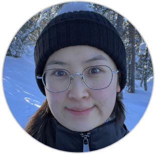
Streamlining Prediction in Bayesian Deep Learning



Rui Li
Aalto University

Marcus Klasson
Aalto University

Arno Solin
Aalto University

Martin Trapp
Aalto University

A! Aalto University

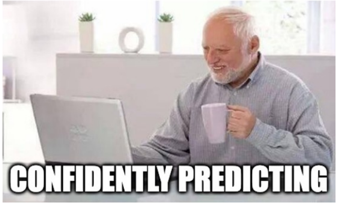
International Conference on
Learning Representations

ICLR

Why Bayesian Deep Learning?

Deep learning is great, but it is not reliable:

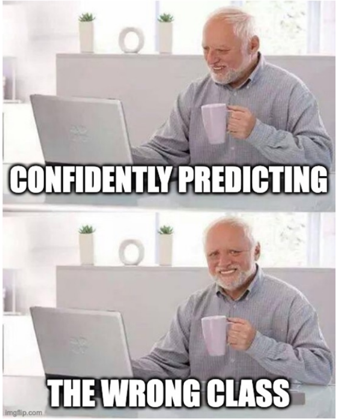
- Sensitive to perturbations
- Don't know when they don't know
- Overconfident predictions



Why Bayesian Deep Learning?

Deep learning is great, but it is not reliable:

- Sensitive to perturbations
- Don't know when they don't know
- Overconfident predictions

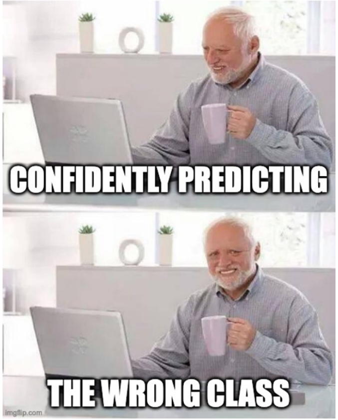


Why Bayesian Deep Learning?

Deep learning is great, but it is not reliable:

- **Sensitive** to perturbations
- **Don't know** when they don't know
- **Overconfident** predictions

💡 Bayesian deep learning aims to solve this.



Bayesian Deep Learning

1. Specify the prior: $p(\boldsymbol{\theta})$
2. Infer the posterior: $p(\boldsymbol{\theta} \mid \mathcal{D})$
3. Make the prediction:

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*) = \int p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}$$

Bayesian Deep Learning

1. Specify the prior: $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_0, \boldsymbol{S}_0)$
2. Infer the posterior: $p(\boldsymbol{\theta} \mid \mathcal{D}) \approx \mathcal{N}(\mathbf{m}, \boldsymbol{S})$
3. Make the prediction:

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*) = \int p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}$$

Bayesian Deep Learning

1. Specify the prior: $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_0, \boldsymbol{S}_0)$
2. Infer the posterior: $p(\boldsymbol{\theta} \mid \mathcal{D}) \approx \mathcal{N}(\mathbf{m}, \boldsymbol{S})$
3. Make the prediction:

$$
\begin{aligned}
p(\boldsymbol{y}^* \mid \boldsymbol{x}^*) &= \int p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta} \\
&\approx \frac{1}{S} \underbrace{\sum_{s=1}^{S} p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{\theta}^{(s)})}_{\text{(Challenge 2)}}, \qquad \underbrace{\boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta} \mid \mathcal{D})}_{\text{(Challenge 1)}}
\end{aligned}
$$

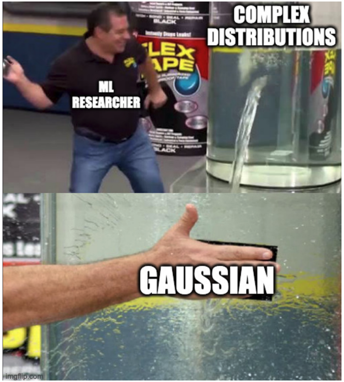
Bayesian Deep Learning

1. Specify the prior: $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_0, \boldsymbol{S}_0)$
2. Infer the posterior: $p(\boldsymbol{\theta} \,|\, \mathcal{D}) \approx \mathcal{N}(\mathbf{m}, \boldsymbol{S})$
3. Make the prediction:

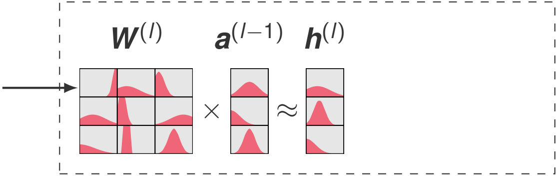
$$p(\boldsymbol{y}^* \,|\, \boldsymbol{x}^*) = \int p(\boldsymbol{y}^* \,|\, \boldsymbol{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}$$

$$\cancel{\approx \frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{y}^* \,|\, \boldsymbol{x}^*, \boldsymbol{\theta}^{(s)}), \quad \boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta} \,|\, \mathcal{D})}$$

We approximate it with a Gaussian

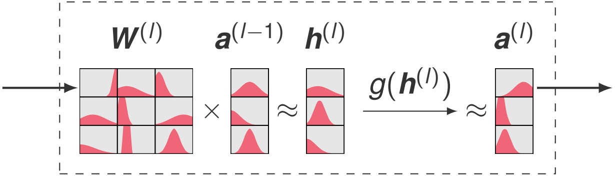


Streamlining Prediction in Bayesian Deep Learning



$$h^{(l)} = \boxed{W^{(l)} a^{(l-1)}} + b^{(l)}$$ approximate as Gaussian

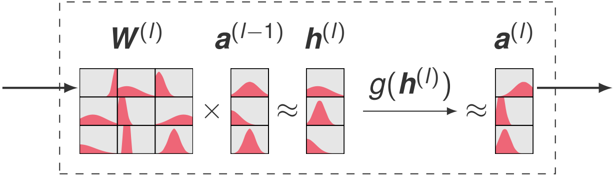
Streamlining Prediction in Bayesian Deep Learning



$$\boldsymbol{h}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}$$ approximate as Gaussian

$$\boldsymbol{a}^{(l)} = g(\boldsymbol{h}^{(l)})$$ first order Taylor expansion

Streamlining Prediction in Bayesian Deep Learning



sampling-free
single forward pass

$$\boldsymbol{h}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}$$ approximate as Gaussian

$$\boldsymbol{a}^{(l)} = g(\boldsymbol{h}^{(l)})$$ first order Taylor expansion

Efficient & Effective Uncertainty Quantification

Metrics	Methods	CIFAR-10	CIFAR-100	DTD	RESISC	IMAGENET-R
ACC ↑	LA Sampling	0.971±0.002	**0.882**±0.003	**0.715**±0.010	**0.892**±0.004	0.731±0.012
	LA GLM	**0.976**±0.002	0.879±0.003	**0.718**±0.010	**0.891**±0.004	**0.739**±0.012
	LA Ours	**0.976**±0.002	**0.880**±0.003	**0.719**±0.010	**0.892**±0.004	**0.739**±0.012
	MFVI Sampling	**0.975**±0.002	**0.880**±0.003	**0.732**±0.010	**0.867**±0.004	**0.730**±0.012
	MFVI Ours	**0.975**±0.002	**0.880**±0.003	**0.734**±0.010	**0.867**±0.004	**0.728**±0.012
NLPD ↓	LA Sampling	0.170±0.004	**0.444**±0.012	1.238±0.028	0.461±0.009	1.208±0.048
	LA GLM	0.092±0.007	0.459±0.012	1.197±0.029	0.385±0.010	**1.180**±0.047
	LA Ours	**0.086**±0.006	**0.456**±0.012	**1.068**±0.035	**0.352**±0.012	1.267±0.043
	MFVI Sampling	0.133±0.011	0.641±0.022	**1.091**±0.048	1.010±0.041	1.577±0.083
	MFVI Ours	**0.088**±0.006	**0.468**±0.013	**1.007**±0.035	**0.617**±0.019	**1.234**±0.052
ECE ↓	LA Sampling	**0.006**	**0.022**	0.197	0.129	0.070
	LA GLM	0.011	0.024	0.155	0.053	**0.057**
	LA Ours	0.008	0.027	**0.040**	**0.016**	0.132
	MFVI Sampling	0.015	0.070	0.075	0.079	0.118
	MFVI Ours	**0.008**	**0.025**	**0.042**	**0.017**	**0.036**

Efficient & Effective Uncertainty Quantification

Metrics	Methods	CIFAR-10	CIFAR-100	DTD	RESISC	IMAGENET-R
ACC ↑	LA Sampling	0.971±0.002	**0.882**±0.003	**0.715**±0.010	**0.892**±0.004	0.731±0.012
	LA GLM	**0.976**±0.002	0.879±0.003	**0.718**±0.010	**0.891**±0.004	**0.739**±0.012
	LA Ours	**0.976**±0.002	**0.880**±0.003	**0.719**±0.010	**0.892**±0.004	**0.739**±0.012
	MFVI Sampling	**0.975**±0.002	**0.880**±0.003	**0.732**±0.010	**0.867**±0.004	**0.730**±0.012
	MFVI Ours	**0.975**±0.002	**0.880**±0.003	**0.734**±0.010	**0.867**±0.004	**0.728**±0.012
NLPD ↓	LA Sampling	0.170±0.004	**0.444**±0.012	1.238±0.028	0.461±0.009	1.208±0.048
	LA GLM	0.092±0.007	0.459±0.012	1.197±0.029	0.385±0.010	**1.180**±0.047
	LA Ours	**0.086**±0.006	**0.456**±0.012	**1.068**±0.035	**0.352**±0.012	1.267±0.043
	MFVI Sampling	0.133±0.011	0.641±0.022	**1.091**±0.048	1.010±0.041	1.577±0.083
	MFVI Ours	**0.088**±0.006	**0.468**±0.013	**1.007**±0.035	**0.617**±0.019	**1.234**±0.052
ECE ↓	LA Sampling	**0.006**	**0.022**	0.197	0.129	0.070
	LA GLM	0.011	0.024	0.155	0.053	**0.057**
	LA Ours	0.008	0.027	**0.040**	**0.016**	0.132
	MFVI Sampling	0.015	0.070	0.075	0.079	0.118
	MFVI Ours	**0.008**	**0.025**	**0.042**	**0.017**	**0.036**

Efficient & Effective Uncertainty Quantification

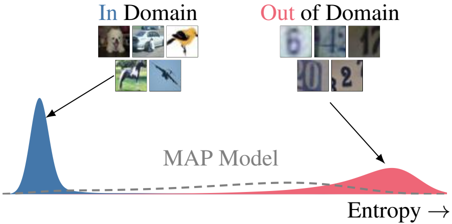
Metrics	Methods	CIFAR-10	CIFAR-100	DTD	RESISC	IMAGENET-R
ACC ↑	LA Sampling	0.971±0.002	**0.882**±0.003	**0.715**±0.010	**0.892**±0.004	0.731±0.012
	LA GLM	**0.976**±0.002	0.879±0.003	**0.718**±0.010	**0.891**±0.004	**0.739**±0.012
	LA Ours	**0.976**±0.002	**0.880**±0.003	**0.719**±0.010	**0.892**±0.004	**0.739**±0.012
	MFVI Sampling	**0.975**±0.002	**0.880**±0.003	**0.732**±0.010	**0.867**±0.004	**0.730**±0.012
	MFVI Ours	**0.975**±0.002	**0.880**±0.003	**0.734**±0.010	**0.867**±0.004	**0.728**±0.012
NLPD ↓	LA Sampling	0.170±0.004	**0.444**±0.012	1.238±0.028	0.461±0.009	1.208±0.048
	LA GLM	0.092±0.007	0.459±0.012	1.197±0.029	0.385±0.010	**1.180**±0.047
	LA Ours	**0.086**±0.006	**0.456**±0.012	**1.068**±0.035	**0.352**±0.012	1.267±0.043
	MFVI Sampling	0.133±0.011	0.641±0.022	**1.091**±0.048	1.010±0.041	1.577±0.083
	MFVI Ours	**0.088**±0.006	**0.468**±0.013	**1.007**±0.035	**0.617**±0.019	**1.234**±0.052
ECE ↓	LA Sampling	**0.006**	**0.022**	0.197	0.129	0.070
	LA GLM	0.011	0.024	0.155	0.053	**0.057**
	LA Ours	0.008	0.027	**0.040**	**0.016**	0.132
	MFVI Sampling	0.015	0.070	0.075	0.079	0.118
	MFVI Ours	**0.008**	**0.025**	**0.042**	**0.017**	**0.036**

Methods	AVG. RUNTIME (± STD) ↓
MAP	3.737±0.093
LA Sampling	190.806±0.137
LA GLM	17.191±0.734
MFVI Sampling	207.854±0.307
Ours (+ Cov)	14.728±0.144
Ours	4.350±0.079

On classification, we achieve better or on-par performance faster than baselines.

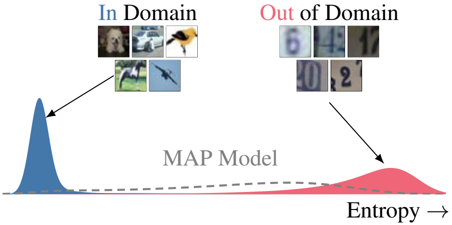
OOD detection and Input Sensitivity Analysis

Practical Outlier Detection

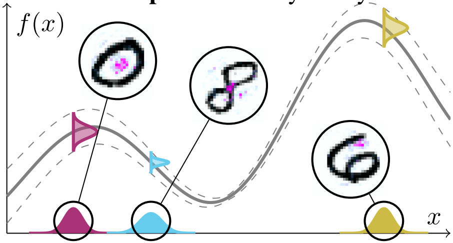


OOD detection and Input Sensitivity Analysis

Practical Outlier Detection



Input Sensitivity Analysis



Take away

Open-source library: `https://github.com/AaltoML/SUQ`

- **Goal:** Make good predictions fast in Bayesian neural networks.
- **Approach:** Locally linearised the neural network for a tractable posterior predictive distribution.
- **Result:** Better or on par performance with faster speed.