Variational Dropout and the Local Reparameterization Trick

Eduard Vladimirov

MIPT, 2023

October 24, 2023

1. Motivation & Background
2. Theory
3. Empirical results

Motivation

Main idea

Efficiency of posterior inference using SGVB can be significantly improved through a local reparameterization.
The authors show how dropout is a special case of SGVB with local reparameterization, and suggest variational dropout, an extension of regular dropout where optimal dropout rates are inferred from the data.

Background

Variational lower-bound

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

$$\text{where } L_\mathcal{D}(\phi) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \mathbb{E}_{q_\psi(\mathbf{w})}(\log p(\mathbf{y}|\mathbf{x},\mathbf{w}))$$

Stochastic Gradient Variational Bayes

$$L_\mathcal{D}(\phi) \approx L_\mathcal{D}^{SGVB}(\phi) = \frac{N}{M}\sum_{i=1}^{M} \log p(\mathbf{y}^i|\mathbf{x}^i, \mathbf{w} = f(\epsilon, \phi))$$

1. $\nabla_\psi L_\mathcal{D}(\phi) \approx \nabla_\psi L_\mathcal{D}^{SGVB}(\phi)$

Variance of the SGVB estimator

Shorthands

$$L_i := \log p(\mathbf{y}^i | \mathbf{x}^i, \mathbf{w} = f(\epsilon, \phi))$$

$$L_{\mathcal{D}}^{SGVB}(\phi) = \frac{N}{M} \sum_{i=1}^{M} L_i$$

$$Var[L_i] = Var_{\epsilon, \mathbf{x}^i, \mathbf{y}^i} \left[\log p(\mathbf{y}^i | \mathbf{x}^i, \mathbf{w} = f(\epsilon, \phi)) \right]$$

Variance

$$Var\left[L_{\mathcal{D}}^{SGVB}(\phi)\right] = N^2 \left(\frac{1}{M} Var[L_i] + \frac{M-1}{M} Cov[L_i, L_j] \right)$$

Local Reparameterization Trick

We want to have $Cov[L_i, L_j] = 0$

Consider simple example:

$\mathbf{B} = \mathbf{A}\,\mathbf{W},$ where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{M\times 1000}, \mathbf{W} \in \mathbb{R}^{1000\times 1000}$

$q_\phi(w_{i,j}) = \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})\ \forall w_{i,j} \in \mathbf{W}$

$w_{i,j} = \mu_{i,j} + \sigma_{i,j}\epsilon_{i,j},$ with $\epsilon_{i,j} \sim \mathcal{N}(0,1)$

We have to sample a separate weight matrix **W** for each example in minibatch. As a result, we would need to sample M million random numbers for just a single layer!!!

Local Reparametrization Trick

Solution: sample the random activations **B** directly!

$$q_\psi(w_{i,j}) = \mathcal{N}(\mu_{i,j}, \sigma_{i,j}^2)\ \forall w_{i,j} \in \mathbf{W} \Longrightarrow q_\phi(b_{m,j}|A) = \mathcal{N}(\gamma_{im,j}, \delta_{m,j}),\ \text{with}$$

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

We only need to sample M thousands random variables

$$b_{m,j} = \gamma_{m,j} + \sqrt{\delta_{m,j}}\zeta_{m,j},\ \text{with}\ \zeta_{m,j} \sim \mathcal{N}(0,1),\ \zeta \in \mathbb{R}^{M\times 1000}.$$

Other advantage: lower variance

Variational Dropout

Dropout

$$\mathbf{B} = (\mathbf{A} \circ \xi)\,\theta \quad \text{with } \xi \sim Bern(1-p),$$
$$\text{where } \mathbf{A} \in \mathbb{R}^{M \times K}, \theta \in \mathbb{R}^{K \times L}, \mathbf{B} \in \mathbb{R}^{M \times L},$$

Gaussian Dropout

$$\xi \sim \mathcal{N}(1, \alpha),\ \alpha = p/(1-p)$$

Variational Dropout

VD with independent weight noise

$$q_\phi(b_{m,j}|A) = \mathcal{N}(\gamma_{im,j}, \delta_{m,j}) \text{ with}$$
$$\gamma_{m,j} = \sum_{i=1}^{K} a_{m,i}\theta_{i,j}, \text{ and } \delta_{m,j} = \alpha \sum_{i=1}^{K} a_{m,i}^2 \theta_{i,j}^2$$

VD with correlated weight noise

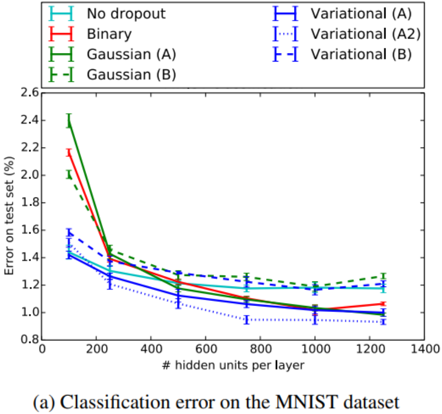
$$\mathbf{B} = (\mathbf{A} \circ \xi)\,\theta, \xi_{i,j} \sim \mathcal{N}(1, \alpha) \iff \mathbf{b}^m = \mathbf{a}^m \mathbf{W}, \text{ with}$$
$$\mathbf{W} = (\mathbf{w}_1^{`}, \ldots, \mathbf{w}_K^{`})^{`}, \text{ and } \mathbf{w}_i = s_i \theta_i,\ q_\phi(s_i) = \mathcal{N}(1, \alpha)$$

Variance comparison

stochastic gradient estimator	top layer 10 epochs	top layer 100 epochs	bottom layer 10 epochs	bottom layer 100 epochs
local reparameterization (ours)	7.8×10^3	1.2×10^3	1.9×10^2	1.1×10^2
separate weight samples (slow)	1.4×10^4	2.6×10^3	4.3×10^2	2.5×10^2
single weight sample (standard)	4.9×10^4	4.3×10^3	8.5×10^2	3.3×10^2
no dropout noise (minimal var.)	2.8×10^3	5.9×10^1	1.3×10^2	9.0×10^0

Figure: Variance comparison

Different graphics



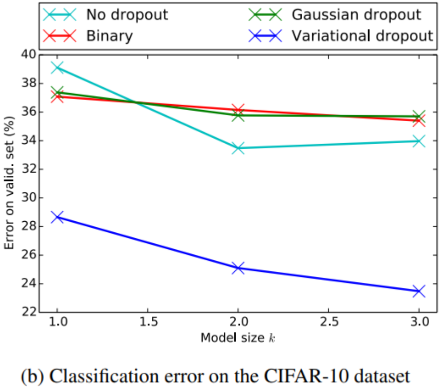


Figure: Different graphics

Literature

1. **Main article** Variational Dropout and the Local Reparameterization Trick.