Weight Uncertainty in Neural Networks

CHARLES BLUNDELL, JULIEN CORNEBISE, KORAY KAVUKCUOGLU, DAAN WIERSTRA

GOOGLE DEEPMIND

Jimyeong Kim

기존 Neural Network

- Overfitting
- Uncertainty를 반영하지 못한다.

- 새로운 Neural Network 제시

Bayes by Backprop(BBB)

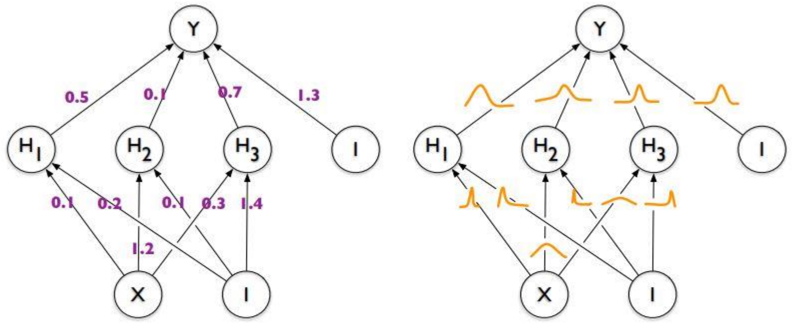


Figure 1. Left: each weight has a fixed value, as provided by classical backpropagation. Right: each weight is assigned a distribution, as provided by Bayes by Backprop.

- 각 weight는 하나의 값이 아닌 하나의 분포를 갖는다.
- Parameter 수를 2배로 늘리면서 무한개의 ensemble 효과를 얻을 수 있다.
- Data에 대한 uncertainty를 얻을 수 있다.

Maximum likelihood estimation(MLE)

$$w^{MLE} = \arg\max_w \log \boxed{P(D|w)}$$

$$= \arg\max_w \sum_{i=1}^{N} \log P(y_i|x_i, w) = \arg\max_w \sum_{i=1}^{N} log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_w(x_i))^2}{2\sigma^2}\right)$$

$$= \arg\max_w \sum_{i=1}^{N} \left(-\frac{(y_i - f_w(x_i))^2}{2\sigma^2}\right)$$

$$= \arg\min_w \sum_{i=1}^{N} (y_i - f_w(x_i))^2$$

Maximum A Posterior(MAP)

$$\boxed{P(w|D)} = \frac{P(D|w)P(w)}{\int P(D|w)P(w)dw}$$

$$w^{MAP} = \arg\max_{w} \log P(D|w) + \log P(w) \qquad (P(w) \sim N(0, \sigma_w^2))$$

$$= \arg\max_{w} \sum_{i=1}^{N} log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_w(x_i))^2}{2\sigma^2}\right) + log \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{(w)^2}{2\sigma_w^2}\right)$$

$$= \arg\max_{w} \sum_{i=1}^{N} \left(-\frac{(y_i - f_w(x_i))^2}{2\sigma^2}\right) + -\frac{(w)^2}{2\sigma_w^2}$$

$$= \arg\min_{w} \sum_{i=1}^{N} (y_i - f_w(x_i))^2 + w^2$$

Being Bayesian by Backpropagation

Bayesian Neural Network에서 사용하는 방법은 무엇인가?

- Posterior($P(w|D)$) 분포를 구해서 unknown X에 대한

  unknown label y의 기댓값을 구하자

  -> 무수히 많은 network 앙상블 시킨 효과

$$P(\hat{\mathbf{y}}|\hat{\mathbf{x}}) = \mathbb{E}_{P(\mathbf{w}|\mathcal{D})}[P(\hat{\mathbf{y}}|\hat{\mathbf{x}}, \mathbf{w})]$$

문제점

Posterior를 구하는 것이 intractable하다.

$$p(w|D) = \frac{p(D|w)p(w)}{\boxed{\int p(D|w)p(w)dw}}$$

→ w의 개수가 너무 많아서
모든 w에 대해 적분 불가능

Variational inference 사용

Variational inference $q(w|\theta) \sim p(w|D)$

$$
\begin{aligned}
\theta^\star &= \arg\min_\theta \mathrm{KL}[q(\mathbf{w}|\theta)||P(\mathbf{w}|\mathcal{D})] \\
&= \arg\min_\theta \int q(\mathbf{w}|\theta) \log \frac{q(\mathbf{w}|\theta)}{P(\mathbf{w})P(\mathcal{D}|\mathbf{w})} \mathrm{d}\mathbf{w} \\
&= \arg\min_\theta \mathrm{KL}\left[q(\mathbf{w}|\theta) \mid\mid P(\mathbf{w})\right] - \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}|\mathbf{w})\right]
\end{aligned}
$$

$\mathcal{F}(\mathcal{D},\theta) = \mathrm{KL}\left[q(\mathbf{w}|\theta) \mid\mid P(\mathbf{w})\right]$ <- Complexity cost

Likelihood cost -> $- \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}|\mathbf{w})\right].$

= ELBO/ variational free energy

$q(w|\theta)$ = variational inference $P(w)$: prior $P(D|w)$: likelihood

Optimization

Loss :

$$\mathcal{F}(\mathcal{D}, \theta) \approx \sum_{i=1}^{n} \log q(\mathbf{w}^{(i)}|\theta) - \log P(\mathbf{w}^{(i)}) - \log P(\mathcal{D}|\mathbf{w}^{(i)})$$

w^i = $i\ th$ Monte Carlo random sample from $q(w^i|\theta)$

$\theta = (\mu, \rho)$, $\sigma = \log(1 + \exp(\rho))$

$w = \mu + \sigma \circ \mathcal{E}$ $\mathcal{E} \sim N(0,1)$

$\circ$: pointwise multiplication

Optimization

Sample $\epsilon \sim \mathcal{N}(0, I)$.
Let $\mathbf{w} = \mu + \log(1 + \exp(\rho)) \circ \epsilon$.
Let $\theta = (\mu, \rho)$.
Let $f(\mathbf{w}, \theta) = \log q(\mathbf{w}|\theta) - \log P(\mathbf{w})P(\mathcal{D}|\mathbf{w})$

Update the variational parameters:

$$\mu \leftarrow \mu - \alpha\Delta_\mu$$

$$\rho \leftarrow \rho - \alpha\Delta_\rho.$$

Scale mixture prior

$$P(\mathbf{w}) = \prod_j \pi \mathcal{N}(\mathbf{w}_j | 0, \sigma_1^2) + (1 - \pi)\mathcal{N}(\mathbf{w}_j | 0, \sigma_2^2) \ ,$$

$\sigma 1 > \sigma 2$
$\sigma 2$ << 1

π, $\sigma 1, \sigma 2$: hyperparameter

$Gaussian$ 2개 더함 -> more uncertainty

Minibatches and KL re-weighting

M = batch size가 아닌 batch의 수

1. $$\mathcal{F}_i^{\mathrm{EQ}}(\mathcal{D}_i, \theta) = \frac{1}{M}\mathrm{KL}\left[q(\mathbf{w}|\theta) \mid\mid P(\mathbf{w})\right] - \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}_i|\mathbf{w})\right]$$

$$\sum_i \mathcal{F}_i^{\mathrm{EQ}}(\mathcal{D}_i, \theta) = \mathcal{F}(\mathcal{D}, \theta)$$

2. $$\mathcal{F}_i^{\pi}(\mathcal{D}_i, \theta) = \pi_i \mathrm{KL}\left[q(\mathbf{w}|\theta) \mid\mid P(\mathbf{w})\right] - \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}_i|\mathbf{w})\right]$$

$$\mathbb{E}_M[\sum_{i=1}^M \mathcal{F}_i^{\pi}(\mathcal{D}_i, \theta)] = \mathcal{F}(\mathcal{D}, \theta)$$

$\pi \in [0,1]^M$ $\quad$ $\boxed{\pi_i = \frac{2^{M-i}}{2^M - 1}}$ $\quad$ $\sum_{i=1}^M \pi_i = 1$

정리

- Posterior($P(w|D)$) 분포를 구해서 unknown X에 대한 unknown label y의 기댓값을 구하자
- Posterior 구하기 힘들다
  -> variational inference $q(w|\theta) \sim p(w|D)$
- $q(w|\theta)$ 와 $p(w|D)$사이 거리를 줄이자
- Sampling을 통해 loss 구하자
- Minibatch loss
- μ , σ update
- Prior: scale mixture prior

$$P(\hat{\mathbf{y}}|\hat{\mathbf{x}}) = \mathbb{E}_{P(\mathbf{w}|\mathcal{D})}[P(\hat{\mathbf{y}}|\hat{\mathbf{x}}, \mathbf{w})]$$

$$p(w|D) = \frac{p(D|w)p(w)}{\boxed{\int p(D|w)p(w)dw}}$$

$$\arg\min_{\theta} \mathrm{KL}[q(\mathbf{w}|\theta)||P(\mathbf{w}|\mathcal{D})]$$

$$w = \mu + \sigma \circ \varepsilon , \qquad \varepsilon \sim N(0,1)$$

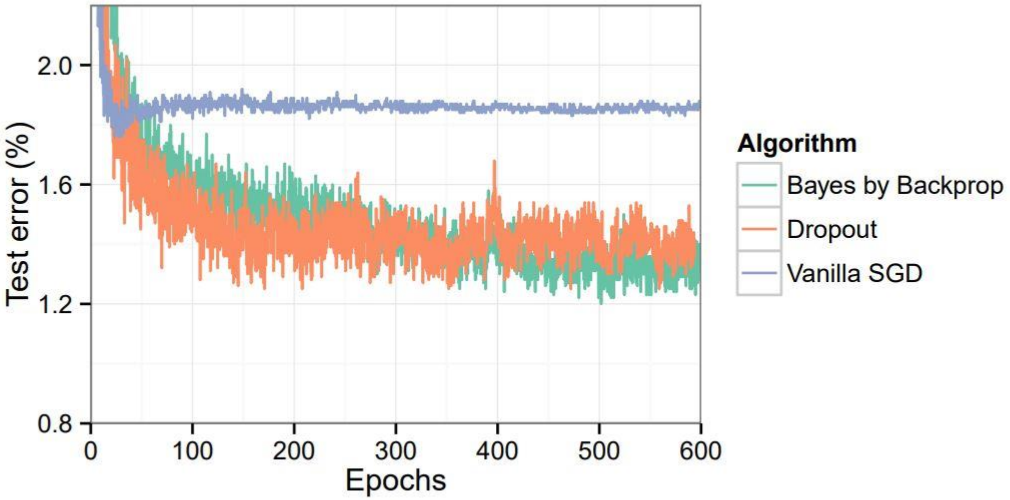
$$\mathcal{F}_i^{\pi}(\mathcal{D}_i, \theta) = \pi_i \mathrm{KL}\left[q(\mathbf{w}|\theta) \mid\mid P(\mathbf{w})\right] - \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}_i|\mathbf{w})\right]$$

$$P(\mathbf{w}) = \prod_j \pi \mathcal{N}(\mathbf{w}_j|0, \sigma_1^2) + (1-\pi)\mathcal{N}(\mathbf{w}_j|0, \sigma_2^2)$$

Experiments

Table 1. Classification Error Rates on MNIST. ⋆ indicates result used an ensemble of 5 networks.

Method	**# Units/Layer**	**# Weights**	**Test Error**
SGD, no regularisation (Simard et al., 2003)	800	1.3m	1.6%
SGD, dropout (Hinton et al., 2012)			≈ 1.3%
SGD, dropconnect (Wan et al., 2013)	800	1.3m	**1.2%**⋆
SGD	400	500k	1.83%
	800	1.3m	1.84%
	1200	2.4m	1.88%
SGD, dropout	400	500k	1.51%
	800	1.3m	1.33%
	1200	2.4m	1.36%
Bayes by Backprop, Gaussian	400	500k	1.82%
	800	1.3m	1.99%
	1200	2.4m	2.04%
Bayes by Backprop, Scale mixture	400	500k	1.36%
	800	1.3m	1.34%
	1200	2.4m	**1.32%**



Experiments

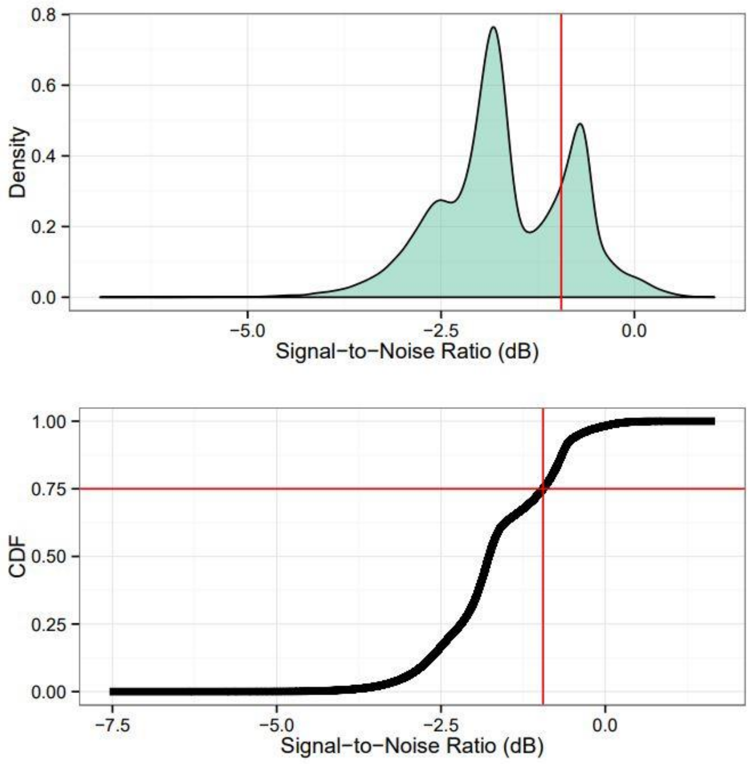


Figure 4. Density and CDF of the Signal-to-Noise ratio over all weights in the network. The red line denotes the 75% cut-off.

Table 2. Classification Errors after Weight pruning

Proportion removed	# Weights	Test Error
0%	2.4m	1.24%
50%	1.2m	1.24%
75%	600k	1.24%
95%	120k	1.29%
98%	48k	1.39%

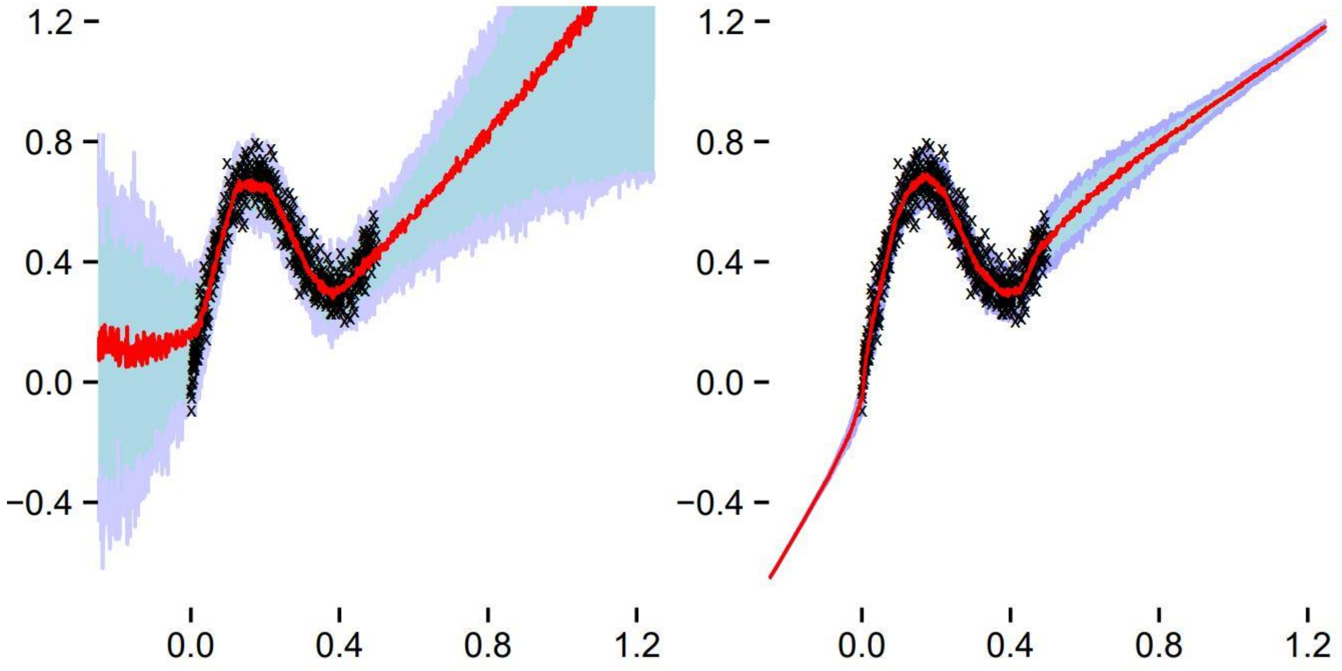
2 layers of 1200 units

Weight마다 signal-to-noise ratio ($|\mu i\ |/\sigma i$)를 계산해서 작은 것부터 지움 (posterior를 0으로 설정)

-> pruning을 통해 더 가벼운 모델로 만들 수 있음

Experiments

$$y = x + 0.3\sin(2\pi(x+\epsilon)) + 0.3\sin(4\pi(x+\epsilon)) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, 0.02)$$



왼쪽: Bayesian
오른쪽: Ordinary

Black: data
Red: median predictions
Blue/purple: interquartile range