[Paper review 18]

Variational Dropout Sparsifies Deep Neural Networks

(Dmitry Molchannov, et al., 2017)

[Contents]

0. Abstract

Extend Variational Dropout to the case when dropout rates are unbounded

Propose a way to reduce the variance of the gradient estimator

1. Introduction

Dropout

- Binary Dropout (Hinton et al., 2012)
- Gaussian Dropout (Srivastava et al, 2014)

  (multiplies the outputs of the neurons by Gaussian random noise)
- Dropout rates are usually optimized by grid-search

  (To avoid exponential complexity, dropout rates are usually shared for all layers)
- can be seen as a Bayesian regularization (Gal & Ghahramani, 2015)

Instead of injecting noise, Sparsity!

- inducing sparsity during training DNN leads regularization (Han et al., 2015a)
- Sparse Bayesian Learning (Tipping, 2001)

  (provies framework for training of sparse models)

This paper

- 1) study Variational Dropout (Kingma et al, 2015) where each weight of a model has its own individual dropout rate

- 2) propose Sparse Variational Dropout

  "extends VD" to all possible values of drop out rates (= α)

  (to do this, provide a new approximation of KL-divergence term in VD objective)
- 3) propose a way to reduce variance of stochastic gradient estimator

  $\rightarrow$ leads to faster convergence

2. Related Work

너무 많아..생략

논문 참조

3. Preliminaries

3.1 Bayesian Inference

HOW to minimize $D_{KL}\left(q_\phi(w) \| p(w \mid \mathcal{D})\right)$?

Maximize ELBO = (1) Expected Log-likelihood - (2) KL-divergence

ELBO : $\mathcal{L}(\phi)=L_{\mathcal{D}}(\phi)-D_{KL}\left(q_\phi(w) \| p(w)\right) \rightarrow \max_{\phi \in \Phi}$

- (1) Expected Log-likelihood : $L_{\mathcal{D}}(\phi)=\sum_{n=1}^N \mathbb{E}_{q_\phi(w)}\left[\log p\left(y_n \mid x_n, w\right)\right]$
- (2) KL-divergence : $D_{KL}\left(q_\phi(w) \| p(w)\right)$

3.2 Stochastic Variational Inference

(a) Reparameterization Trick (Kingma & Welling, 2013)

- obtain unbiased differentiable minibatch-based MC estimator of expected log-likelihood

  (that is, find $\nabla_\phi L_{\mathcal{D}}\left(q_\phi\right)$)
- trick : decompose into (1) deterministic & (2) stochastic part

  $w=f(\phi, \epsilon)$ where $\epsilon \sim p(\epsilon)$
- number of data in one mini-batch : M

  $\mathcal{L}(\phi) \simeq \mathcal{L}^{SGVB}(\phi)=L_{\mathcal{D}}^{SGVB}(\phi)-D_{KL}\left(q_\phi(w) \| p(w)\right)$

  $L_{\mathcal{D}}(\phi) \simeq L_{\mathcal{D}}^{SGVB}(\phi)=\frac{N}{M} \sum_{m=1}^M \log p\left(\tilde{y}_m \mid \tilde{x}_m, f\left(\phi, \epsilon_m\right)\right)$

  $\nabla_\phi L_{\mathcal{D}}(\phi) \simeq \frac{N}{M} \sum_{m=1}^M \nabla_\phi \log p\left(\tilde{y}_m \mid \tilde{x}_m, f\left(\phi, \epsilon_m\right)\right)$

(b) Local Reparameterization Trick (Kingma et al., 2015)

- sample separate weight matrices for each data-point inside mini-batch
- done efficiently by moving the noise from "weights" to "activation"

3.3 Variational Dropout

$B=(A \odot \Xi) W, \text{ with } \xi_{mi} \sim p(\xi)$ .... putting noise on INPUT

Bernoulli(Binary) Dropout

- Hinton et al., 2012
- $\xi_{mi} \sim$ Bernoulli $(1-p)$

Gaussian Dropout with continuous noise

- Srivastava et al, 2014
- $\xi_{mi} \sim \mathcal{N}(1, \alpha=\frac{p}{1-p})$
- continuous noise is better than discrete noise

  (multiplying the inputs by Gaussian noise = putting Gaussian noise on the weights)
- can be used to obtain posterior distribution over model's weight! (Wang & Manning, 2013), (Kingma et al., 2015)

  ($\xi_{ij} \sim \mathcal{N}(1, \alpha)$ = sampling w_{ij} from $q\left(w_{ij} \mid \theta_{ij}, \alpha\right)=\mathcal{N}\left(w_{ij} \mid \theta_{ij}, \alpha \theta_{ij}^{2}\right)$.)

  (Then, $w_{ij}=\theta_{ij} \xi_{ij}=\theta_{ij}\left(1+\sqrt{\alpha} \epsilon_{ij}\right) \sim \mathcal{N}\left(w_{ij} \mid \theta_{ij}, \alpha \theta_{ij}^{2}\right) \quad \text{where} \quad \epsilon_{ij} \sim \mathcal{N}(0,1)$)

Variational Dropout

- (use reparam trick + draw single sample $W \sim q(W \mid \theta, \alpha)$)

  $\rightarrow$ Gaussian dropout = stochastic optimization of exxpected log likelihood
- VD extends this technique!

  use $q(W \mid \theta, \alpha)$ as an approximate posterior with special prior,
  $p\left(\log \left|w_{ij}\right|\right)=\text{const} \Leftrightarrow p\left(\left|w_{ij}\right|\right) \propto \frac{1}{\left|w_{ij}\right|}$

GD Training = VD Training (when α is fixed)

However, VD provides a way to train dropout rate α by optimizing the ELBO

4. Sparse Variational Dropout

difficulties in training the model with large values of α

$\rightarrow$ have considered the case of $\alpha \leq 1$ ($\leftrightarrow$ $p \leq 0.5$ in binary dropout)

High dropout rate $\alpha_{ij} \rightarrow +\infty$ = $p = 1$

(meaning : corresponding weight is always ignored & can be removed)

4.1 Additive Noise Reparameterization

$$\frac{\partial \mathcal{L}^{SGVB}}{\partial \theta_{ij}} = \frac{\partial \mathcal{L}^{SGVB}}{\partial w_{ij}} \cdot \frac{\partial w_{ij}}{\partial \theta_{ij}} = (1) \times (2)$$

(2) is very noisy if α_{ij} is large.

$w_{ij} = \theta_{ij} \left(1 + \sqrt{\alpha_{ij}} \cdot \epsilon_{ij}\right)$

$\frac{\partial w_{ij}}{\partial \theta_{ij}} = 1 + \sqrt{\alpha_{ij}} \cdot \epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}(0,1)$

How to reduce variance when α_{ij} is large ?

replace multiplicative noise term $1 + \sqrt{\alpha_{ij}} \cdot \epsilon_{ij}$ .... with $\sigma_{ij} \cdot \epsilon_{ij}$,

(where $\sigma_{ij}^2 = \alpha_{ij} \theta_{ij}^2$)

$$\begin{aligned} w_{ij} = & \theta_{ij} \left(1 + \sqrt{\alpha_{ij}} \cdot \epsilon_{ij}\right) \\ & = \theta_{ij} + \sigma_{ij} \cdot \epsilon_{ij} \end{aligned}$$

Thus, $\frac{\partial w_{ij}}{\partial \theta_{ij}} = 1, \quad \epsilon_{ij} \sim \mathcal{N}(0,1)$
(has no injection noise!)

avoid the problem of large gradient variance!

can train the model within the full range of $\alpha_{ij} \in (0, +\infty)$

4.2. Approximation of the KL Divergence

full KL-divergence term in ELBO

$D_{KL}(q(W \mid \theta, \alpha) \| p(W)) = \sum_{ij} D_{KL}\left(q\left(w_{ij} \mid \theta_{ij}, \alpha_{ij}\right) \| p\left(w_{ij}\right)\right)$

log-scale uniform prior distribution is an improper prior

$-D_{KL}\left(q\left(w_{ij} \mid \theta_{ij}, \alpha_{ij}\right) \| p\left(w_{ij}\right)\right) = \frac{1}{2} \log \alpha_{ij} - \mathbb{E}_{\epsilon \sim \mathcal{N}(1, \alpha_{ij})}$

Term above is intractable in VD

need to be sampled & approximated

$$-D_{KL}\left(q\left(w_{ij} \mid \theta_{ij}, \alpha_{ij}\right) \| p\left(w_{ij}\right)\right) \approx \approx k_1 \sigma\left(k_2 + k_3 \log \alpha_{ij}\right)\right) - 0.5 \log\left(1 + \alpha_{ij}^{-1}\right) + \mathrm{C}$$
$$k_1 = 0.63576 \quad k_2 = 1.87320 \quad k_3 = 1.48695$$