Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness

Fanny Yang[†,°], **Zuowen Wang**[°], **Christina Heinze-Deml**[⋆]

[†]Stanford; [⋆]Seminar for Statistics, [°]Foundations of Data Science, ETH Zürich

Problem we aim to solve

- Make neural networks invariant against perturbations $T(X)$
- In this work: small translations ($\pm 3px$) and rotations ($\pm 30°$)

- Invariance evaluated using the adversarial loss (in practice: grid search over 775 values)

$$\min_{f\in\mathcal{F}} \mathbb{E}_{X,Y} \sup_{x'\in T(X)} \ell(f(x'), Y). \tag{1}$$

- Also care about standard accuracy (std)

$$\min_{f\in\mathcal{F}} \mathbb{E}_{X,Y} \ell(f(X), Y) \tag{2}$$

Algorithm: Regularizing standard and adversarial augmentation

Gradient update on a minibatch:

Usually:
$$\theta^{t+1} = \theta^t - \gamma_t \sum [\nabla \ell(x, y; \theta^t) + \nabla \ell(\tilde{x}, y; \theta^t)$$

Invariance-regularizer:
$$+ \boldsymbol{\lambda \nabla h(f_{\theta^t}(x), f_{\theta^t}(\tilde{x}))}]$$

Choices of augmented $\tilde{x}$:

- $\tilde{x} \sim \text{Unif}(T(X))$: data augmentation (D.A.)
- $\arg\max_{x'\in T(x)} \ell(x', y; \theta^t)$: adversarial training (A.T.) (e.g. [1] for ℓ_2)
- $\arg\max_{x'\in T(x)} h(f_{\theta^t}(x), f_{\theta^t}(x'))$: modified adv. training (e.g. [2] for KL)

Practical benefits of regularization

High-level take-aways

- Regularization on top of augmentation always increases robustness
- at almost no computational overhead
- Augmentation-based training of vanilla network outperform selected handcrafted networks

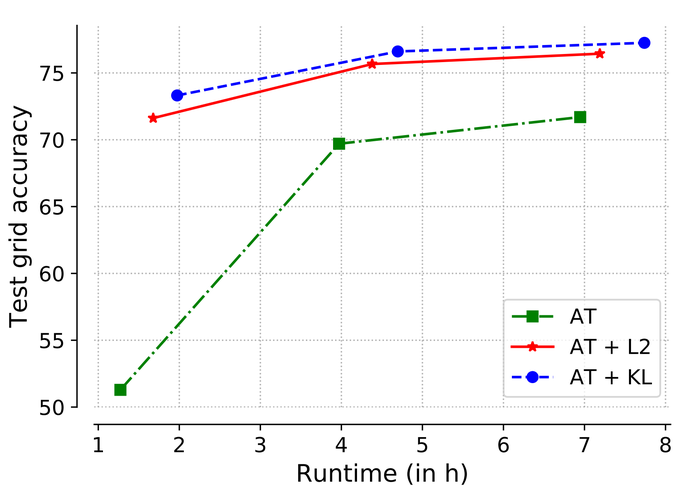
Unregularized vs. regularized augmented training for NN

- Here we choose KL-div. and ℓ_2-dist. as semimetric h
- $f_\theta(x)$ are logits (ℓ_2) and post-softmax activations (KL)
- GRN: G-Resnet44 [3] has convolutions with multiple rotated filters

	std	D.A.	D.A. + KL	A.T.	A.T. + KL	A.T. + L2	GRN
SVHN (std)	95.48	93.97	96.16	96.03	96.14	96.53	95.0
(grid)	18.85	82.60	90.69	90.35	92.42	**92.55**	84.90
CIFAR-10 (std)	92.11	89.93	89.19	91.78	89.82	90.53	93.08
(grid)	9.52	58.29	73.32	70.97	**78.79**	77.06	71.64
CIFAR-100 (std)	70.23	66.62	68.69	68.79	69.99	67.11	–
(grid)	5.09	28.53	48.69	38.21	**53.70**	50.82	–

Computational comparison

- Mean runtime for different methods on CIFAR-10
- Points with increasing runtime use worst-of-k defense, $k \in \{1, 10, 20\}$



Experimental details

Solving the maximization for $\tilde{x}$ via

- **random**: A uniformly drawn transformation
- **worst-of-k**: Worst of k randomly drawn transformation [4]
- **S-PGD**: Projected Gradient Descent with respect to the transformation parameters

Theoretical framework

- **Regularized algorithm** with $\tilde{x} = \arg\max_{x'\in G^x} h(f_{\theta^t}(x), f_{\theta^t}(x'))$ and

$$\theta^{t+1} = \theta^t - \gamma_t \mathbb{E}\left[\nabla \ell(x, y; \theta^t) + \lambda \nabla h(f_{\theta^t}(\tilde{x}), f_{\theta^t}(x))\right]$$

is a first-order method for minimizing the **penalized loss**

$$\min_{f\in\mathcal{F}} \mathbb{E}\, \ell(f(X), Y) + \lambda \sup_{\tilde{x}\in G^X} h(f(\tilde{x}), f(x)).$$

- For some λ, this is the dual of the **constrained problem**

$$\min_{f\in\mathcal{F}} \mathbb{E}\, \ell(f(X), Y) \text{ s.t. } f \in \mathcal{V} \tag{OP}$$

- $\mathcal{V}$ is space of all invariant functions f for which (for all semimetrics h)

$$\sup_{\tilde{x}\in G^x} h(f(\tilde{x}), f(x)) = 0 \qquad \forall x \in \text{supp}(\mathbb{P}_X).$$

Technical note on G^X compared to $T(X)$:

- every X belongs to unique transformation set G^X, a small subset of group transformations of X
- set of transformation sets $\{G^X\}$ partitions $\text{supp}(\mathbb{P}_X)$
- (symmetric) transformation sets $T(X)$ in practice cover G^X, that is $T(X) \supset G^X$ for all $X \in \text{supp}(\mathbb{P}_X)$

Theoretical statements for regularization

Theorem (Robust minimizers are invariant)

- *If $\mathcal{V} \subset \mathcal{F}$, all minimizers of the adversarial loss (1) are in $\mathcal{V}$*
- *Any solution of (OP) minimizes the adversarial loss*

Theorem (Trade-off natural vs. robust accuracy)

- *If $\mathcal{V} \subset \mathcal{F}$ and $Y \perp\!\!\!\perp X | G^X$, the adversarial minimizer of (1) also minimizes the standard loss (2)*
- *If moreover $\bigcup G^X = supp(\mathbb{P}_X)$, and loss ℓ is injective, then every standard minimizer of (2) is in $\mathcal{V}$*

References

[1] H. Kannan *et al.*, "Adversarial logit pairing," *arXiv:1803.06373*, 2018.

[2] H. Zhang *et al.*, "Theoretically principled trade-off between robustness and accuracy," *ICML*, 2019.

[3] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *ICML*, 2016.

[4] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International Conference on Machine Learning*, 2019.