Deep Weight Prior

Andrei Atanov*, Arsenii Ashukha*,
Kirill Struminsky, Dmitry Vetrov, Max Welling

Bayesian Inference

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Likelihood $p(y_i \mid x_i, W)$

Pior $p(W)$

Bayes rule $$p(W \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid W)p(W)}{p(\mathcal{D})}$$

Variatonal Inference

Dataset	$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$
Likelihood	$p(y_i \mid x_i, W)$
Pior	$p(W)$
Variational posterior	$q_\theta(W)$

$$D_{\mathrm{KL}}(q_\theta(W) \| p(W \mid \mathcal{D})) \to \min_\theta \Leftrightarrow \mathcal{L}(\theta) \to \max_\theta$$

ELBO $$\mathcal{L}(\theta) = \sum_{i=1}^N \mathbb{E}_{q_\theta(W)} \log p(y_i \mid x_i, W) - D_{\mathrm{KL}}(q_\theta(W) \| p(W))$$

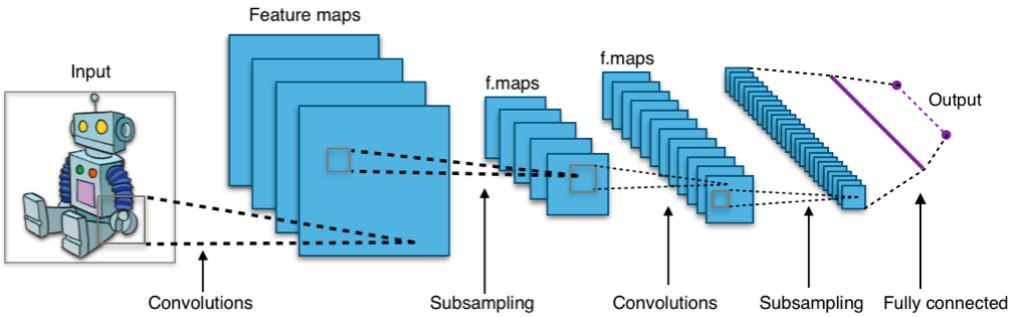
Prior distribution

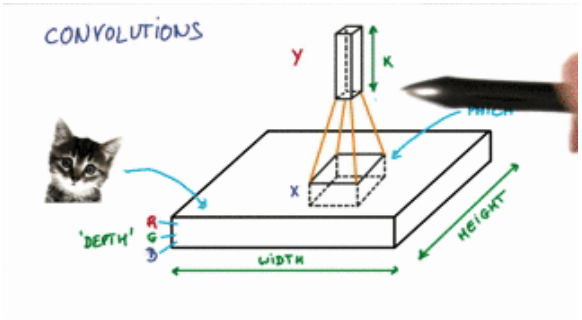
Incorporate prior knowledge or specific properties into a model:

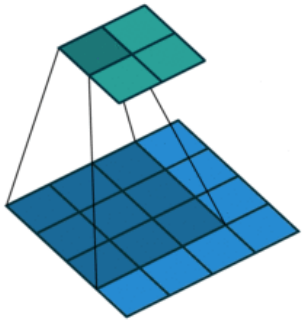
- Sparsification (Molchanov et al. 2017; Neklyudov et al. 2017)
- Quantization (Ullrich et al., 2017)
- Compression (Louizos et al., 2017)

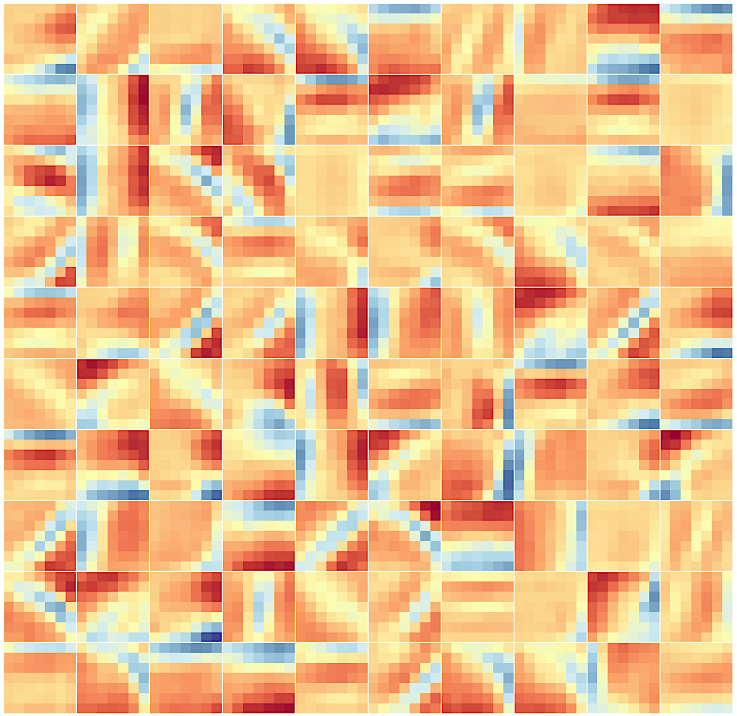
- Usually fully-factorized
- Does not take into account intrinsic structure of the weight

Convolutional Neural Networks



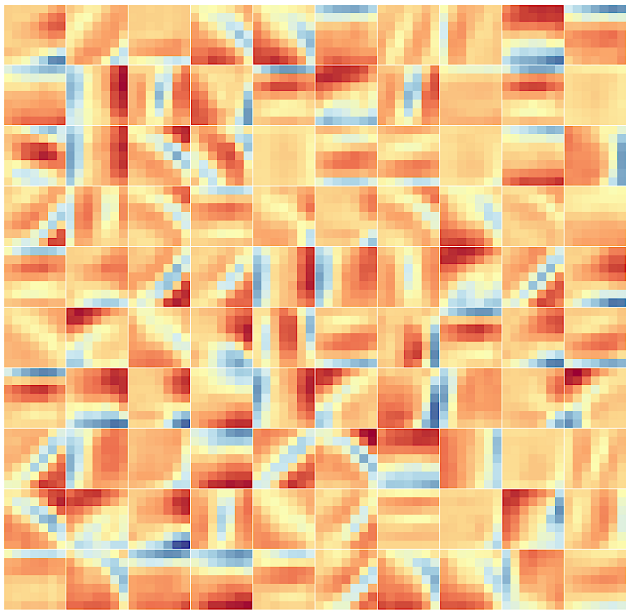




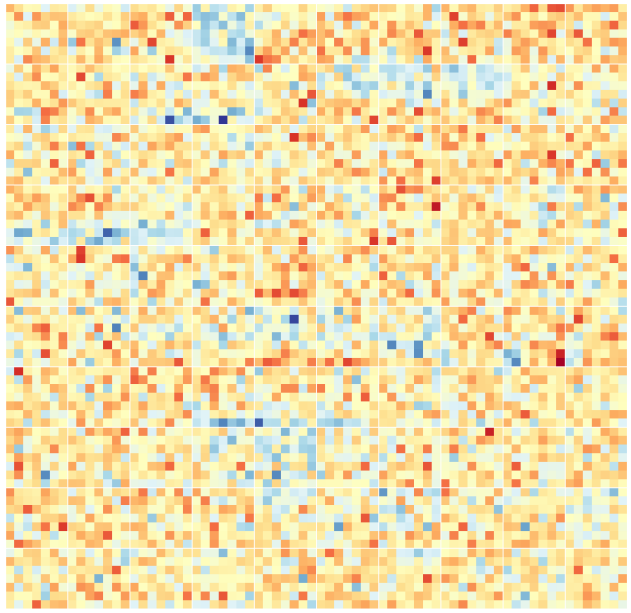


Learned convolutional kernels

Convolutional Neural Networks



Kernels learned on
large dataset



Kernels learned on
small dataset

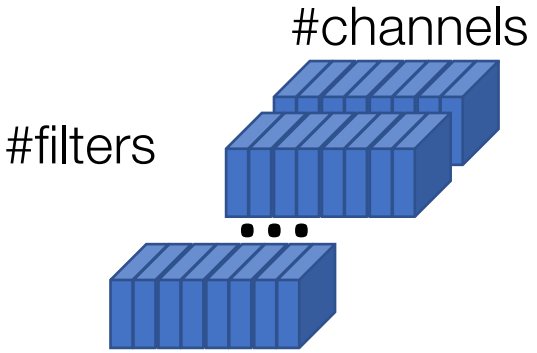
Deep Weight Prior

Consider prior factorized over layers, channels and filters:

#layers #channels #filters

$$p(W) = \prod_{l=1}^{L} \prod_{i=1}^{I_l} \prod_{j=1}^{O_l} p_l(w_{ij}^l)$$

kernel

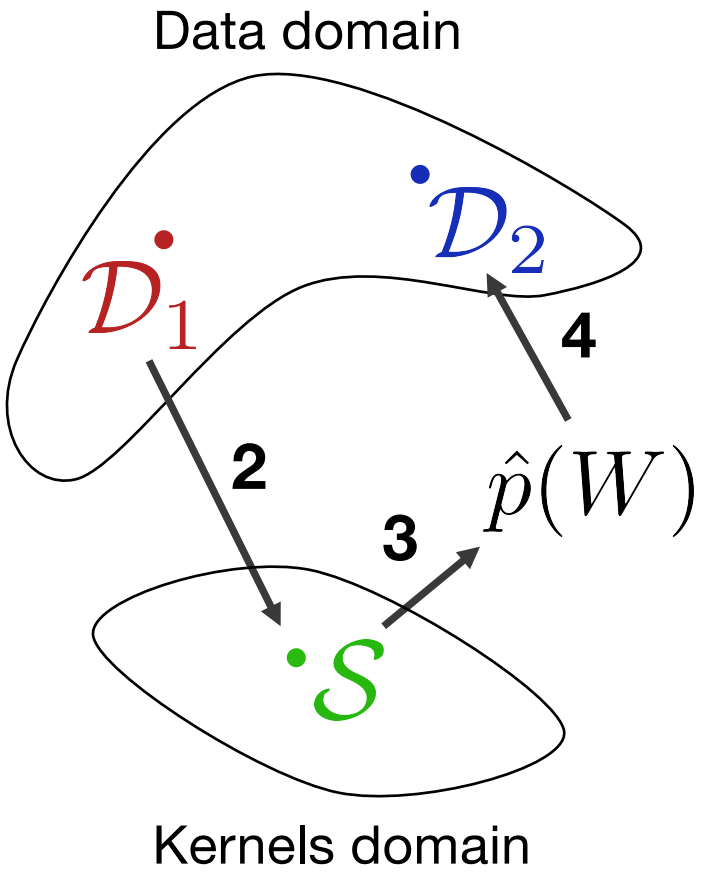


How to develop distribution that will favor learned kernels?

Generative models!

Learning Deep Weight Prior

1. Train CNNs on a given source dataset $\mathcal{D}_1$
2. Collect datasets $\mathcal{S}$ of learned kernels
3. Train prior $\hat{p}(W)$ distributions on $\mathcal{S}$
4. Use this distribution for Bayesian inference on a new dataset $\mathcal{D}_2$



Generative models

Explicit models:

$$p(w)$$

- Kernel Density Estimation
- Normalizing Flows
- PixelCNN\RNN

\+ Access to the density function

\- High memory\computational cost

Implicit models:

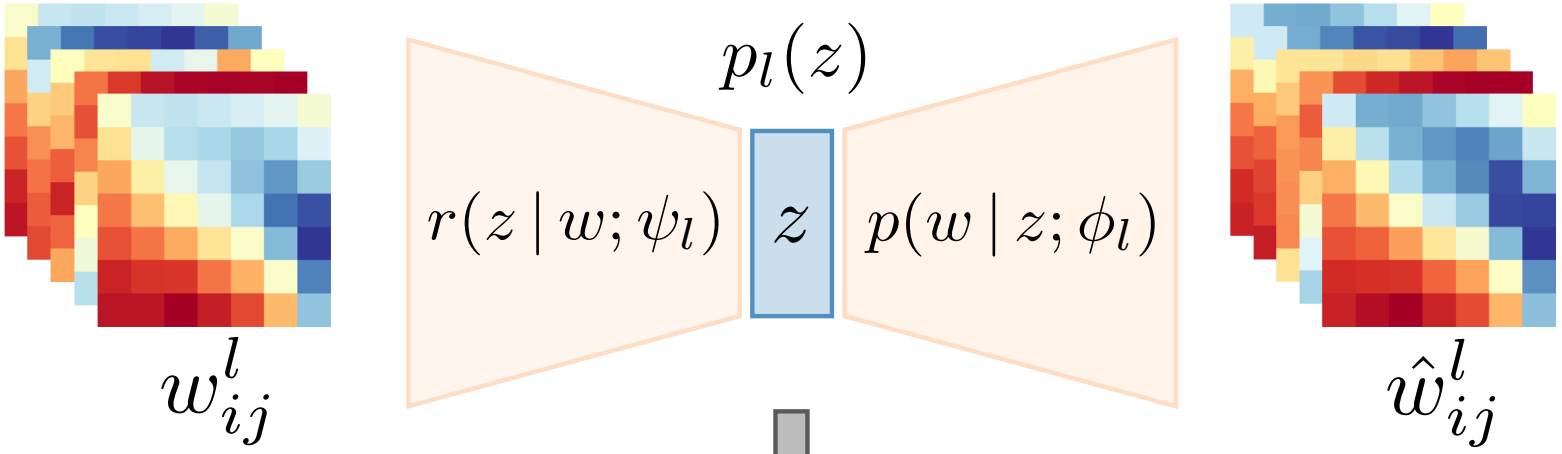
$$p(w) = \int p(w \mid z) p(z) dz$$

- Variational auto-encoder

\+ Memory and computational efficient

\- No explicit form of the density

Learning Deep Weight Prior



$$\hat{p}_l(w) = \int p(w \mid z; \phi_l) p_l(z)\, dz$$

Make ELBO tractable again

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \mathbb{E}_{q_\theta(W)} \log p(y_i \mid x_i, W) - D_{\mathrm{KL}}(q_\theta(W) \| p(W))$$

Does not depend on a prior

No more tractable for DWP!

KL-divergence for one kernel of l-th layer:

$$D_{\mathrm{KL}}(q(w|\theta_{ij}^l) \| \hat{p}_l(w))$$

$$= -H(q(w \mid \theta_{ij}^l)) - \mathbb{E}_{q(w \mid \theta_{ij}^l)} \log \mathbb{E}_{p_l(z)} p(w \mid z; \phi_l)$$

$$= \log \hat{p}_l(w)$$

Make ELBO tractable again

$$
\begin{aligned}
& D_{\mathrm{KL}}(q(w) \| \hat{p}_l(w)) \\
&= -H(q(w)) - \mathbb{E}_{q(w)} \log \hat{p}_l(w) \\
&= -H(q(w)) - \mathbb{E}_{q(w)} \mathbb{E}_{r(z|w)} \log \frac{p(w \mid z; \phi_l) p_l(z)}{p(z|w)} \frac{r(z \mid w)}{r(z \mid w)} \\
&= -H(q(w)) + \mathbb{E}_{q(w)} \left[D_{\mathrm{KL}}(r(z|w) \| p_l(z)) - \mathbb{E}_{r(z|w)} \log p(w|z; \phi_l) \right] \\
&\qquad\qquad - \mathbb{E}_{q(w)} D_{\mathrm{KL}}(r(z|w) \| p(z|w))
\end{aligned}
$$

Make ELBO tractable again

$$D_{\mathrm{KL}}(q(w)\|\hat{p}_l(w))$$

$$= -H(q(w)) - \mathbb{E}_{q(w)} \log \hat{p}_l(w)$$

$$= -H(q(w)) - \mathbb{E}_{q(w)}\mathbb{E}_{r(z|w)} \log \frac{p(w \mid z;\phi_l)p_l(z)}{p(z|w)} \frac{r(z \mid w)}{r(z \mid w)}$$

$$= -H(q(w)) + \mathbb{E}_{q(w)} \left[D_{\mathrm{KL}}(r(z|w)\|p_l(z)) - \mathbb{E}_{r(z|w)} \log p(w|z;\phi_l)\right]$$

KL upper bound

$$-\mathbb{E}_{q(w)} D_{\mathrm{KL}}(r(z|w)\|p(z|w))$$

<= 0

Make ELBO tractable again

- **Upper bound on the whole KL-term:**

$$D_{\mathrm{KL}}(q(W)\|\hat{p}(W)) = \sum_{l,i,j} D_{\mathrm{KL}}(q(w_{ij}^l|\theta_{ij}^l)\|\hat{p}_l(w_{ij}^l)) \leq \sum_{l,i,j} \big(- H(q(w_{ij}^l \mid \theta_{ij}^l)) +$$

$$+\mathbb{E}_{q(w_{ij}^l \mid \theta_{ij}^l)} \big[D_{\mathrm{KL}}(r(z \mid w_{ij}^l; \psi_l)\|p_l(z)) - \mathbb{E}_{r(z \mid w_{ij}^l;\psi_l)} \log p(w_{ij}^l \mid z; \phi_l) \big] \big) = D_{\mathrm{KL}}^{bound}$$

- **Final auxiliary lower bound:**

$$\mathcal{L}^{aux}(\theta, \psi) = L_D - D_{\mathrm{KL}}^{bound} \leq L_D - D_{\mathrm{KL}}(q_\theta(W)\|\hat{p}(W)) = \mathcal{L}(\theta)$$

$$\mathcal{L}(\theta, \psi) \to \max_{\theta, \psi}$$

Make ELBO tractable again

- **Upper bound on the whole KL-term:**

$$D_{\mathrm{KL}}(q(W)\|\hat{p}(W)) = \sum_{l,i,j} D_{\mathrm{KL}}(q(w_{ij}^l|\theta_{ij}^l)\|\hat{p}_l(w_{ij}^l)) \leq \sum_{l,i,j} \big(- H(q(w_{ij}^l \,|\, \theta_{ij}^l)) +$$

$$+\mathbb{E}_{q(w_{ij}^l \,|\, \theta_{ij}^l)} \big[D_{\mathrm{KL}}(r(z \,|\, w_{ij}^l; \psi_l)\|p_l(z)) - \mathbb{E}_{r(z \,|\, w_{ij}^l; \psi_l)} \log p(w_{ij}^l \,|\, z; \phi_l) \big] \big) = D_{\mathrm{KL}}^{bound}$$

- **Final auxiliary lower bound:**

$$\mathcal{L}^{aux}(\theta, \psi) = L_D - D_{\mathrm{KL}}^{bound} \leq L_D - D_{\mathrm{KL}}(q_\theta(W)\|\hat{p}(W)) = \mathcal{L}(\theta)$$

$$\mathcal{L}(\theta, \psi) \to \max_{\theta, \psi}$$

ϕ **is fixed!**

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$
Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$
  while not converged **do**
    $\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$
    $\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization
    $\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization
    $\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$
    Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$
    Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm
  end while
  return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$

Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$

while not converged **do**

$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$

$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization

$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization

$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$

Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$

Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm

end while

return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$

Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$

while not converged **do**

$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$

$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization

$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization

$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$

Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$

Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm

end while

return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$

Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$

while not converged **do**

$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$

$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization

$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization

$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$

Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$

Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm

end while

return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$
Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$
while not converged **do**
$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$
$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization
$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization
$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$
Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$
Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm
end while
return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$
Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$
while not converged **do**
$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$
$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization
$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization
$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$
Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$
Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm
end while
return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$
Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$
while not converged **do**
$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$
$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization
$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization
$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$
Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$
Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm
end while
return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$

Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$

while not converged **do**

$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$

$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization

$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization

$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$

Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$

Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm

end while

return Parameters θ, ψ

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$

Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$

Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$

while not converged **do**

$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$

$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization

$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization

$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} \boxed{-\log q(\hat{w}_{ij}^l \mid \theta_{ij}^l)} - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$

Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$

Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm

end while

return Parameters θ, ψ

$$\boxed{-H(q(w))} + \mathbb{E}_{q(w)} \left[D_{\mathrm{KL}}(r(z|w) \| p_l(z)) - \mathbb{E}_{r(z|w)} \log p(w|z; \phi_l) \right]$$

KL upper bound

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$
Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$
while not converged **do**
$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$
$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization
$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization
$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$
Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$
Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm
end while
return Parameters θ, ψ

$$-H(q(w)) + \mathbb{E}_{q(w)} \left[D_{\mathrm{KL}}(r(z|w) \| p_l(z)) - \mathbb{E}_{r(z|w)} \log p(w|z; \phi_l) \right]$$

KL upper bound

Algorithm 1 Stochastic Variational Inference With Implicit Prior Distribution

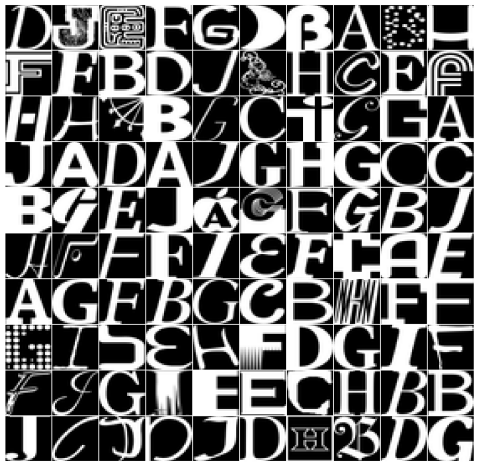
Require: dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
Require: variational approximations $q(w \mid \theta_{ij}^l)$ and reverse models $r(z \mid w; \psi_l)$
Require: reconstruction models $p(w \mid z; \phi_l)$, priors for auxiliary variables $p_l(z)$
while not converged **do**
$\hat{M} \leftarrow$ mini-batch of objects form dataset $\mathcal{D}$
$\hat{w}_{ij}^l \leftarrow$ sample weights from $q(w|\theta_{ij}^l)$ with reparametrization
$\hat{z}_{ij}^l \leftarrow$ sample auxiliary variables from $r(z \mid \hat{w}_{ij}^l; \psi_l)$ with reparametrization
$\hat{\mathcal{L}}^{aux} \leftarrow L_{\hat{M}} + \sum_{l,i,j} - \log q(\hat{w}_{ij}^l \mid \theta_{ij}^l) - \log r(\hat{z}_{ij}^l \mid \hat{w}_{ij}^l; \psi_l) + \log p_l(\hat{z}_{ij}^l) + \log p(\hat{w}_{ij}^l \mid \hat{z}_{ij}^l; \phi_l)$
Obtain unbiased estimate $\hat{g}$ with $\mathbb{E}[\hat{g}] = \nabla \mathcal{L}^{aux}$ by differentiating $\hat{\mathcal{L}}^{aux}$
Update parameters θ and ψ using gradient $\hat{g}$ and a stochastic optimization algorithm
end while
return Parameters θ, ψ

$$-H(q(w)) + \mathbb{E}_{q(w)} \left[D_{\mathrm{KL}}(r(z|w) \| p_l(z)) - \mathbb{E}_{r(z|w)} \log p(w|z; \phi_l) \right]$$

KL upper bound

Experiments

Train prior on:



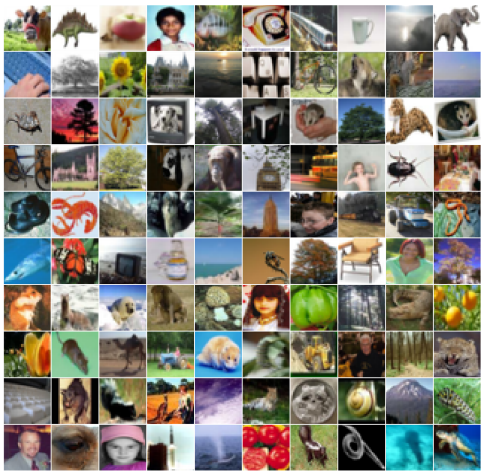
$\mathcal{D}_1$ **notMNIST**

Use prior for:

$\mathcal{D}_2$ **small MNIST**

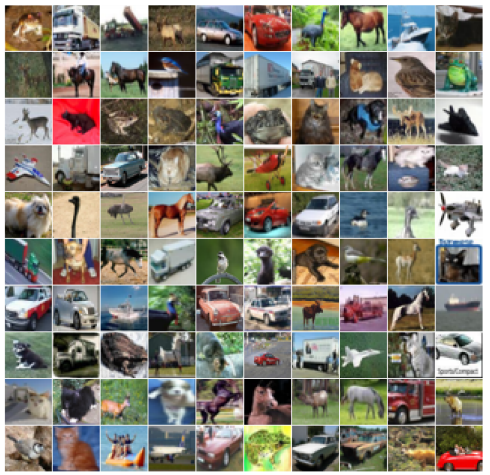
$\hat{p}(W)$

Train prior on:



$\mathcal{D}_1$ **CIFAR100**

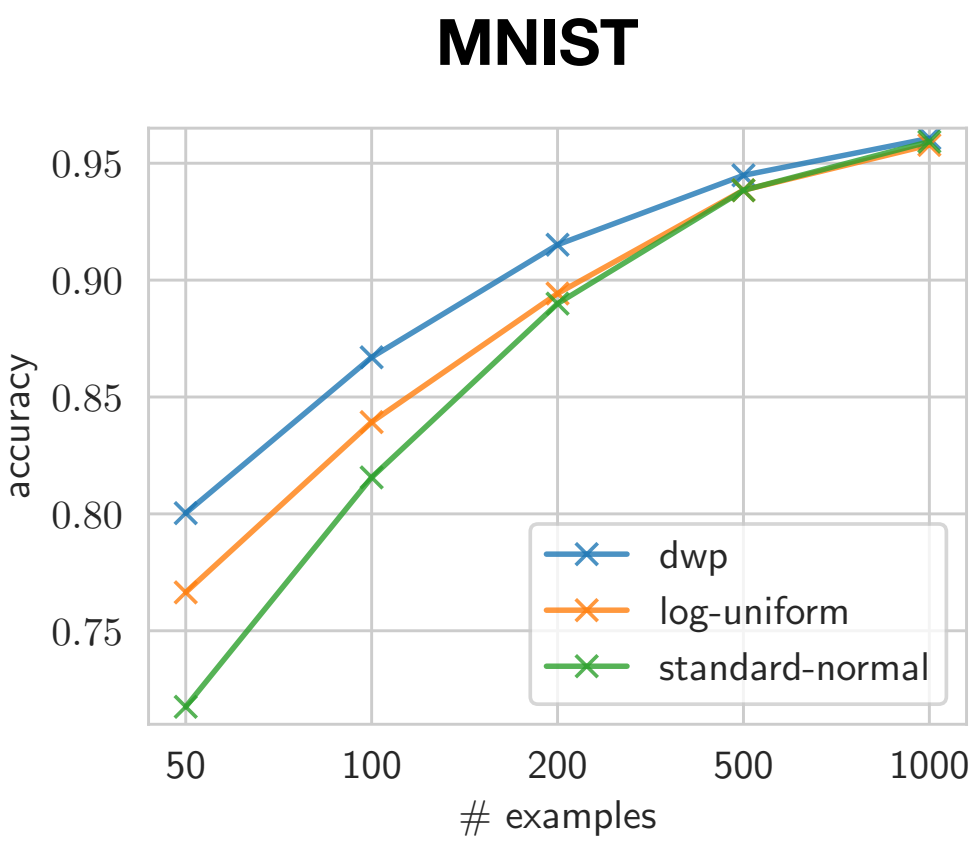
Use prior for:

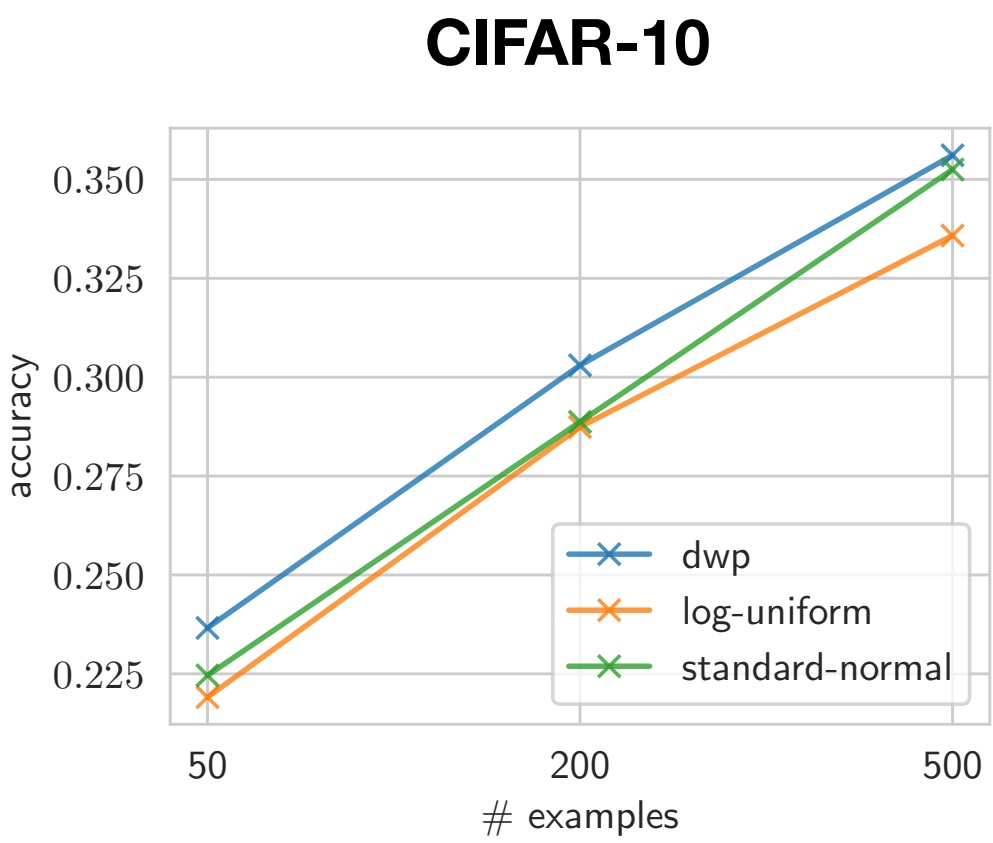


$\mathcal{D}_2$ **small CIFAR10**

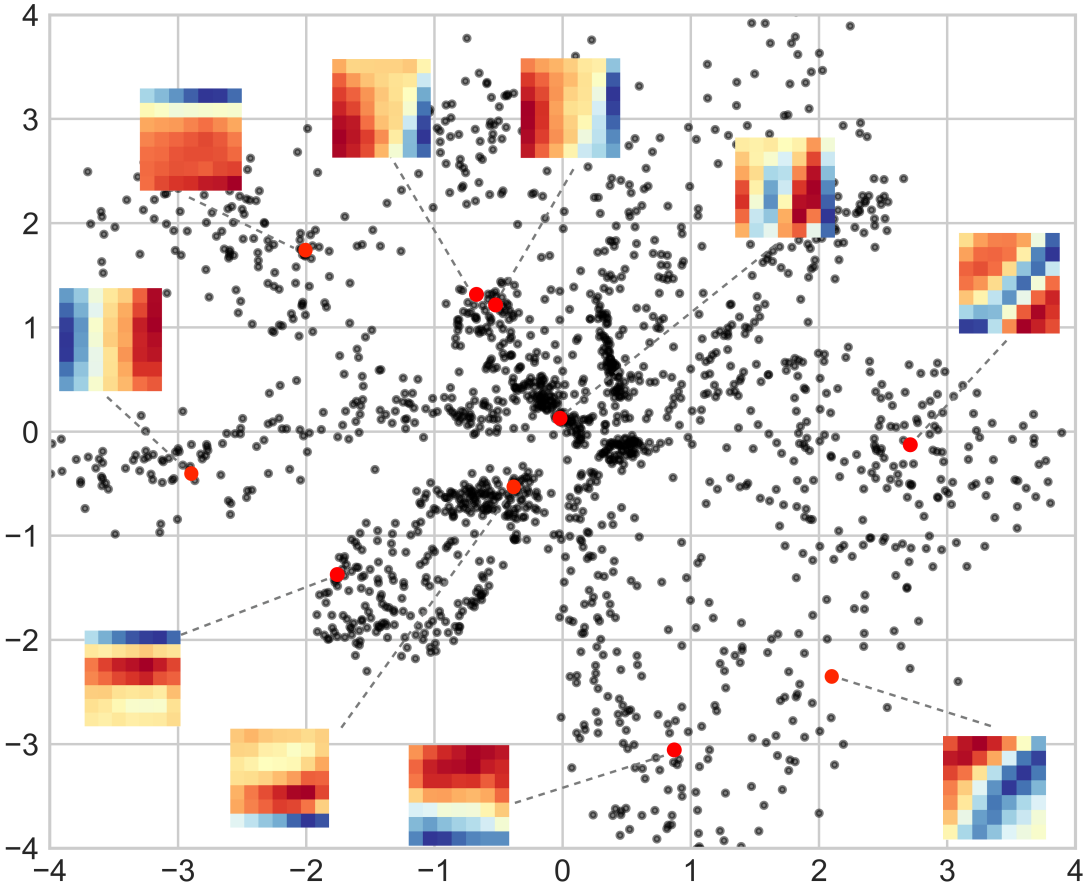
$\hat{p}(W)$

Classification

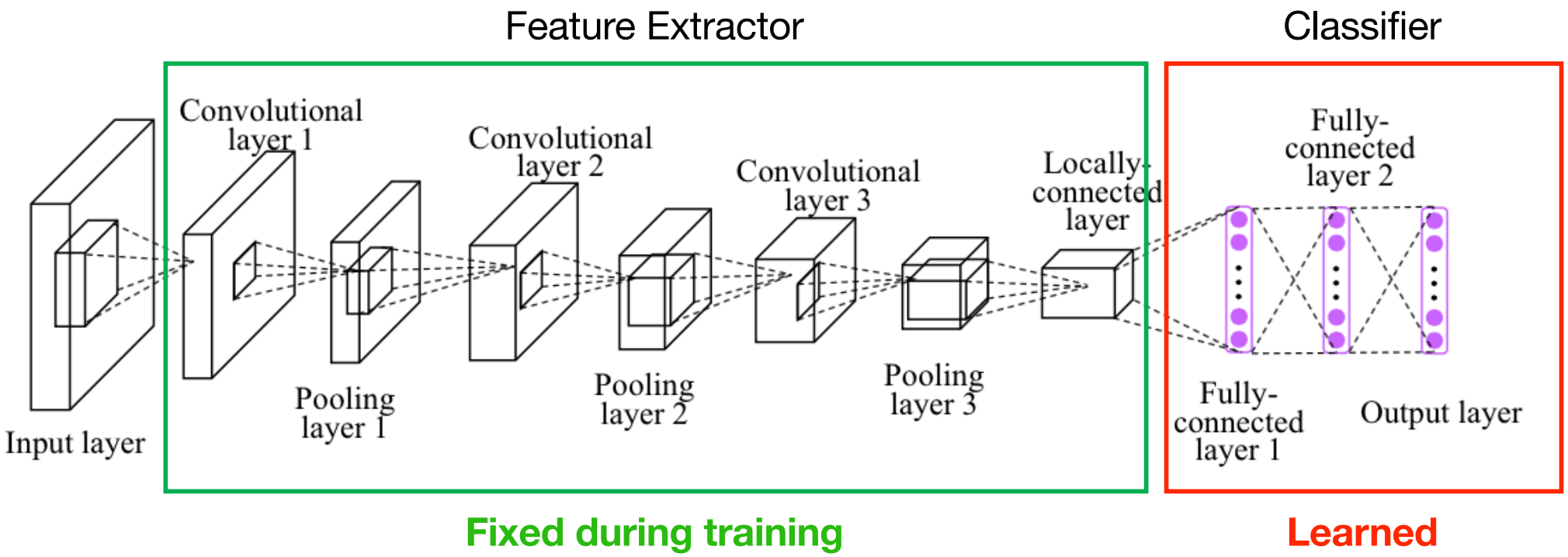




DWP Visualization

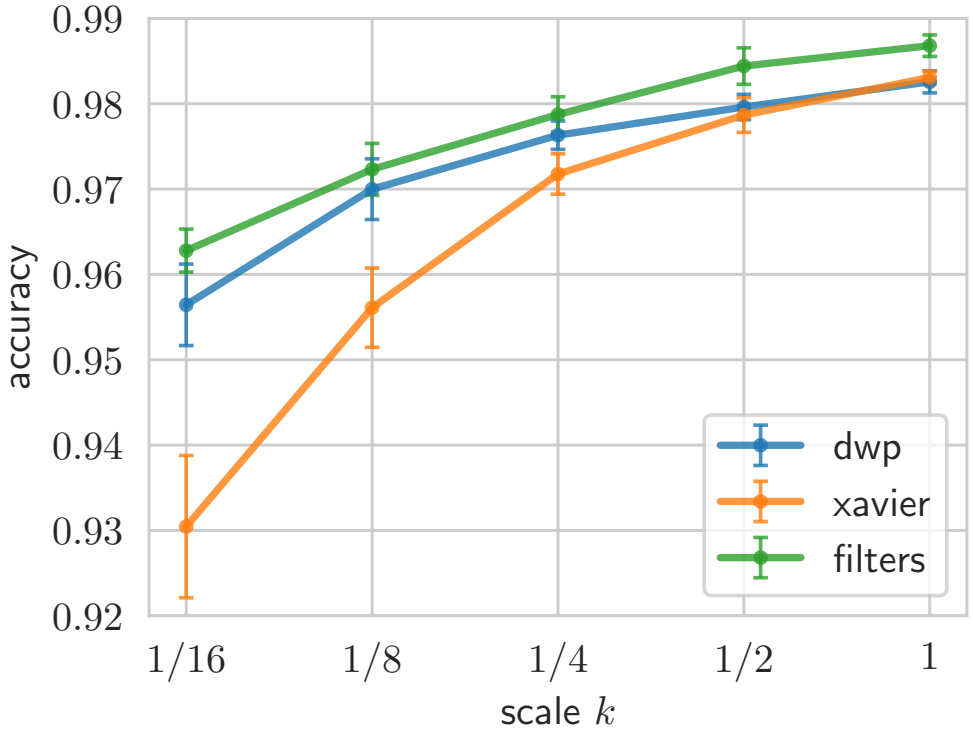


Random Feature Extraction

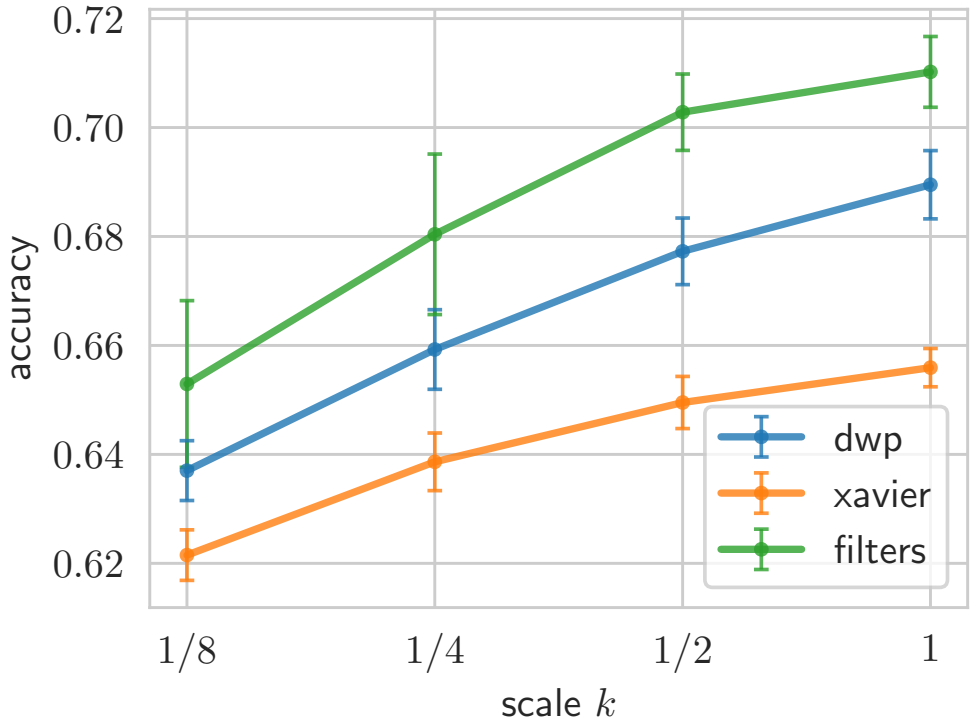


Random Feature Extraction

MNIST

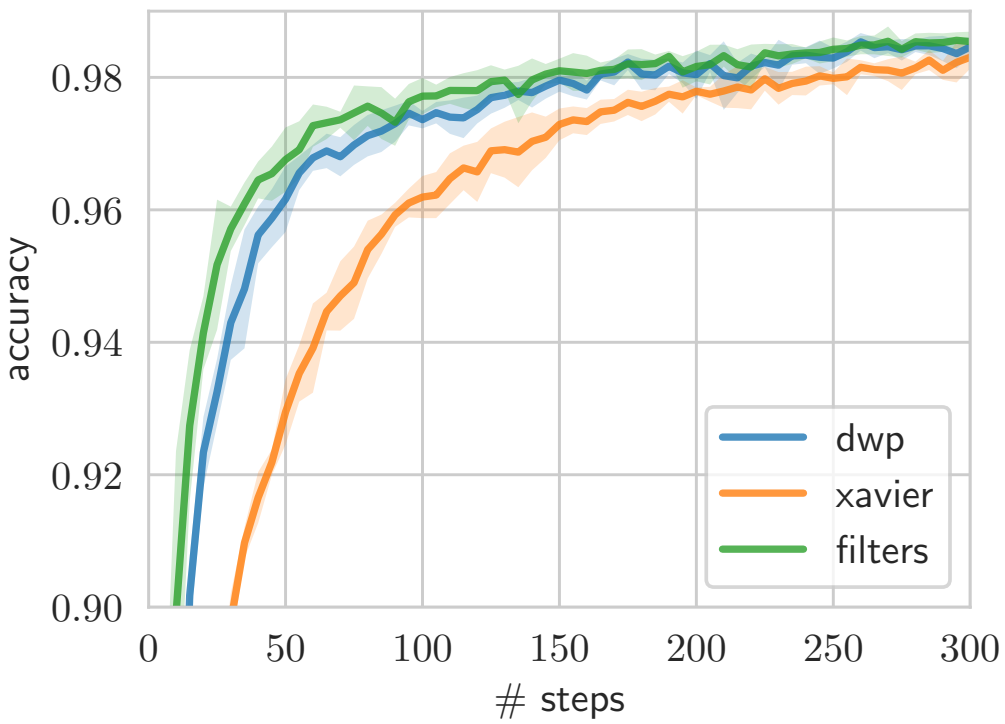


CIFAR-10

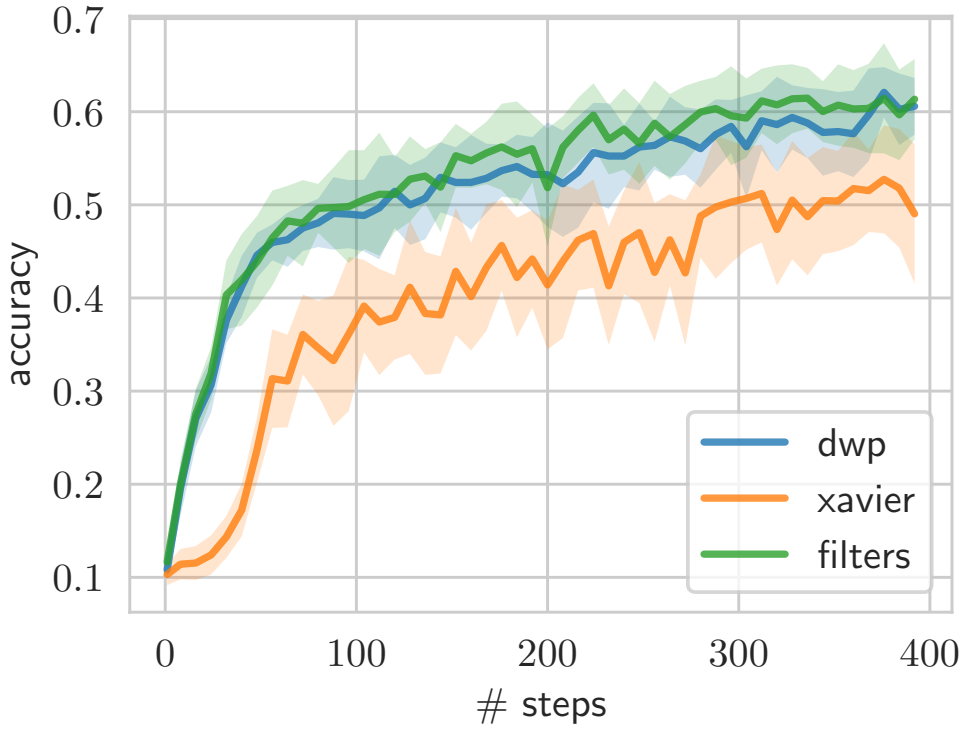


Fast Convergence

MNIST



CIFAR-10



Deep Weight Prior

- Propose the expressive prior distribution capable to encode interdependencies of the learned convolutional filters.
- Develop the method for variational inference with the proposed implicit prior distribution.