[Paper review 16]

Variational Dropout and the Local Reparameterization Trick

(Kingma et al., 2015)

[Contents]

0. Abstract

propose LRT for reducing variance of SGVB

- LRT : Local Reparameterization Trick
- SGVB : Stochastic Gradients for Variational Bayesian inference

LRT

- translates uncertainty about global parameters into local noise (which is independent across mini-batchs)
- can be parallelized
- have variance, which is inversely proportional to the mini-batch size(M)

Explore a connection with dropout

- Gaussian dropout objectives correspond to SGVB with LRT
- propose "Variational Dropout"

  (= generalization of Gaussian Dropout)

1. Introduction

Gaussian Dropout :

- regular(binary) dropout has Gaussian approximation
- faster convergence
- optimizes a lower bound on the marginal likelihood of the data

In this paper...

"relationship between dropout & Bayesian Inference" can be extended and exploited to greatly improve the efficiency of variational Bayesian Inference!

Previous works

- MCMC
- Variational Inference...

$\Rightarrow$ those have not been shown to outperform simpler methods such as "drop out"

Proposes...

- trick for improving the efficiency of stochastic gradient-based variational inference with "mini-batches" of data

  (by translating uncertainty about global parameters into local noise)

  $\rightarrow$ has a optimization speed on the same level as "fast dropout" (Wang, 2013)

2. Efficient and Practical Bayesian Inference

Variational Inference

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

  (where $L_{\mathcal{D}}(\phi)=\sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} \mathbb{E}_{q_{\phi}(\mathbf{w})}[\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{w})]$.... called "expected log-likelihood")
- have to maximize ELBO

2.1 SGVB (Stochastic Gradient Variational Bayes)

Two key points

- 1) parameterize the random parameters
  - from : $w \sim q_{\phi}(w)$
  - to : $w=f(\epsilon, \phi)$, where $\epsilon \sim p(\epsilon)$
- 2) unbiased differentiable minibatch-based MC estimator (of the expected log likelihood) :

  $L_{\mathcal{D}}(\phi) \simeq L_{\mathcal{D}}^{\mathrm{SGVB}}(\phi)=\frac{N}{M} \sum_{i=1}^{M} \log p\left(\mathbf{y}^{i} \mid \mathbf{x}^{i}, \mathbf{w}=f(\epsilon, \phi)\right)$
  - (M : # of data in a mini-batch)
  - differentiable w.r.t ϕ
  - unbaised

  Thus, its gradient is also unbiased! : $\nabla_{\phi} L_{\mathcal{D}}(\phi) \simeq \nabla_{\phi} L_{\mathcal{D}}^{\mathrm{SGVB}}(\phi)$

2.2 Variance of the SGVB estimator

(theory) stochastic approximation $\rightarrow$ asymptotically converge to a local optimum

(practice) depends on "the variance of the gradient"

Assume we draw a mini-batch with replacement & let
$L_{\mathcal{D}}^{\mathrm{SGVB}}(\phi)=\frac{N}{M} \sum_{i=1}^{M} \log p\left(\mathbf{y}^{i} \mid \mathbf{x}^{i}, \mathbf{w}=f(\epsilon, \phi)\right)=\frac{N}{M} \sum_{i=1}^{M} L_{i}$,

$$
\begin{aligned}
\operatorname{Var}\left[L_{\mathcal{D}}^{\mathrm{SGVB}}(\phi)\right] & =\frac{N^{2}}{M^{2}}\left(\sum_{i=1}^{M} \operatorname{Var}\left[L_{i}\right]+2 \sum_{i=1}^{M} \sum_{j=i+1}^{M} \operatorname{Cov}\left[L_{i}, L_{j}\right]\right) \\
& =N^{2}\left(\frac{1}{M} \operatorname{Var}\left[L_{i}\right]+\frac{M-1}{M} \operatorname{Cov}\left[L_{i}, L_{j}\right]\right)
\end{aligned}
$$

- term (1) $\frac{1}{M} \operatorname{Var}\left[L_{i}\right]$
  - inversely proportional to the mini-batch size M
- term (2) $\frac{M-1}{M} \operatorname{Cov}\left[L_{i}, L_{j}\right]$
  - does not decrease with M

$\Rightarrow \operatorname{Var}\left[L_{\mathcal{D}}^{\mathrm{SGVB}}(\phi)\right]$ can be dominated by the covariances for even moderately large M

2.3 LRT (Local Reparameterization Trick)

To solve the problem above...

[1] propose alternative estimator which $\operatorname{Cov}\left[L_{i}, L_{j}\right]=0$

$\rightarrow$ SG scales as $1/M$!

[2] Then, for efficiency,

- do not sample ϵ directly
- rather, sample the intermediate values $f(\epsilon)$

by doing so ([1] & [2])

"the global uncertainty in the weights is translated into a form of local uncertainty", that is independent across examples

Example

structure :

- 1000 neurons
- input feature dim : $M \times 1000$
- weight matrix (W) dim : 1000×1000
- $B=AW$
- posterior approximation on weights = fully factorized Gaussian

- $q_\phi\left(w_{i,j}\right)=\dot{N}\left(\mu_{i,j}, \sigma_{i,j}^2\right) \quad , \quad \forall w_{i,j} \in \mathbf{W}$
- $w_{i,j}=\mu_{i,j}+\sigma_{i,j} \epsilon_{i,j}$

  $\epsilon_{i,j} \sim N(0,1)$.
- Then, $\operatorname{Cov}\left[L_i, L_j\right]=0$

property 1) EFFICIENT

Method 1) sample a separate weight matrix $\mathbf{W}$

- computationally inefficient

  (need to sample M million random numbers for just a single layer of NN)
- $q_\phi\left(w_{i,j}\right)=N\left(\mu_{i,j}, \sigma_{i,j}^2\right) \forall w_{i,j} \in \mathbf{W}$

Method 2) sample the random activations B directly

- without sampling $\mathbf{W}$ or ϵ
- how is it possible? "weights only influence the expected log-likelihood through B "
- more efficient MC estimator!
- $q_\phi\left(b_{m,j} \mid \mathbf{A}\right)=N\left(\gamma_{m,j}, \delta_{m,j}\right)$

  $\gamma_{m,j}=\sum_{i=1}^{1000} a_{m,i} \mu_{i,j}, \quad \text{and} \quad \delta_{m,j}=\sum_{i=1}^{1000} a_{m,i}^2 \sigma_{i,j}^2$

In short...

- rather than sampling the Gaussian weights,
- sample the "activations" from their implied Gaussian dist'n,

  using $b_{m,j}=\gamma_{m,j}+\sqrt{\delta_{m,j}} \zeta_{m,j}$, with $\zeta_{m,j} \sim N(0,1)$
- ζ is an $M \times 1000$ matrix
  - "only need to sample M thousand random variables instead of M million"
- "MUCH MORE EFFICIENT"

property 2) LOWER VARIANCE

Method 1) sample a separate weight matrix $\mathbf{W}$

- 1000 $\epsilon_{i,j}$ influencing each gradient term

$$\frac{\partial L_{\mathcal{D}}^{\mathrm{SGVB}}}{\partial \sigma_{i,j}^2}=\frac{\partial L_{\mathcal{D}}^{\mathrm{SGVB}}}{\partial b_{m,j}} \frac{\epsilon_{i,j} a_{m,i}}{2 \sigma_{i,j}}$$

Method 2) sample the random activations B directly (use LRT)

- $\zeta_{m,j}$ is the only r.v. influencing the gradient (via $b_{m,j}$)

$$\frac{\partial L_{\mathcal{D}}^{\mathrm{SGVB}}}{\partial \sigma_{i,j}^{2}}=\frac{\partial L_{\mathcal{D}}^{\mathrm{SGVB}}}{\partial b_{m,j}} \frac{\zeta_{m,j} a_{m,i}^{2}}{2 \sqrt{\delta_{m,j}}}$$

3. Variational Dropout

Dropout : $\mathbf{B}=(\mathbf{A} \circ \xi) \theta, \text{ with } \xi_{i,j} \sim p\left(\xi_{i,j}\right)$

- by adding noise $\rightarrow$ less likely to overfit
- ex) Gaussian $N(1, \alpha)$ with $\alpha=p/(1-p)$

3.1 Variational Dropout with "INDEPENDENT" weight noise

$\mathbf{B}=(\mathbf{A} \circ \xi) \theta$, where noise matrix follows $\xi \sim N(1, \alpha)$

(marginal dist'n of $b_{m,j}$ is also Gaussian)

$q_{\phi}\left(b_{m,j} \mid \mathbf{A}\right)=N\left(\gamma_{m,j}, \delta_{m,j}\right)$

- $\gamma_{m,j}=\sum_{i=1}^{K} a_{m,i} \theta_{i,j}$
- $\delta_{m,j}=\alpha \sum_{i=1}^{K} a_{m,i}^{2} \theta_{i,j}^{2}$

3.2 Variational Dropout with "CORRELATED" weight noise

$\mathbf{B}=(\mathbf{A} \circ \xi) \theta$, where noise matrix follows $\xi_{i,j} \sim N(1, \alpha)$

$\Longleftrightarrow \mathbf{b}^{m}=\mathbf{a}^{m} \mathbf{W}$

- where $\mathbf{W}=\left(\mathbf{w}_{1}^{\prime}, \mathbf{w}_{2}^{\prime}, \ldots, \mathbf{w}_{K}^{\prime}\right)^{\prime}, \text{ and } \mathbf{w}_{i}=s_{i} \theta_{i}, \text{ with } q_{\phi}\left(s_{i}\right)=N(1, \alpha)$