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

Fanny Yang, Zuowen Wang, Christina Heinze-Deml
1Stanford; *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 (±3px) and rotations (±30°)

\[ \begin{align*}
\text{Truth: ship} & \quad \text{Pred.: airplane}
\end{align*} \]

Invariance evaluated using the adversarial loss
(in practice: grid search over 775 values)
\[
\min_{f \in F} \mathbb{E}_{X,Y} \sup_{x \in \mathcal{T}(X)} \ell(f(x'), Y),
\]
Also care about standard accuracy (std)
\[
\min_{f \in F} \mathbb{E}_{X,Y} \ell(f(X), Y)
\]

Algorithm: Regularizing standard and adversarial augmentation
Gradient update on a minibatch:
\[
\theta^{t+1} = \theta^t - \gamma_\ell \nabla \mathbb{E}_{X,Y} \ell(f(x), Y) + \lambda \nabla h(f_\theta(x), f_\theta(x'))
\]

Invariance-regularizer:
- \( \lambda \nabla h(f_\theta(x), f_\theta(x')) \): modified adv. training (e.g. \([1]\) for \( \ell_2 \))
- \( \lambda \nabla h(f_\theta(x), f_\theta(x')) \): mod. adv. training (e.g. \([2]\) for KL)

Practical benefits of regularization
- 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 \( \ell_2 \)-dist. as semimetric \( h \)
- \( f_\theta(x) \) are logits (\( \ell_2 \)) and post-sofmax activations (\( KL \))
- GRN: G-ResNet44 \([3]\) has convolutions with multiple rotated filters

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

Theoretical framework
- Regularized algorithm with \( \tilde{x} = \arg \max_{x \in \mathcal{G}_h} h(f_\theta(x), f_\theta(x')) \) and
\[
\theta^{t+1} = \theta^t - \gamma_\ell \mathbb{E} \left[ \nabla \ell(x, y; \theta^t) + \lambda \nabla h(f_\theta(x), f_\theta(x')) \right]
\]
is a first-order method for minimizing the penalized loss
\[
\min_{f \in F} \mathbb{E} \ell(f(X), Y) + \lambda \sup_{x \in \mathcal{G}_h} h(f_\theta(x), f_\theta(x'))
\]
- For some \( \lambda \), this is the dual of the constrained problem
\[
\min_{f \in F} \mathbb{E} \ell(f(X), Y) \quad \text{s.t.} \quad f \in \mathcal{V}
\] (OP)
- \( \mathcal{V} \) is space of all invariant functions \( f \) for which (for all semimetrics \( h \))
\[
\sup_{x \in \mathcal{G}_h} h(f_\theta(x), f_\theta(x')) = 0 \quad \forall x \in \text{supp}(\mathbb{P}_X).
\]

Technical note on \( \mathcal{G}_h \) compared to \( T(X) \):
- every \( X \) belongs to unique transformation set \( \mathcal{G}_h \),
- a small subset of group transformations of \( X \)
- set of transformation sets \( \{\mathcal{G}_h\} \) partitions \( \text{supp}(\mathbb{P}_X) \)
- (symmetric) transformation sets \( T(X) \) in practice cover \( \mathcal{G}_h \),
- that is \( T(X) \supset \mathcal{G}_h \) 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 \((\text{OP})\) minimizes the adversarial loss

Theorem (Trade-off natural vs. robust accuracy)
- If \( \mathcal{V} \subset \mathcal{F} \) and \( Y \perp \cond X | \mathcal{G}_h \), the adversarial minimizer of \((1)\)
  also minimizes the standard loss \((2)\)
- If moreover \( \bigcup \mathcal{G}_h = \text{supp}(\mathbb{P}_X) \), and loss \( \ell \) is injective,
  then every standard minimizer of \((2)\) is in \( \mathcal{V} \)

References