深度学习理论

孔之丰
z4kong@eng.ucsd.edu

University of California San Diego

August 8, 2018

主要内容

- 深度学习的理论研究
- Example: Margin Bounds for Neural Networks
- 一些讨论与展望

深度学习的理论研究

神经网络的性质

- 深度神经网络的逼近性
  - 对具有特定性质连续函数的逼近理论
  - GAN的逼近理论
- 深度神经网络的收敛性
  - 浅层全连接网络的完整理论分析
  - 深层网络的试探性理论分析
  - 鞍点的分析
- 深度神经网络的(各种)上下界
  - VC-dimension, Rademacher Complexity
  - Sample Complexity, Margin Bounds, ...

神经网络的理论保证

- 深度神经网络的速度
  - 训练(收敛)速度
  - 推理速度
- 深度神经网络的解释性
  - 可视化
  - 逐层分析
- 深度神经网络的泛化性
  - 泛化误差界
  - 流形结构分析

神经网络与其它方法的关联

- 新的优化方法
  - 各种trick(Momentum, Dropout, BN, Over parametrization...)
  - 基于梯度的训练方法
  - 不基于梯度的训练方法
- 与统计学习的关系
  - 探究相关性(Kernel Learning, ...)
  - 严格的比较(ResNet, ...)
- 实验的验证
  - 正面的结果(太多了)
  - 负面的结果(逐渐变多, CIFAR10, FFT/Parity)

深度学习的改进(例)

- 新的架构
  - GAN
  - Capsule
  - D^2NN
- 性能提升
  - 模型压缩
  - 计算加速
- 理论指导
  - WGAN
  - Selu

Example: Margin Bounds for Neural Networks

Peter L. Bartlett, Dylan J. Fostery, Matus Telgarskyz,
Spectrally normalized margin bounds for neural networks,
arXiv:1706.08498v2

Margin Bounds

- 问题描述

$$P\left\{\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right\} \leq ?$$

- 结论：以概率不低于$1-\delta$

$$P\{\cdot\} \leq \hat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + \tilde{\mathcal{O}}\left(\frac{\|X\|_2 R_{\mathcal{A}}}{\gamma^n} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

其中

$$R_{\mathcal{A}} = \left(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma\right)\left(\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\right)^{3/2}$$

$$\hat{\mathcal{R}}_\gamma(f) \leq \frac{1}{n}\sum_i \mathbf{1}[f(x_i)_{y_i} \leq \gamma + \max_{j \neq y_i} f(x_i)_j]$$

求解步骤

- 问题转化：$\mathcal{F}_\gamma = \{(x,y) \mapsto \ell_\gamma(-\mathcal{M}(f(x),y)), f \in \mathcal{F}\}$, 则以概率$1-\delta$

$$P\{\cdot\} \leq \mathcal{R}_\gamma(f) \leq \hat{\mathcal{R}}_\gamma(f) + 2\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) + 3\sqrt{\frac{\log(1/\delta)}{2n}}$$

其中$\mathfrak{R}(\mathcal{H}_{|S}) = \frac{1}{n}\mathbb{E}\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\epsilon_i h(x_i,y_i)$.

- 下一步：寻找中间量 $\geq \mathfrak{R}$ 且$\leq$ 某个可以直接计算的值。引入Matrix Covering Number

$$\mathcal{N}(U,\epsilon,\|\cdot\|) = \min\{|V| : \forall A \in U,\ \exists B \in V\ s.t.\ \|A-B\| \leq \epsilon\}$$

求解步骤

- 对单层$\mathcal{N}$的估计

$$\log \mathcal{N}\left(\left\{XA : A \in \mathbb{R}^{d\times m}, \|A\|_{q,s} \le a\right\}, \epsilon, \|\cdot\|_2\right)$$

$$\le \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \right\rceil \log(2dm)$$

- 对整个网络$\mathcal{N}$的估计

$$\begin{aligned}\log \mathcal{N}(\mathcal{H}_X, \epsilon, \|\cdot\|_2) \quad &\le \quad \tfrac{1}{\epsilon^2}\|X\|_2^2 \log(2W^2) \\ &\quad \times \left(\textstyle\prod_{j=1}^{L} s_j^2 \rho_j^2\right)\left(\sum_{i=1}^{L}\left(\frac{b_i}{s_i}\right)^{2/3}\right)^3\end{aligned}$$

- 对$\mathfrak{R}$的估计

$$\mathfrak{R}(\mathcal{F}_{|S}) \le \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \epsilon, \|\cdot\|_2)}\, d\epsilon\right)$$

讨论与展望

- 这一代深度学习
  - 更统一的架构/ 新的数学框架
  - DL作为一门科学的定位/ 标准实验过程的引入
- 下一代深度学习
  - 强推理能力/ Wolfram语言
  - 全新的架构与训练方式/ 动态变化, 模糊性
- 与产业的关系

参考文献(arXiv preprint)

- 1709.02540 (Expressive power)
- 1705.08991 (GAN theory)
- 1703.00560 (ReLU network convergence)
- 1702.07966 (ConvNet convergence)
- 1706.08498 (Margin bound)
- 1712.06541 (Sample complexity)
- 1802.01396 (DL $\leftrightarrow$ kernel learning)
- 1802.06509 (Overparameterization)
- 1806.00451 (CIFAR10 $\rightarrow$ CIFAR10)
- 1804.08838 (Intrinsic dimension)
- 1706.02515 (SELU)
- 1701.07875 (WGAN)

Thanks for your attention!
Q & A