# The Deep Weight Prior


SAMSUNG AI Center - Moscow
UNIVERSITY OF AMSTERDAM
p(B|A)yesgroup.ru

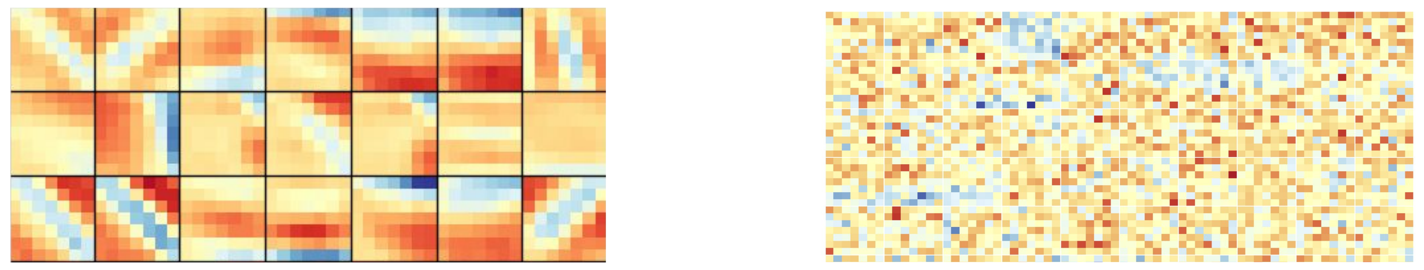
Andrei Atanov*, Arsenii Ashukha*, Kirill Struminsky, Dmitry Vetrov, Max Welling


Paper bit.ly/2VgDQiJ

Code bit.ly/2LllxnV

## Motivation

Kernels learned on large and small datasets:



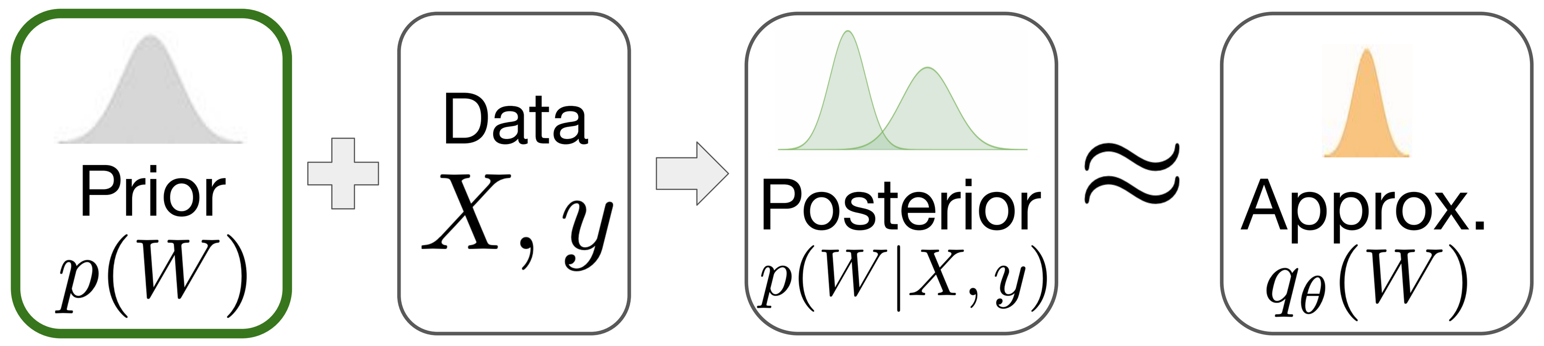

How to construct a **prior** that will favor the **specific structure** of learned kernels?

## Contributions

Propose a ***Deep Weight Prior*** that:

- **Favors** the structure of learned convolution kernels
- **Allows** learning hierarchical prior with a stochastic VI
- **Improves** few-shot classification performance

## Bayesian Neural Networks




Aims to approximate $p(W|X,y)$ via minimization:

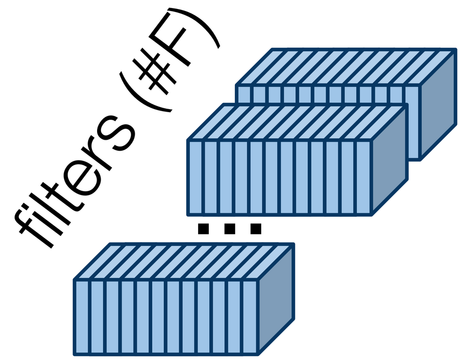
$$KL(q_\theta(W)\,||\,p(W\,|\,X,Y)) \to \min_\theta$$

**Variational Inference** reduces the problem to maximization of *variational lower bound* (vlb):

$$\mathcal{L}(\theta) = L_D - KL(q_\theta(W)\,||\,p(W)) \to \max_\theta$$

$$L_D = \mathbb{E}_{q_\theta(W)} \log p(Y\,|\,X,W)$$

## Deep Weight Prior




Kernels $w^l_{ij} \in \mathbb{R}^{k\times k}$

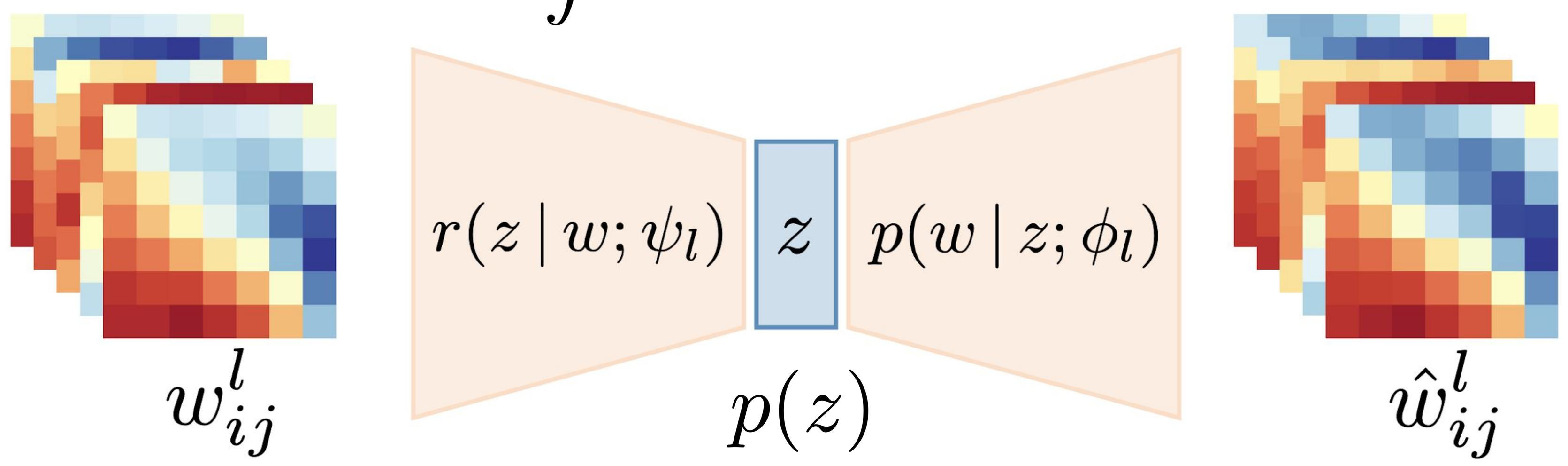
We share the same prior over all kernels within the layer:

$$p(w^l) = \prod_{i=1}^{\#F}\prod_{j=1}^{\#C} p_l(w^l_{ij})$$

How to find a distribution $p_l(w)$ that has a high density for **kernels of learned CNNs**?

Let's use generative models (VAE)!

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




## Variational Inference for Hierarchical Prior

$$KL(q_\theta(w^l_{ij})\,||\,\hat{p}_l(w^l_{ij})) = -H(q_\theta) + \mathbb{E}_{q_\theta}\log\hat{p}_l(w^l_{ij})$$

Intractable

**Upper bound** the intractable term:

$$\mathbb{E}_{q_\theta}\log\hat{p}_l(w^l_{ij}) \le \mathbb{E}_{q_\theta}[KL(r_l(z\,|\,w^l_{ij},\psi_l)||p(z)) - \mathbb{E}_{r_{\psi_l}}\log p(w^l_{ij}|z,\phi_l)]$$

Reverse Model

Learned part

Construct **an auxiliary lower bound**:

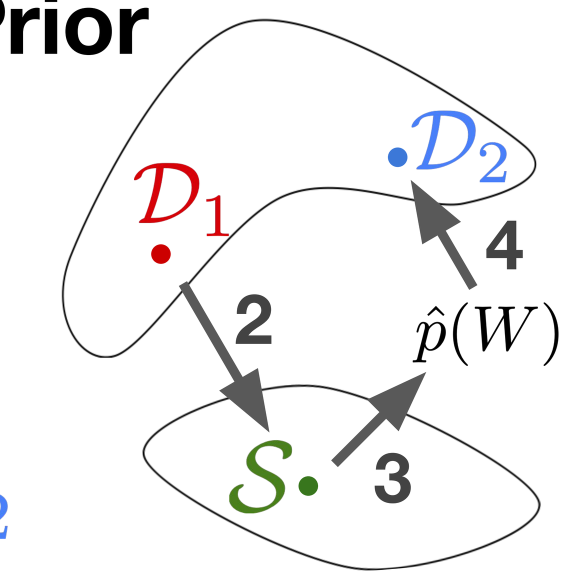
$$\mathcal{L}(\theta) = L_D + H(q_\theta) - \sum_{l,i,j}\mathbb{E}_{q_\theta}\log p^l_\phi(w^l_{ij}) \ge$$

$$\ge L_D + H(q_\theta) - \sum_{l,i,j}\mathbb{E}_{q_\theta}[KL(r_l(z\,|\,w^l_{ij},\psi_l)||p(z)) -$$

$$- \mathbb{E}_{r_{\psi_l}}\log p(w^l_{ij}|z,\phi_l)] = \mathcal{L}^{aux}(\theta,\psi) \to \max_{\theta,\psi}$$

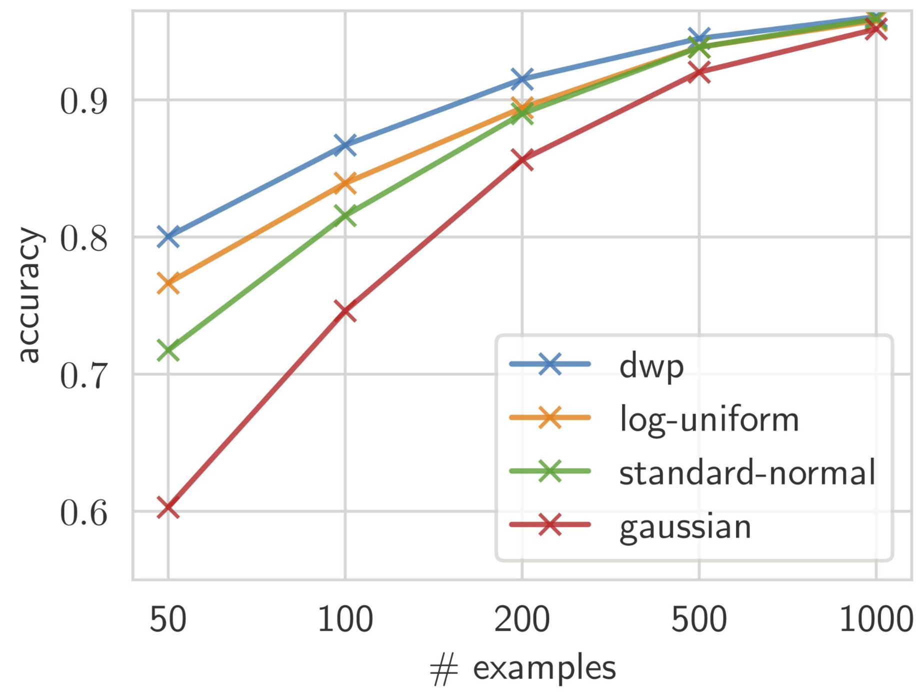
## Learning Deep Weight Prior

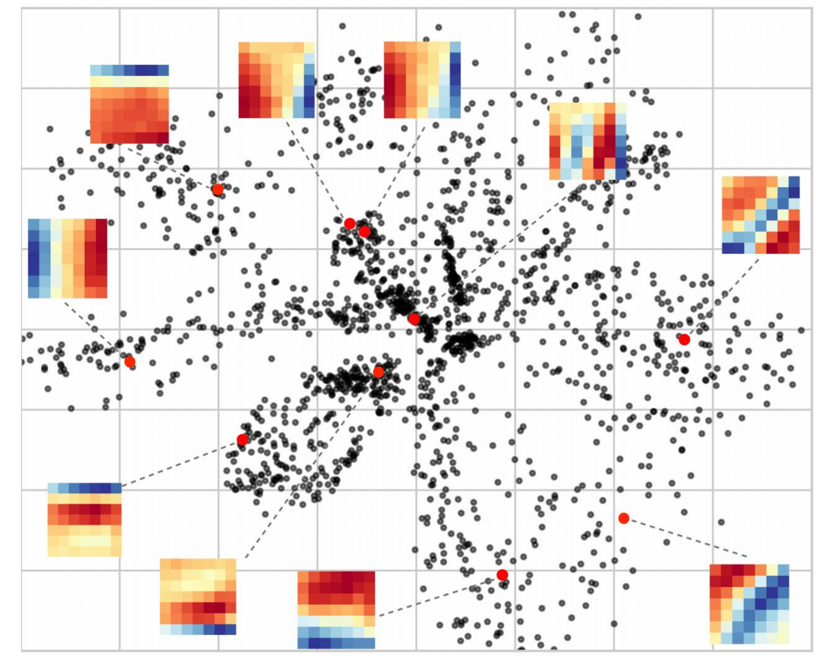
1. Train CNNs on large source $\mathcal{D}_1$
2. Collect dataset $\mathcal{S}$ of kernels
3. Train dwp $\hat{p}(W)$ using $\mathcal{S}$
4. Use the prior for VI on a small $\mathcal{D}_2$

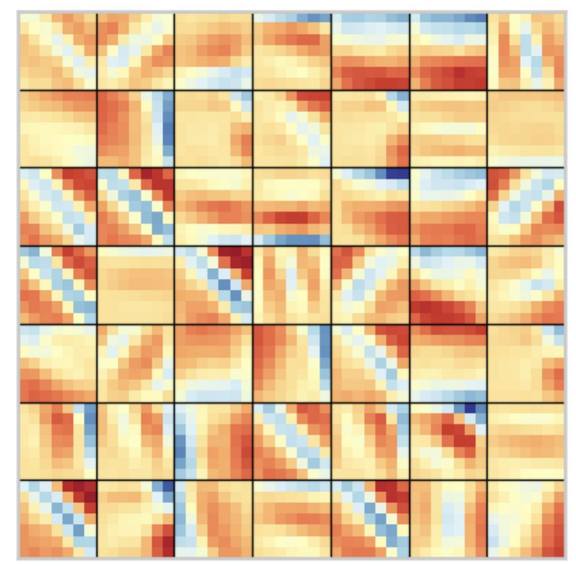



## MNIST Few-Short Classification

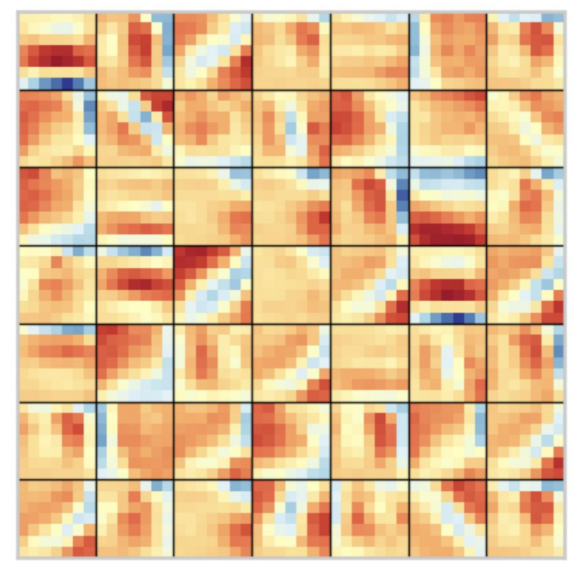
We compare the performance of a Bayesian CNN with 4 different prior distributions with limited training data:





(b) Learned filters



(c) Samples from DWP

## Fast Convergence: VAE and ConvNet

We compare different kernel **initialization techniques** :

- Vanilla Xavier
- Learned filters
- Samples from dwp