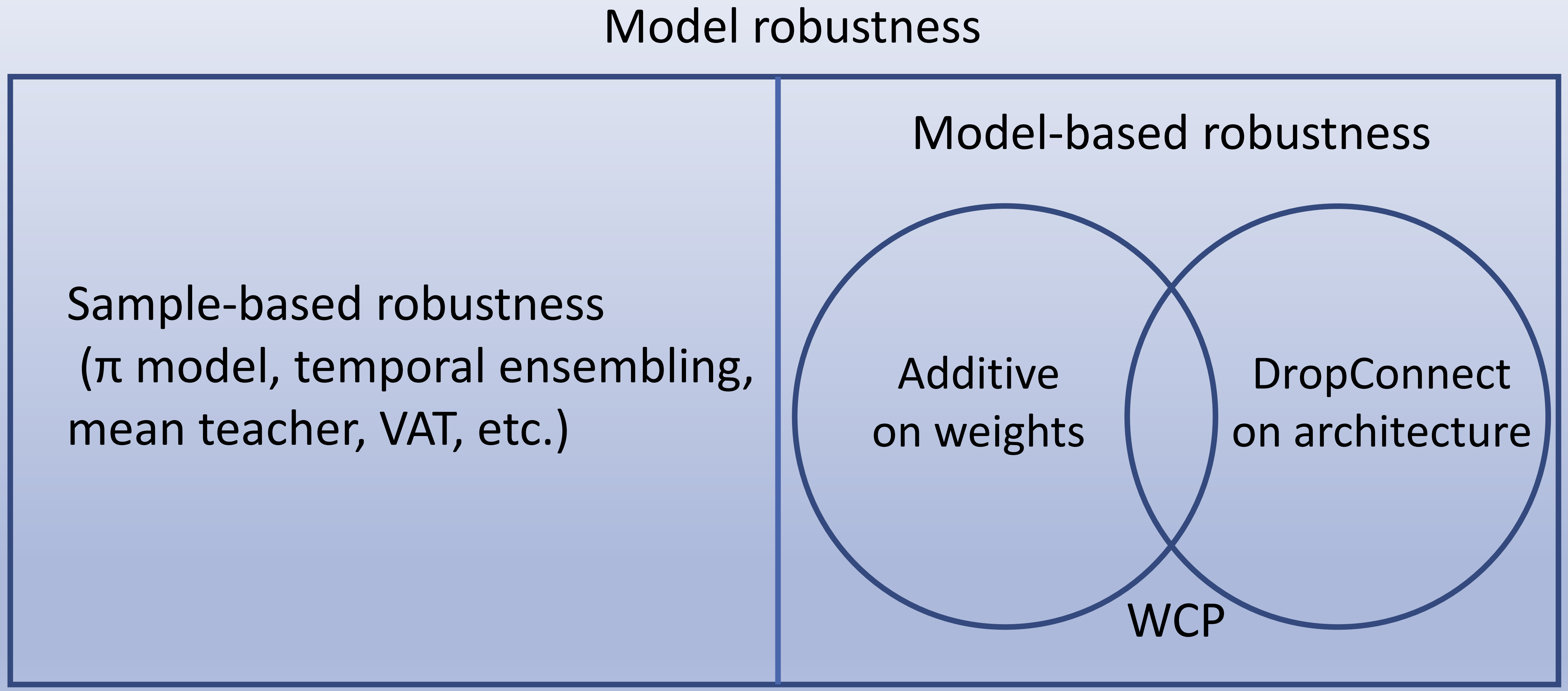
Liheng Zhang[1] Guo-Jun Qi[1, 2]

[1]Laboratory of Machine Perception and Learning (MAPLE)
[2]Futurewei Technologies

Outline

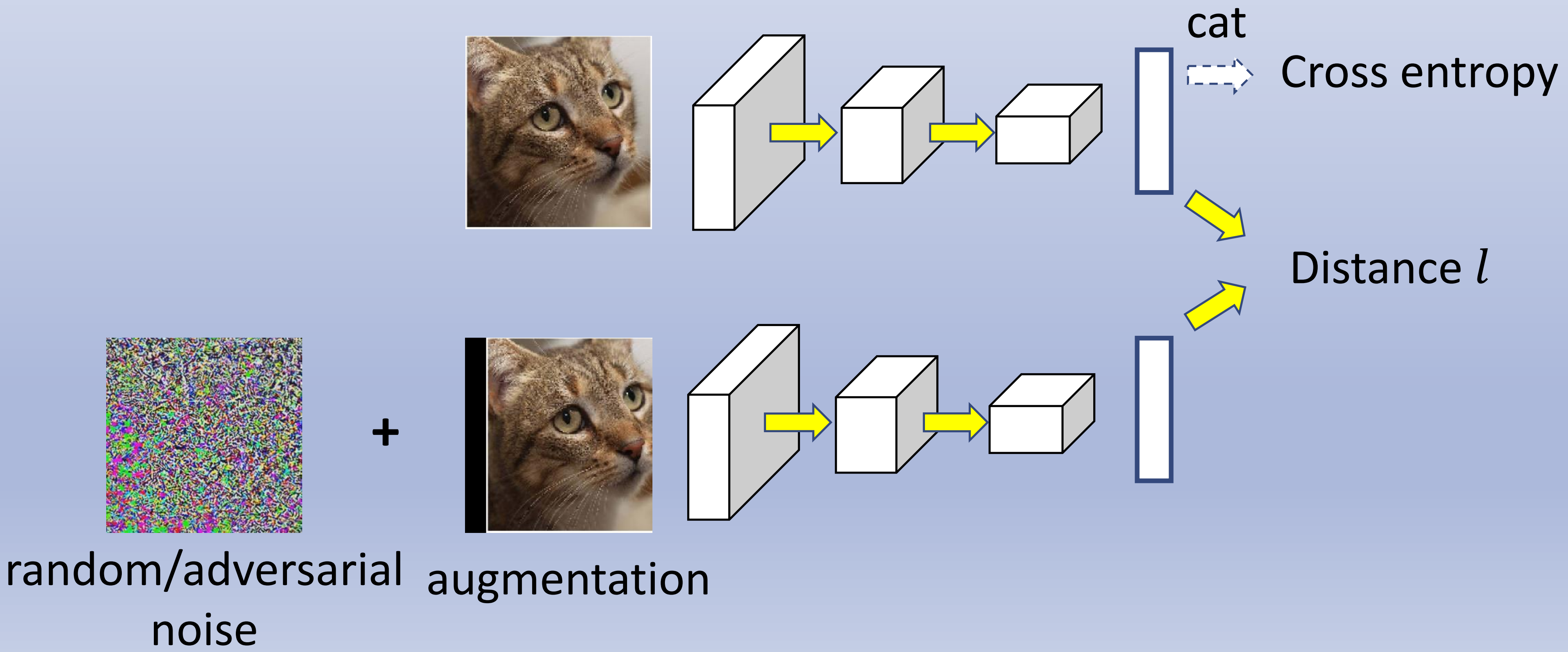
- Previous methods: Sample-based robustness
- Worst-Case Perturbations: Model-based robustness
  - ➢Additive perturbations on model weights
  - ➢DropConnect Perturbations on model architecture
- Experiments
- Conclusion

Model-based robustness vs. Sample-based robustness



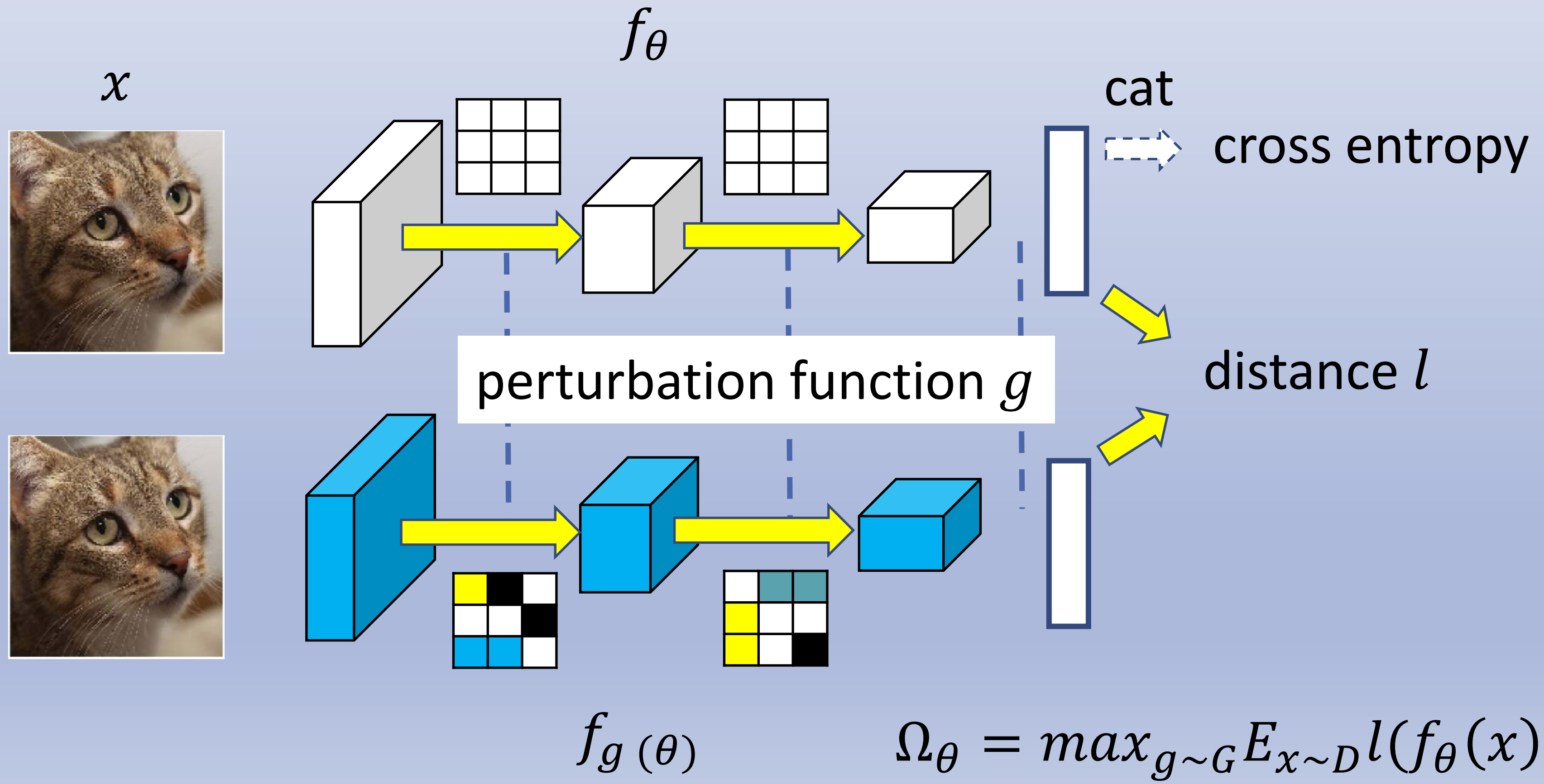
Previous methods: Sample-based robustness

- Explore unlabeled data via label invariance against perturbations on data
  - Augmentation
  - Random noise (e.g. π model, temporal ensembling, mean teacher, etc.)
  - Adversarial noise (e.g. VAT)



Worst-Case Perturbations: Model-based robustness

- Model-based robustness: Invariance against perturbations on model
  - ***Worst*** perturbations on model weights (***Additive perturbations***)
  - ***Worst*** perturbations on model architecture (***DropConnect perturbations***)

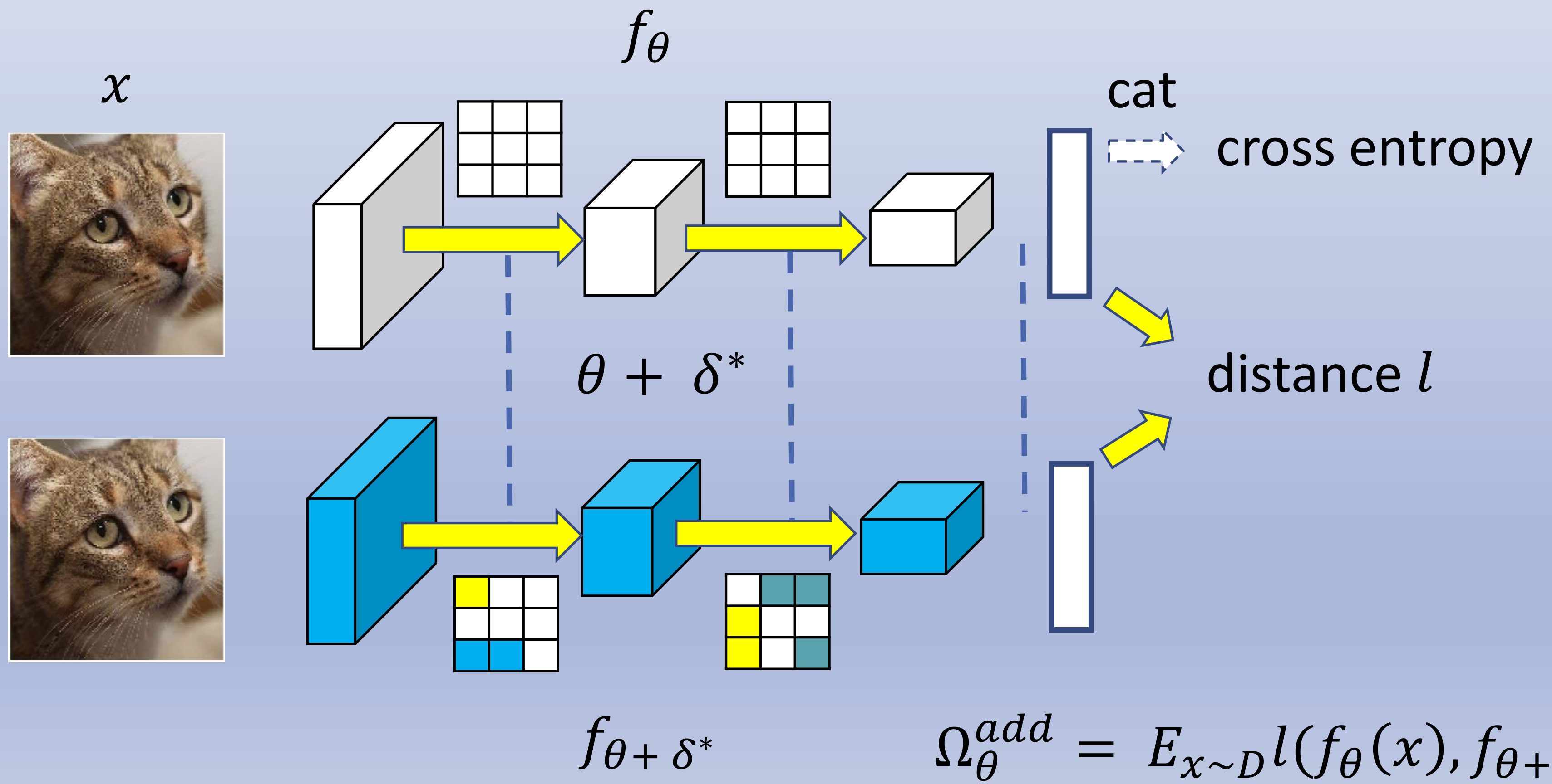


$$\Omega_\theta = max_{g \sim G} E_{x \sim D}\, l(f_\theta(x), f_{g(\theta)}(x))$$

Perturbations on model weights

- Additive perturbation

$$g(\theta) = \theta + \delta, \text{ with noise } ||\delta|| < \epsilon$$



$$\Omega_\theta^{add} = E_{x\sim D}\, l(f_\theta(x), f_{\theta+\delta^*}(x))$$

Derivation of δ^*

We assume:

- $l(y, z) = 0$ when $y = z$
- $l(y, z) \geq 0$, i.e., its minimal value is zero
- $l(y, z)$ is at least twice differentiable

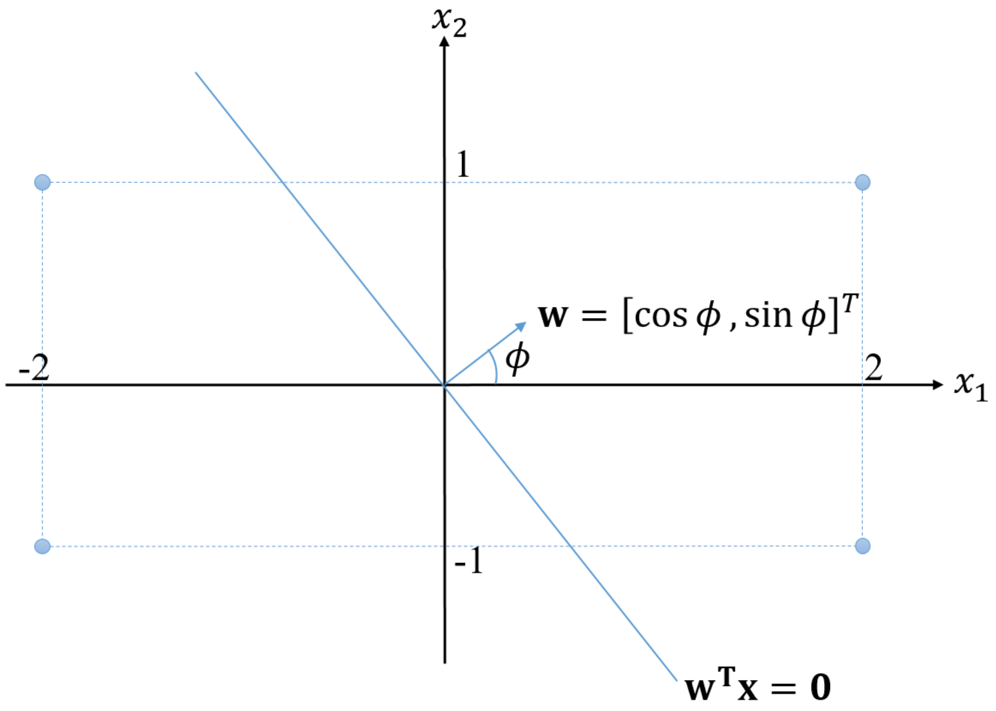
- Taking the Taylor expansion

$$\Omega_\theta^{add} = max_{||\delta||<\epsilon}\, E_{x\sim D}\, l\big(f_\theta(x), f_{\theta+\delta}(x)\big)$$
$$\approx max_{||\delta||<\epsilon}\, E_{x\sim D}\, \frac{1}{2}\delta^T S_\theta \delta$$

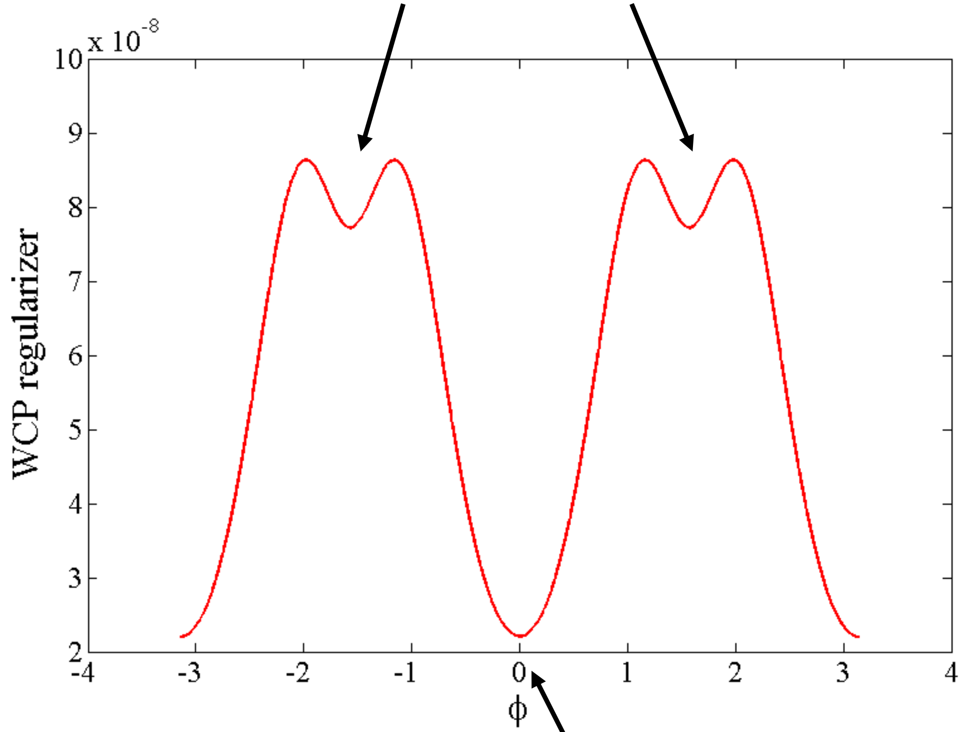
where $S_\theta = E_{x\sim D} \nabla^2 l\big(f_\theta(x), f_{\theta+\delta}(x)\big)|_{\delta=0}$

- Optimal $\delta^* = \epsilon u_\theta$, where u_θ is the singular vector of the largest singular value of S_θ, it can be efficiently computed by power iteration.

A sigmoid example: connection with max margin



Local minima at ϕ = $\pm\frac{\pi}{2}$ (x_2 = 0)

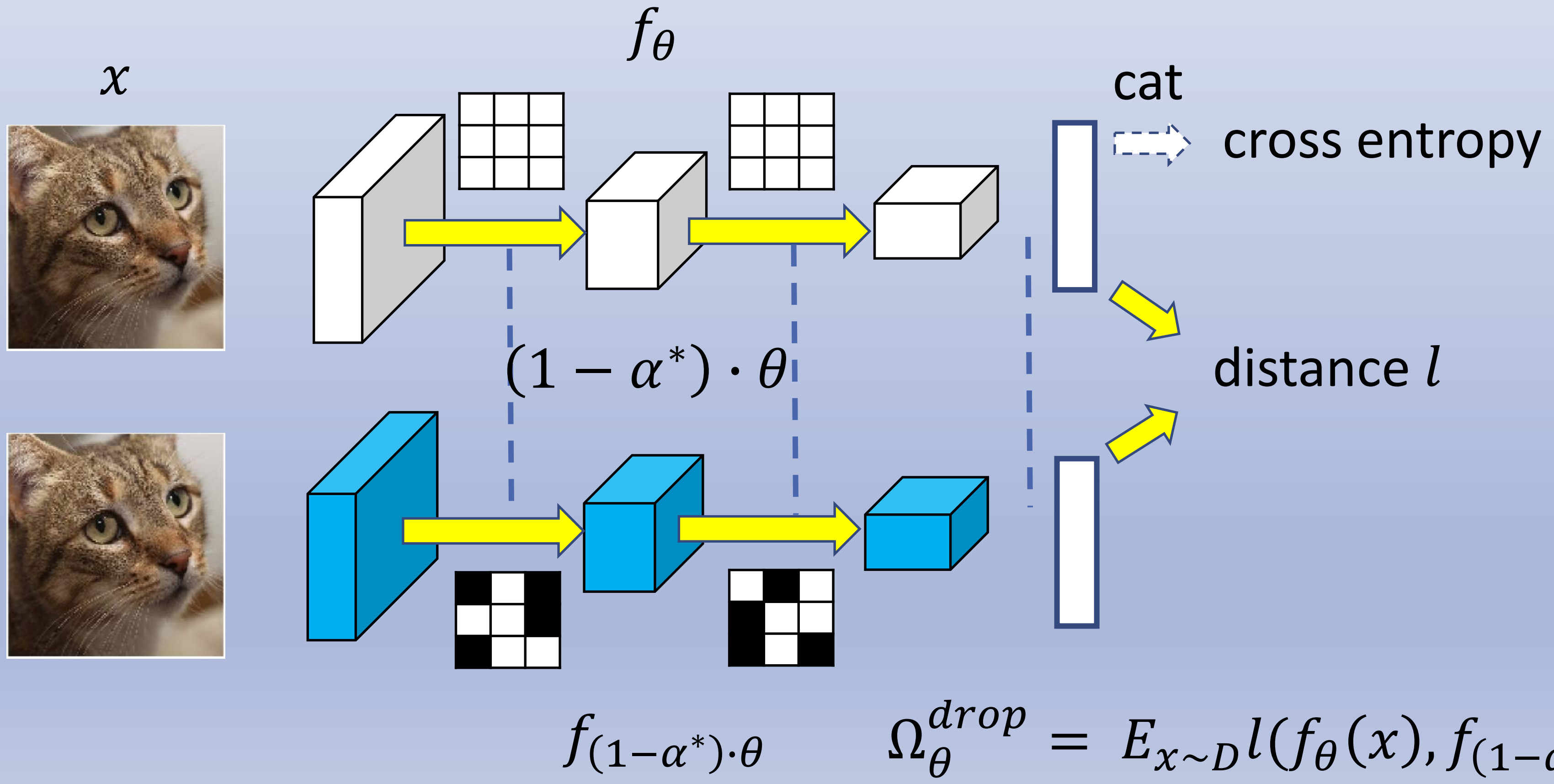


minima at ϕ = 0 (x_1 = 0)

Perturbations on model architecture

- DropConnect perturbation

$$g(\theta) = (1 - \alpha) \cdot \theta, \text{ with } G_\alpha = \{\alpha | \alpha \in \{0,1\}^N, ||\alpha||_0 = [\sigma N]\}$$



$$\Omega_\theta^{drop} = E_{x\sim D} l(f_\theta(x), f_{(1-\alpha^*)\cdot\theta}(x))$$

Derivation of α^*

- Taking the Taylor expansion

$$\alpha^* = argmax_{\alpha \in G_\alpha} E_{x \sim D}\, l\left(f_\theta(x), f_{(1-\alpha)\cdot\theta}\,(x)\right)$$
$$\approx argmax_{\alpha \in G_\alpha} \frac{1}{2} \alpha^T Q_\theta \alpha$$

where $Q_\theta = E_{x \sim D} \nabla^2 l\left(f_\theta(x), f_{(1-\alpha)\cdot\theta}(x)\right) |_{\alpha=0}$

- We can get α^* through solving the constraint Binary Quadratic Programming (BQP) problem by spectral method. (See details in the paper.)

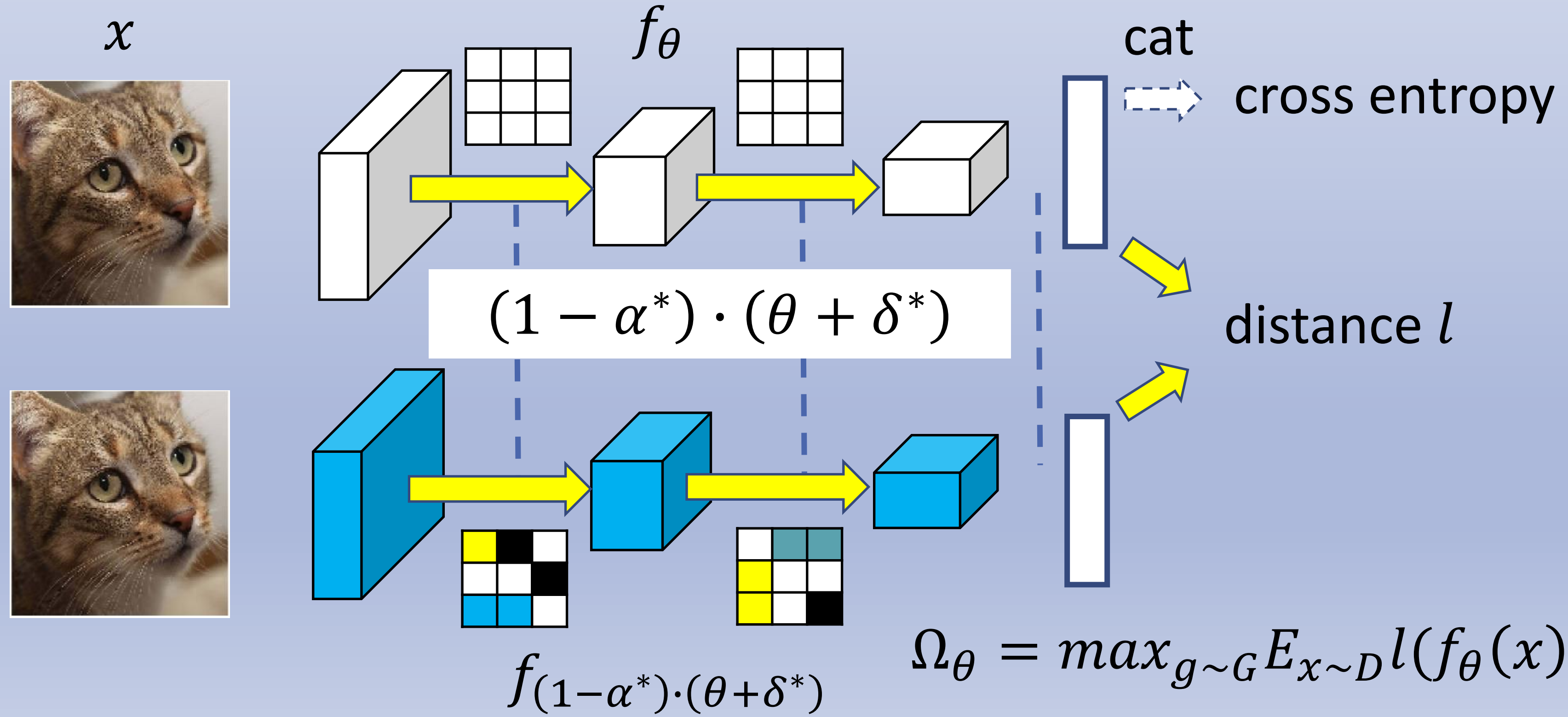
Integrating Additive and DropConnect

- Perturbation function

$$g(\theta) = (1 - \boxed{\alpha^*}) \cdot (\theta + \boxed{\delta^*})$$

WCP on model architecture WCP on model weights

- Semi-supervised objective: $min_\theta E_{(x,y)\sim T} \varepsilon_\theta(x,y) + \gamma\Omega_\theta$



$$\Omega_\theta = max_{g\sim G} E_{x\sim D}\, l(f_\theta(x), f_{(1-\alpha^*)\cdot(\theta+\delta^*)}(x))$$

CIFAR-10 Experiments

Error rate over 10 runs with the same 13-layer architecture

Method	1000 labels	2000 labels	4000 labels
GAN			18.63 ± 2.32
π model			12.36 ± 0.31
Temporal Ensembling			12.16 ± 0.31
VAT			11.36
VAT+EntMin			10.55
Supervised-only	46.43 ± 1.21	33.96± 0.73	20.66 ± 0.57
π model	27.26 ± 1.20	18.02 ± 0.60	13.20 ± 0.27
Mean Teacher	21.55 ± 1.48	15.73 ± 0.31	12.31 ± 0.28
The proposed WCP	**17.62 ± 1.52**	**11.93 ± 0.39**	**9.72 ± 0.31**

SVHN Experiments

Error rate over 10 runs with the same 13-layer architecture

Method	250 labels	500 labels	1000 labels
GAN		18.44 ± 4.8	8.11 ± 11.3
π model		6.65 ± 0.53	4.82 ± 0.17
Temporal Ensembling		5.12 ± 0.13	4.42 ± 0.16
VAT			5.42
VAT+EntMin			3.86
Supervised-only	27.77 ± 3.18	16.88± 1.30	12.32 ± 0.95
π model	9.69 ± 0.92	6.83 ± 0.66	4.95 ± 0.26
Mean Teacher	4.35 ± 0.50	4.18 ± 0.27	3.95 ± 0.19
The proposed WCP	**4.29 ± 0.10**	**3.75 ± 0.11**	**3.58 ± 0.186**

Ablation Study

Impact of different model components (CIFAR-10 with 4000 labels)

Components			
Additive Perturbation	√	√	√
DropConnect Perturbation		√	√
Entropy Minimization (EntMin)			√
Error rate	10.15	9.85	**9.51**

DropConnect on different layers

DropConnect	**Error rate**
1^{st} layer	9.77
2^{nd} layer	**9.51**
3^{rd} layer	10.08

DropConnect ratio

ratio	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.7**
Error rate	9.81	**9.51**	9.66	9.78	9.92	10.26

Conclusion

- Model-based robustness vs. sample-based robustness

  WCP previous methods

- We introduce two forms of WCP regularizations:
  - ***Additive*** perturbations on model weights
  - ***DropConnect*** perturbations on model architecture

- Experiments demonstrate the WCP outperforms many state-of-the-art models in literature.

Thanks!

Code is released at:
https://github.com/maple-research-lab/WCP