[Paper review 15]

Fast Dropout Training

(Sida I. Wang, Christopher D. Manning, 2013)

[Contents]

0. Abstract

- Dropout (Hinton et al., 2012)

  but, repeatedly sampling makes much slower!

- This paper shows how to do fast dropout training!

  "by sampling from, or integrating a GAUSSIAN APPROXIMATION" (instead of doing MC optimization) (by CLT)

1. Introduction

Dropout

- prevent feature co-adaptation $\rightarrow$ regularization
- can be seen as "averaging over many NN with shared weights"

Problem with Dropout

- makes training SLOWER!
- loss of information

  (drop out rate of p : proportion of data still not seen after n passes is p^n)

This paper suggests "benefit of dropout training without actually sampling", thereby using ALL data efficiently

$\rightarrow$ use Gaussian Approximation

2. Fast approximations to dropout

2.1 The implied objective function

example) Logistic Regression with dropout

- m dimension data
- $z_i \sim$ Bernoulli (p_i) , where $i = 1 \ldots m$
- SGD update : $\Delta w = \left(y - \sigma\left(w^T D_z x\right)\right) D_z x$
  - $D_z = \operatorname{diag}(z) \in \mathbb{R}^{m \times m}$
  - $\sigma(x) = 1/\left(1 + e^{-x}\right)$
- MC approximation : $\Delta \bar{w} = E_{z; z_i \sim \text{ Bernoulli } (p_i)}\left[\left(y - \sigma\left(w^T D_z x\right)\right) D_z x\right]$

Objective function of gradient above :

- $y \sim \operatorname{Bernoulli}\left(\sigma\left(w^T D_z x\right)\right)$

$$\begin{aligned} L(w) & = E_z\left[\log(p\left(y \mid D_z x; w\right)\right] \\ & = E_z\left[y \log\left(\sigma\left(w^T D_z x\right)\right) + (1-y) \log\left(1 - \sigma\left(w^T D_z x\right)\right)\right] \end{aligned}$$

- complexity : $O(2^m m)$
- can be reduced to $O(m)$. .... HOW?

2.2 The Gaussian approximation

(now, let $Y(z) = w^T D_z x = \sum_i^m w_i x_i z_i$ ..... weighted sum of Bernoulli r.v)

Y can be approximated by Normal distribution (as $m \rightarrow \infty$)

let $Y \xrightarrow{d} S$

$S = E_z[Y(z)] + \sqrt{\operatorname{Var}[Y(z)]} \epsilon = \mu_S + \sigma_S \epsilon$

- $\epsilon \sim \mathcal{N}(0,1)$
- $E_z[Y(z)] = \sum_{i=1}^m p_i w_i x_i$,
- $\operatorname{Var}[Y(z)] = \sum_{i=1}^m p_i\left(1 - p_i\right)\left(w_i x_i\right)^2$

2.3 Gradient computation by sampling from Gaussian

BEFORE) sample from $Y(z)$ directly

- time : $O(m)$
- d

AFTER) sample from S

- especially good for high dimensional case

- time : $O(1)$ (m times faster !)

$$L(w)=E_z\left[y\log\left(\sigma\left(w^TD_zx\right)\right)+(1-y)\log\left(1-\sigma\left(w^TD_zx\right)\right)\right]$$

$$\nabla L(w)=E_z\left[(y-\sigma(Y(z)))D_zx\right]$$

- $f(Y(z))=y-\sigma(Y(z))$
- $g(z)=D_zx$

$$\begin{aligned}\nabla L(w)=&E_z\left[(y-\sigma(Y(z)))D_zx\right]\\ =&E_z\left[f(Y(z))x_iz_i\right]\\ =&\sum_{z_i\in\{0,1\}}p\left(z_i\right)z_ix_iE_{z_{-i}|z_i}[f(Y(z))]\\ =&p\left(z_i=1\right)x_iE_{z_{-i}|z_i=1}[f(Y(z))]\\ \approx&p_ix_i\left(E_{S\sim\mathcal{N}\left(\mu_S,\sigma_S^2\right)}[f(S)]+\Delta\mu_i\left.\frac{\partial E_{S\sim\mathcal{N}\left(\mu,\sigma_S^2\right)}[f(S)]}{\partial\mu}\right|_{\mu=\mu_S}+\Delta\sigma_i^2\left.\frac{\partial E_{S\sim\mathcal{N}\left(\mu_S,\sigma^2\right)}[f(S)]}{\partial\sigma^2}\right|_{\sigma^2=\sigma_S^2}\right)\\ =&p_ix_i\left(\alpha\left(\mu_S,\sigma_S^2\right)+\Delta\mu_i\beta\left(\mu_S,\sigma_S^2\right)+\Delta\sigma_i^2\gamma\left(\mu_S,\sigma_S^2\right)\right)\end{aligned}$$

- $\Delta\mu_i=\left(1-p_i\right)x_iw_i$
- $\Delta\sigma_i^2=-p_i\left(1-p_i\right)x_i^2w_i^2$

α,β,γ can be computed by drawing K samples from $S\rightarrow$ takes $O(K)$ time

($\leftrightarrow$ if sample from $Y(z)$, takes $O(mK)$ time!)

- α only need to be computed ONE per training case
- $\beta\left(\mu,\sigma^2\right)=\frac{\partial\alpha\left(\mu,\sigma^2\right)}{\partial\mu}$

  $\gamma\left(\mu,\sigma^2\right)=\frac{\partial\alpha\left(\mu,\sigma^2\right)}{\partial\sigma^2}$
- $\alpha\left(\mu_S,\sigma_S^2\right)=E_{S\sim\mathcal{N}\left(\mu_S,\sigma_S^2\right)}[f(S)]$

  $\alpha\left(\mu,\sigma^2\right)=y-E_{S\sim\mathcal{N}(0,1)}\left[\frac{1}{1+e^{-\mu-\sigma_SS}}\right]$

$$\begin{aligned}L(w)\quad&=E_z\left[\log(p\left(y\mid D_zx;w\right)\right]\\ &=E_z\left[y\log\left(\sigma\left(w^TD_zx\right)\right)+(1-y)\log\left(1-\sigma\left(w^TD_zx\right)\right)\right]\\ &\approx E_{S\sim\mathcal{N}\left(\mu_S,\sigma_S\right)}[y\log(\sigma(S))+(1-y)\log(1-\sigma(S))]\end{aligned}$$

2.4 A closed- form approximation

$\Phi(x)$: CDF of Gaussian ($=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt$)

$\sigma(x)$: sigmoid (logistic) function

Since $\sigma(x)\approx\Phi(\sqrt{\pi/8}x)$,

- $\int_{-\infty}^{\infty}\Phi(\lambda x)\mathcal{N}(x\mid\mu,s)dx=\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2}+s^2}}\right)$
- $\int_{-\infty}^{\infty}\sigma(x)\mathcal{N}\left(x\mid\mu,s^2\right)dx\approx\sigma\left(\frac{\mu}{\sqrt{1+\pi s^2/8}}\right)$

Apply the above to our case ...

$$\begin{aligned} E_{X\sim\mathcal{N}(\mu,s^2)}[\log(\sigma(X))] &= \int_{-\infty}^{\infty} \log(\sigma(x))\mathcal{N}\left(x \mid \mu, s^2\right) dx \\ &\approx \sqrt{1+\pi s^2/8}\log\sigma\left(\frac{\mu}{\sqrt{1+\pi s^2/8}}\right) \end{aligned}$$