Deep Weight Prior

Kotova Daria

May, 2019

Contents

Original research is available on:
`https://openreview.net/pdf?id=ByGuynAct7`

1. Variational inference
2. Reparametrisation trick
3. Variational autoencoder
4. Results
5. Conclusion

Variational Inference: Problem

$p(w)$ - prior knowledge of parameters w of the model
D - dataset, which should be described by the model
Goal: transform prior knowledge $p(w)$ to the posterior distribution $p(w|D)$ with Bayes rule:

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

The problem is $p(D)$ is intractable $\rightarrow$ $p(w|D)$ is intractable.
Moreover, we have no analytical expression for $p(w)$.

Variational lower bound

Firstly, let's approximate $p(w|D)$ with proposal distribution $q_\theta(w)$.
To make it we minimize KL-divergence:

$$D_{KL}(q_\theta(w)||p(w|D)) = \mathbb{E}_{q_\theta} log \frac{q_\theta(w)}{p(w|D)} \to \min_\theta$$

It is equivalent to **maximization of variational lower bound** of the marginal log-likelihood of the data :

$$VLB := \mathbb{E}_{q_\theta} \log p(D|w) - D_{KL}(q_\theta(w)||p(w)) \to \max_\theta$$

Reparametrisation trick

$$VLB = \mathbb{E}_{q_\theta} \log p(D|w) - D_{KL}(q_\theta(w)||p(w)) \to \max_\theta$$

$$\nabla_\theta \mathbb{E}_{q_\theta} \log p(D|w) = \int \nabla_\theta q_\theta(w) \log p(D|w) \neq \mathbb{E}_{q_\theta} \nabla_\theta \log p(D|w)$$

So we can not obtain unbiased gradients and perform mini-batch training.

Reparametrisation trick

Idea:
Let's represent q_θ as deterministic differentiable function $f(\theta, \epsilon)$.
For example:

$$\epsilon \sim \mathcal{N}(0, 1) \quad \xi \sim \mathcal{N}(\mu, \sigma^2)$$

$$\epsilon = \frac{\xi - \mu}{\sigma} \quad \xi = \mu + \sigma \cdot \epsilon$$

Now we can compute gradients of VLB, where instead of q_θ now is $f(\theta, \epsilon)$, since we take math expectation over ϵ but not q_θ.

Variational lower bound

To make it we minimize KL-divergence:

$$D_{KL}(q_\theta(w)||p(w|D)) = \mathbb{E}_{q_\theta} log \frac{q_\theta(w)}{p(w|D)} \to \min_\theta$$

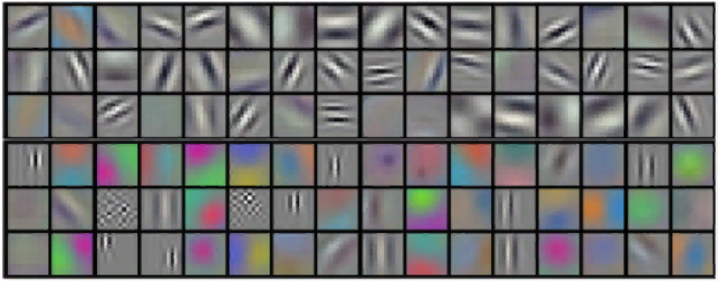
It is equivalent to **variational lower bound maximization**:

$$VLB = \mathbb{E}_{q_\theta} \log p(D|w) - D_{KL}(q_\theta(w)||p(w)) \to \max_\theta$$

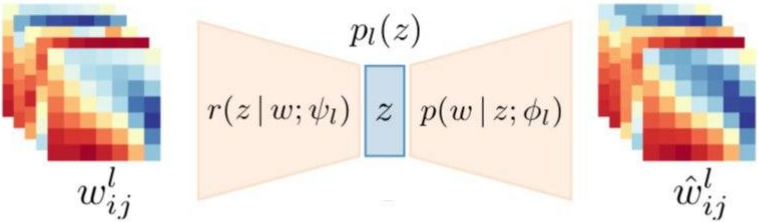
The second problem is that we do not have analytical expression for $p(w)$.

Prior distribution

We suggest that networks learned on similar datasets will have similar kernel's structures in convolutional layers.
That means that we can use kernels from already learned networks to approximate prior distribution $p(w)$.

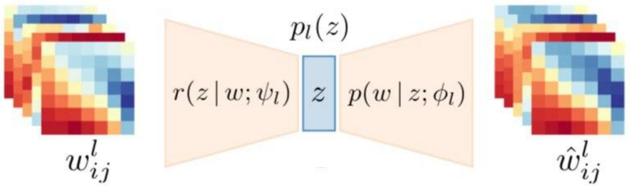


Variational autoencoder



VAE helps us to build upper bound for KL-divergence $D_{KL}(q_\theta(w)||p(w))$.

Variational inference with implicit prior distribution



VAE helps to make auxilary bound for VLB, which can be optimized unlike VLB:

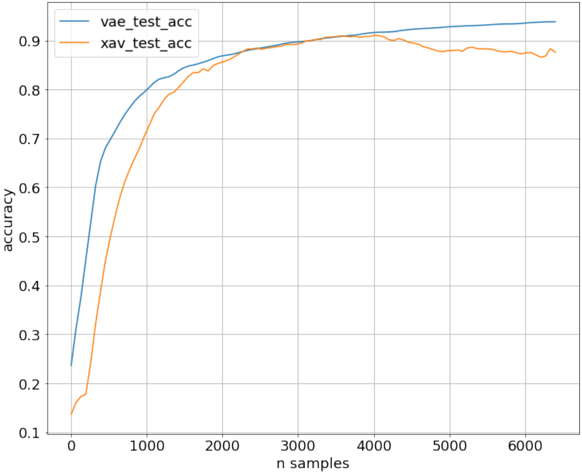
$$VLB = \mathbb{E}_{q_\theta} \log p(D|w) - D_{KL}(q_\theta(w)||p(w)) =$$
$$= L^{aux} + \mathbb{E}_{q(w)} D_{KL}(r(z|w)||p(z|w)) \geq L^{aux}$$

Results

Code and different examples are available on:
`https://github.com/DahaKot/Deep-Weight-Prior`

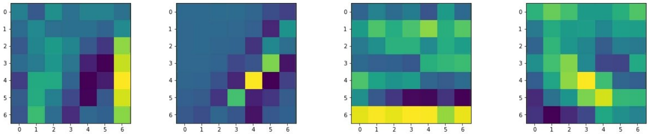
- trained 100 CNNs on notMNIST
- trained vae on source kernels
- used samples from vae for CNN learned on MNIST
- compared performance of CNNs with different initializations

The performance of convolutional network with two different priors: deep weight prior (dwp) and xavier:

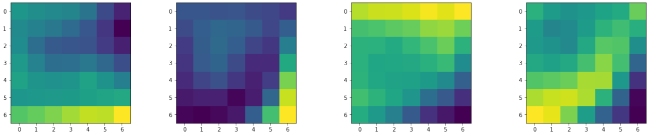


Kernel examples

There are samples of the source kernels:



And kernels got from vae:



Conclusion

We considered article deep weight prior, which proposes modification of variational inference with implicit prior distribution $p(w)$:

- Modify variational lower bound
- To perform mini-batch training use reparametrisation trick
- Perform better than 'default' random initialization