Adversarial Training with Unlabeled Data

Narsil Zhang

Department of Computer Science
The University of Texas at Austin

Introduction

- **Adversarial Training**:
  At each iteration, using attack methods (e.g. PGD, C&W) to augment data, and training the model with these generated data and clean data.
- **Objective**:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\eta\|\leq\epsilon} \ell(\theta; x+\eta, y), \tag{1}$$

Semi Supervised Learning

$\mathcal{D}^l$: labeled data; $\mathcal{D}^{ul}$: unlabeled data
Virtual Adversarial Training (VAT):

$$\mathbb{E}_{(x,y)\sim\mathcal{D}^l}\ell(\theta; x, y) + \lambda\mathbb{E}_{x\sim\mathcal{D}^l\bigcup\mathcal{D}^{ul}}\mathbb{D}\{p(y|x)||p(y|x')\} \tag{2}$$

Generally speaking, $\mathbb{D}$ can be any divergence measure. The authors take KL divergence in their paper.

Adv. Methodology

x' is an adv. example of x, $f(x)$ denotes the logits, i.e. $p(y|x)$.
Adversarial Logit Pairing:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\theta; x, y) + \lambda \|f(x) - f(x')\|_2] \tag{3}$$

TRADES:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\theta; x, y) + \lambda \cdot \mathrm{KL}(f(x)||f(x'))] \tag{4}$$

The second term can be used with unlabeled data.

BTW...

There are 3 NIPS submission papers talking about the last point.

- Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi. Unlabeled data improves adversarial robustness.
  arXiv preprint arXiv:1905.13736, 2019
- R. Zhai, T. Cai, D. He, C. Dan, K. He, J. Hopcroft, and L. Wang. Adversarially robust generalization just requires more unlabeled data.
  arXiv preprint arXiv:1906.00555, 2019.
- J.Uesato,J.Alayrac,P.Huang,R.Stanforth,A.Fawzi, and P. Kohli. Are labels required for improving adversarial robustness?
  arXiv preprint arXiv:1905.13725, 2019.

Theoretical Framework

Now we introduce a framework for analysing the relationship between **number of data** and **classification risk**:
(Gaussian model).

- Let $\mu \in \mathbb{R}^d$ be the per-class mean vector. Let $\sigma > 0$ be the variance parameter.
  Then the (μ, σ) -Gaussian model is defined by the following distribution over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$:
  First, draw a label $y \in \{\pm 1\}$ uniformly at random. Then sample the data point $x \in \mathbb{R}^d$ from $\mathcal{N}\left(y \cdot \mu, \sigma^2 I\right)$
- classifier $f_\theta(x) = \operatorname{sign}(\theta^T x)$

Theoretical Framework

- $\mathrm{err}_{\mathrm{standard}}\ (f_\theta) := \mathbb{P}_{(x,y)\sim P_{x,y}}\left(f_\theta(x) \neq y\right)$
- $\mathrm{err}^{\infty,\epsilon}_{\mathrm{robust}}\ (f_\theta) := \mathbb{P}_{(x,y)\sim P_{x,y}}\left(\exists x' \in \mathcal{B}^{\infty}_{\epsilon}(x), f_\theta\left(x'\right) \neq y\right)$
  for $\mathcal{B}^{\infty}_{\epsilon}(x) := \{x' \in \mathcal{X} \mid \|x' - x\|_\infty \leq \epsilon\}$

Algorithm

- Supervised:$\hat{\theta}_n := \frac{1}{n}\sum_i y_i x_i$
- Semi-Supervised: self-labeling:
  1. $\hat{\theta}_{\text{intermediate}} = \hat{\theta}_n = \frac{1}{n}\sum_i y_i x_i$ for labeled data
  2. $\tilde{y}_j = \text{sign}(\hat{\theta}^T_{\text{intermediate}} x_j)$ for unlabeled data
  3. $\hat{\theta}_{\text{final}} := \frac{1}{\tilde{n}}\sum_j y_j x_j$ for all data

Analysis Sketch: Step 1

Transform the upper bound of risk

$$
\begin{aligned}
\mathrm{err}_{\text{robust}}^{\infty,\epsilon}(f_\theta) &= \mathbb{P}\left(\inf_{\|\nu\|_\infty \le \epsilon} \left\{ y \cdot (x+\nu)^\top \theta \right\} < 0\right) \\
&= \mathbb{P}\left(y \cdot x^\top \theta - \epsilon \|\theta\|_1 < 0\right) = \mathbb{P}\left(\mathcal{N}\left(\mu^\top \theta, (\sigma\|\theta\|)^2\right) < \epsilon\|\theta\|_1\right) \\
&= Q\left(\frac{\mu^\top \theta}{\sigma\|\theta\|} - \frac{\epsilon\|\theta\|_1}{\sigma\|\theta\|}\right) \le Q\left(\frac{\mu^\top \theta}{\sigma\|\theta\|} - \frac{\epsilon\sqrt{d}}{\sigma}\right)
\end{aligned}
$$

into the lower bound of $\frac{\mu^\top \theta}{\sigma\|\theta\|}$.
($Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$ is monotonously decreasing)

Analysis Sketch: Step 2

$\hat{\theta}_n = \frac{1}{n}\sum_i y_i x_i \sim N(\mu, \frac{\sigma^2}{n} I)$

$$\therefore \delta := \hat{\theta}_n - \mu \sim N(0, \frac{\sigma^2}{n} I)$$

$$\frac{\left\|\hat{\theta}_n\right\|^2}{\left(\mu^\top \hat{\theta}_n\right)^2} = \frac{\|\delta + \mu\|^2}{\left(\|\mu\|^2 + \mu^\top \delta\right)^2} = \cdots = \frac{1}{\|\mu\|^2} + \frac{\|\delta\|^2 - \frac{1}{\|\mu\|^2}\left(\mu^\top \delta\right)^2}{\left(\|\mu\|^2 + \mu^\top \delta\right)^2}$$

$$\leq \frac{1}{\|\mu\|^2} + \frac{\|\delta\|^2}{\left(\|\mu\|^2 + \mu^\top \delta\right)^2}$$

Using $\|\delta\|^2 \sim \frac{\sigma^2}{n}\chi_d^2$ and $\frac{\mu^\top \delta}{\|\mu\|} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$ and standard concentration can give a bound