Low-rank Bayesian neural networks

Markus Heinonen, Phd

Academy Research Fellow

markus.o.heinonen@aalto.fi

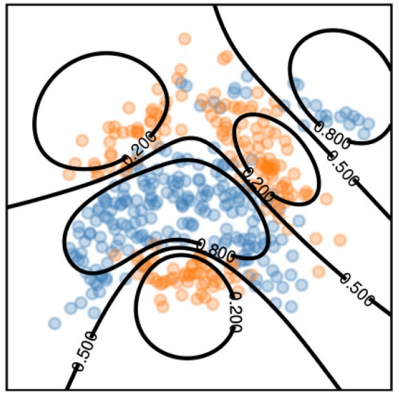
Aalto University

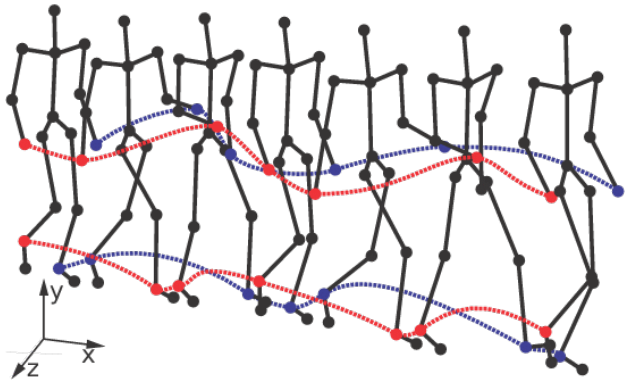
Finnish Center for Artificial Intelligence FCAI

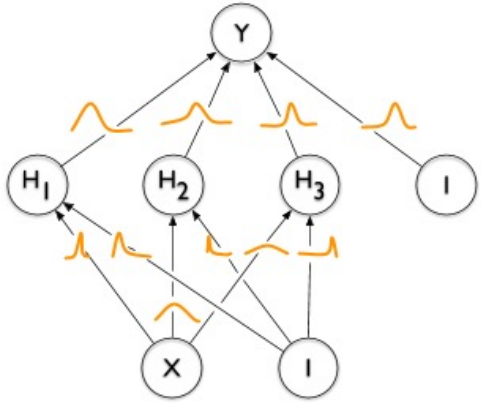
UnGroundNLP-2022, 29.4.2022

My background

- Gaussian processes
  - Rossi, Heinonen, Bonilla, Shen, Filippone. Sparse Gaussian Processes Revisited: Bayesian Approaches to Inducing-Variable Approximations, AISTATS'21
- Dynamical Systems
  - Yildiz, Heinonen, Lähdesmäki. ODE2VAE: Deep generative second order ODEs with Bayesian neural networks, NIPS'19
- Bayesian deep learning
  - Trinh, Kaski, Heinonen. Scalable Bayesian neural networks by layer-wise input augmentation. Arxiv'20
- Bioinformatics, drug development, robotics, etc

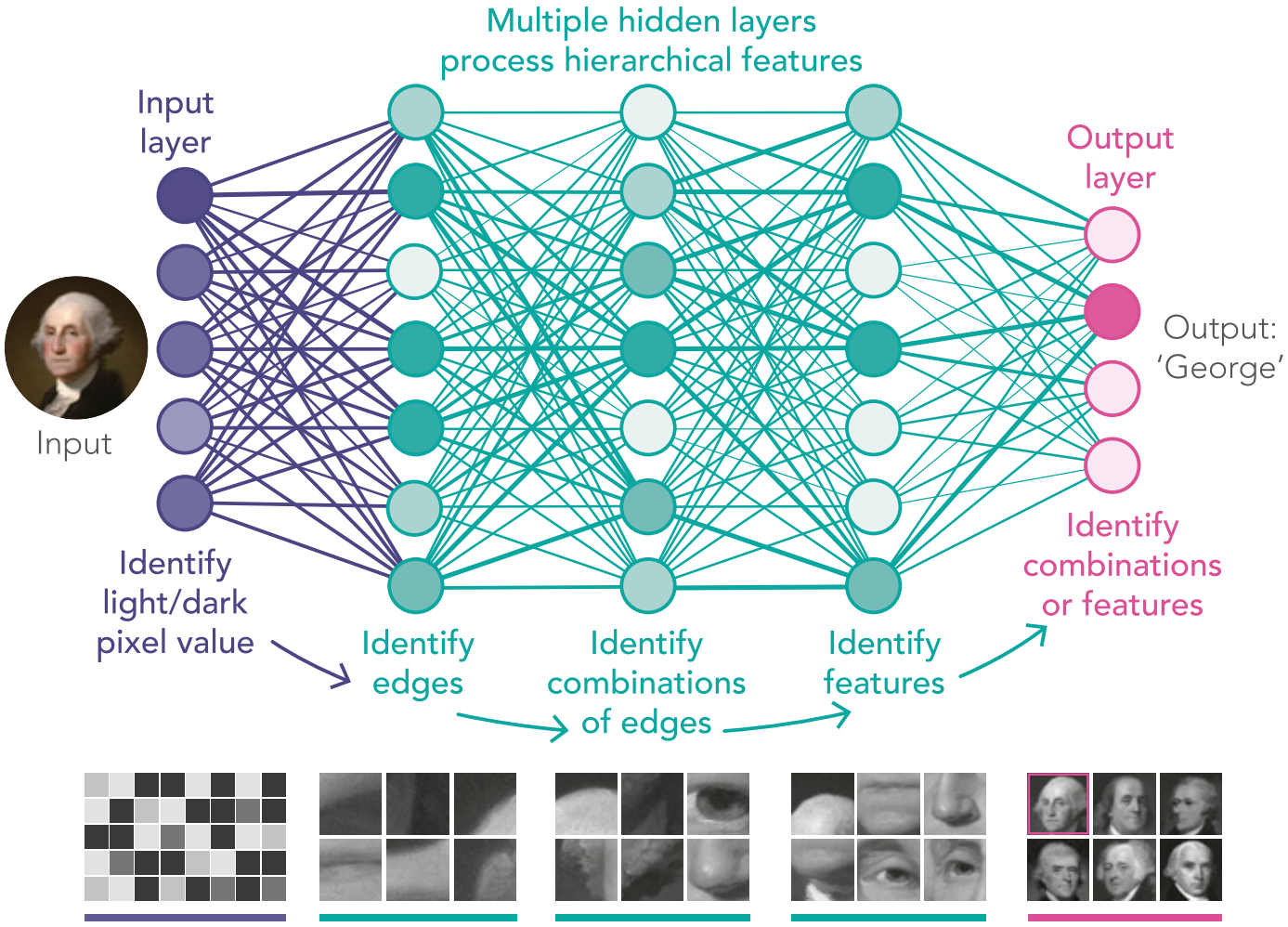






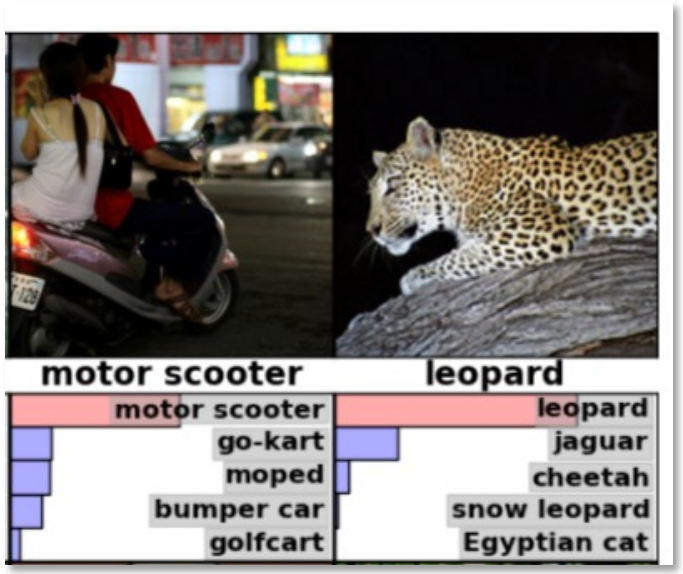
Deep learning

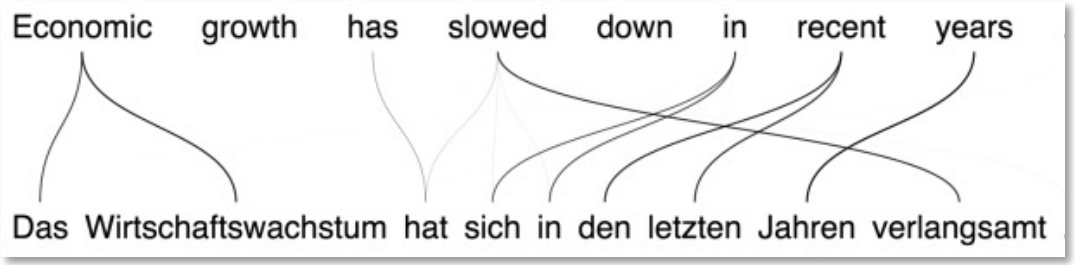
- Learn to explain what we can't explain
- Classic neural networks
  - Massive capacity (... to make mistakes)
  - Massive amounts of data
  - Huge parameter spaces
  - Deterministic: no uncertainty
  - Black box'y



PNAS 2019: What are the limits of deep learning?

Deep learning breakthroughs





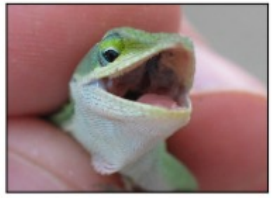
Artificial intelligence yields new antibiotic

A deep-learning model identifies a powerful new drug that can kill many species of antibiotic-resistant bacteria.

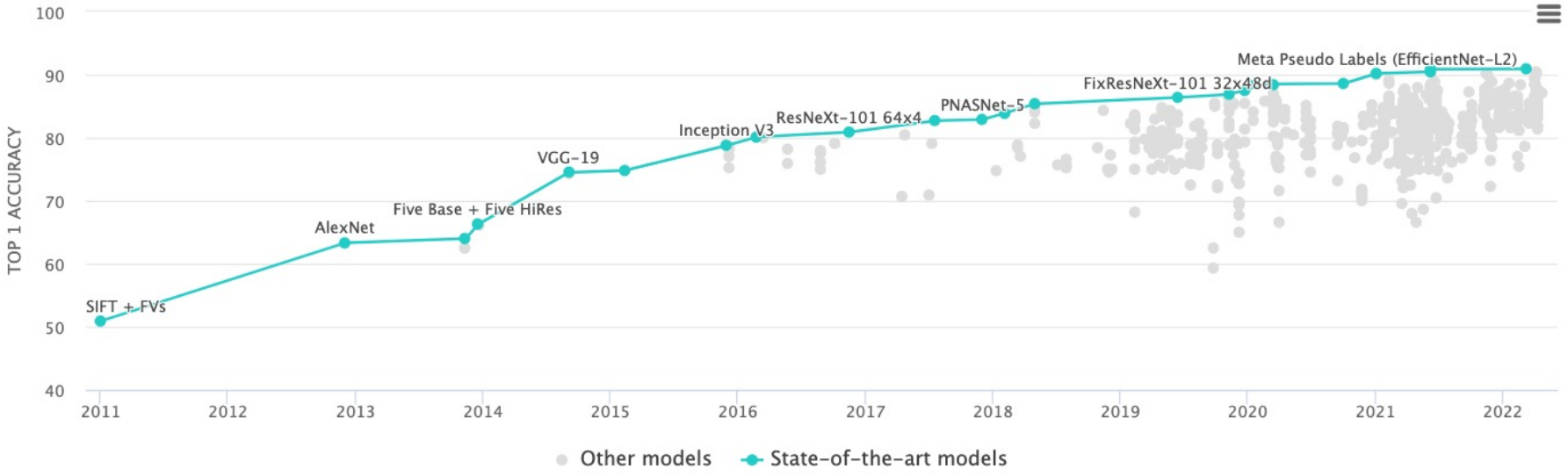
Benchmarks drive deep learning

- Standard benchmarks and metrics
- ImageNet top-1 accuracy has risen 50% -> 90% in 10 years
- .. is the accuracy real?

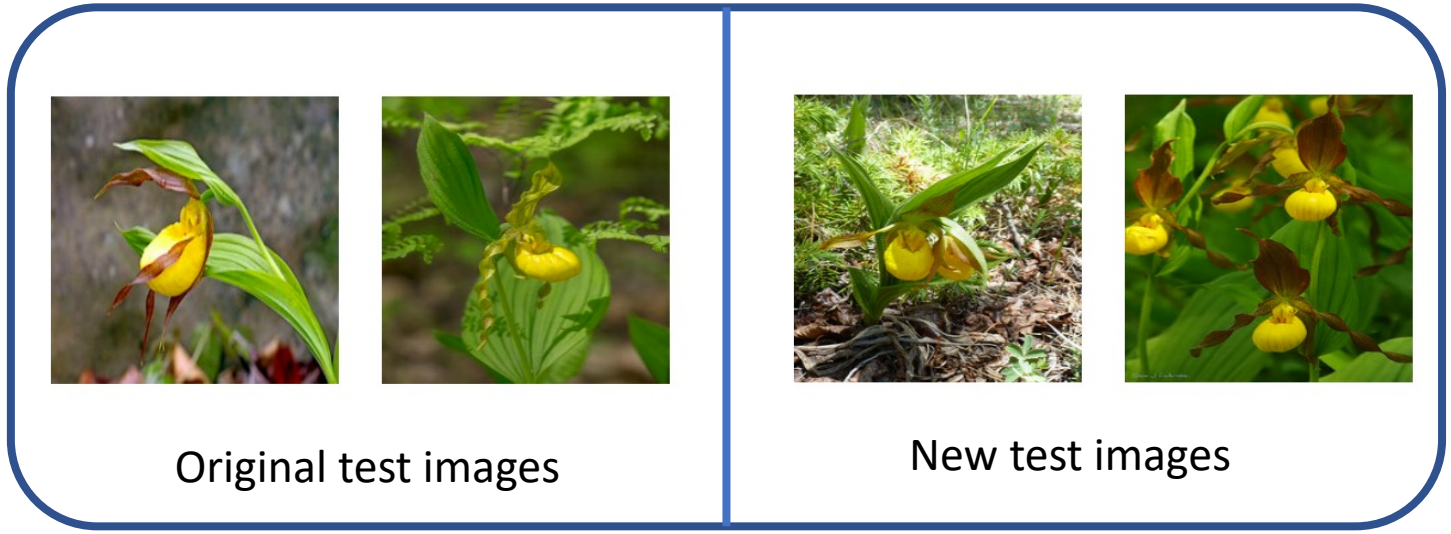

n02097047 (196)


n01682714 (40)


n03134739 (522)



Uh oh…

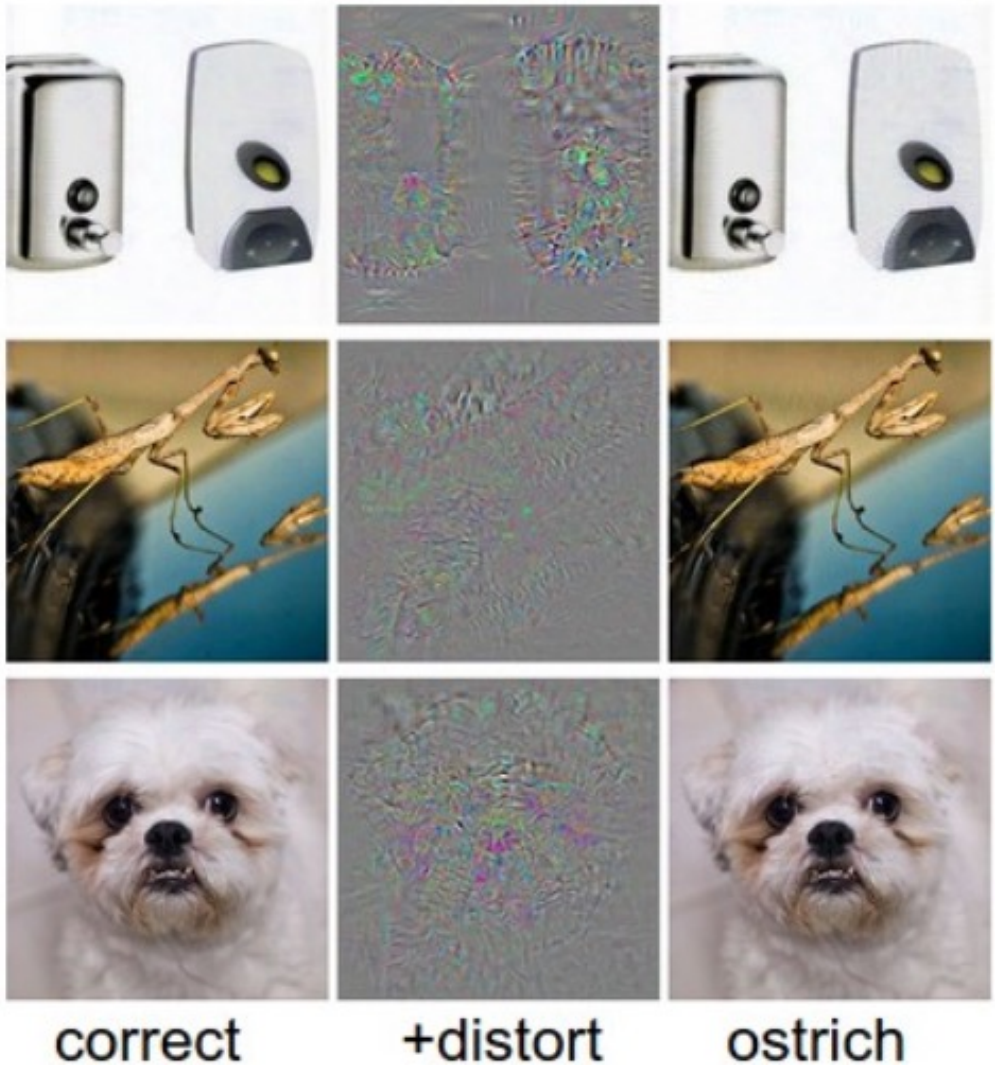


Model	Orig. Accuracy	New Accuracy
pnasnet_large_tf	82.9 [82.5, 83.2]	72.2 [71.3, 73.1]
nasnetalarge	82.5 [82.2, 82.8]	72.2 [71.3, 73.1]
resnet152	78.3 [77.9, 78.7]	67.0 [66.1, 67.9]
inception_v3_tf	78.0 [77.6, 78.3]	66.1 [65.1, 67.0]
densenet161	77.1 [76.8, 77.5]	65.3 [64.4, 66.2]
vgg19_bn	74.2 [73.8, 74.6]	61.9 [60.9, 62.8]

Recht, Roelofs, Schmidt, Shankar. Do ImageNet Classifiers Generalize to ImageNet?, ICML 2019

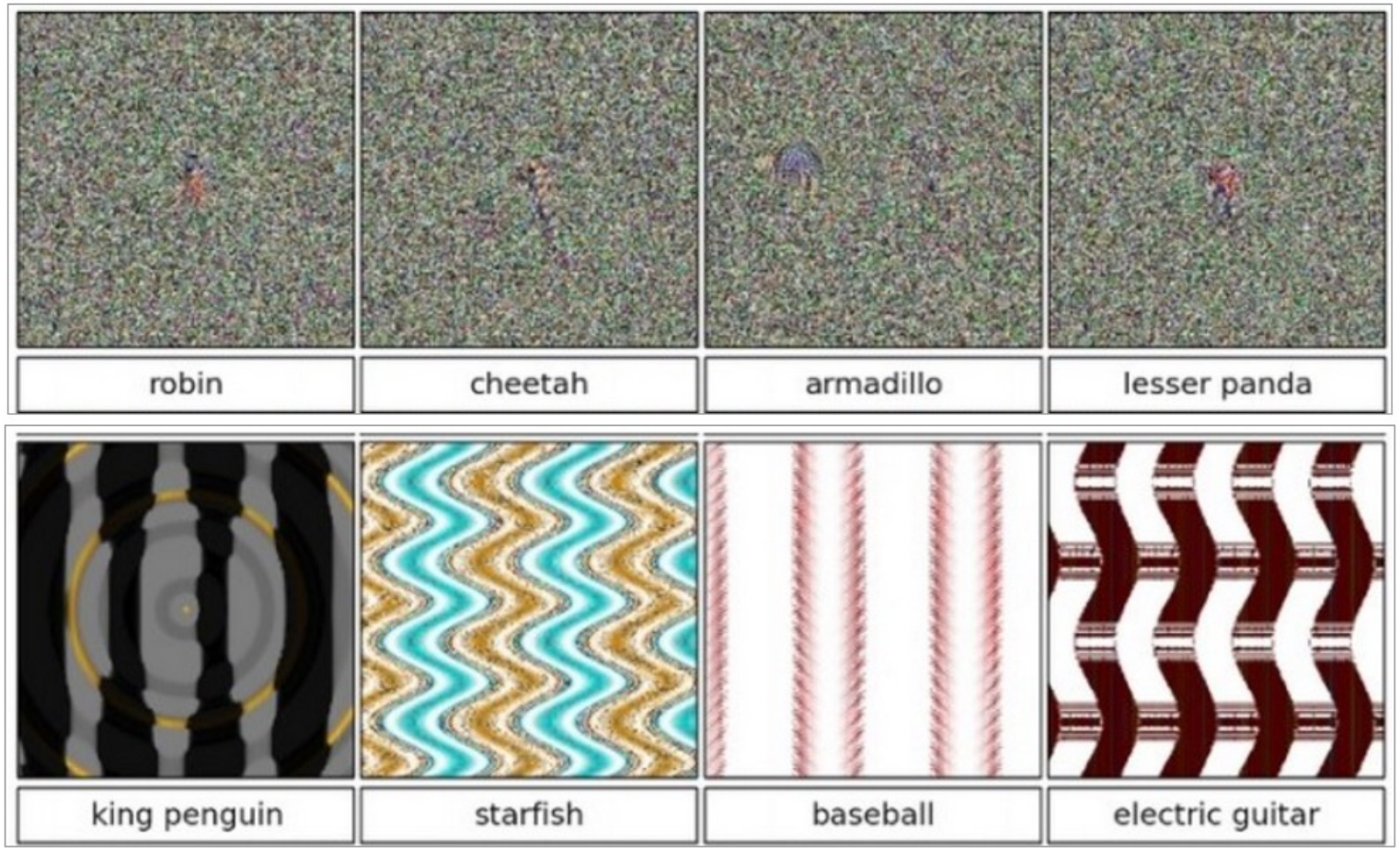
Adversarial attacks

- Neural networks can be fooled by hostile perturbations
  - Conventional training error is oblivious to attacks
    - More data would not help!
  - Non-smooth decision boundaries
  - Prediction overconfidence



Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus. Intriguing properties of neural networks, ICLR 2014

Sensitivity to patterns

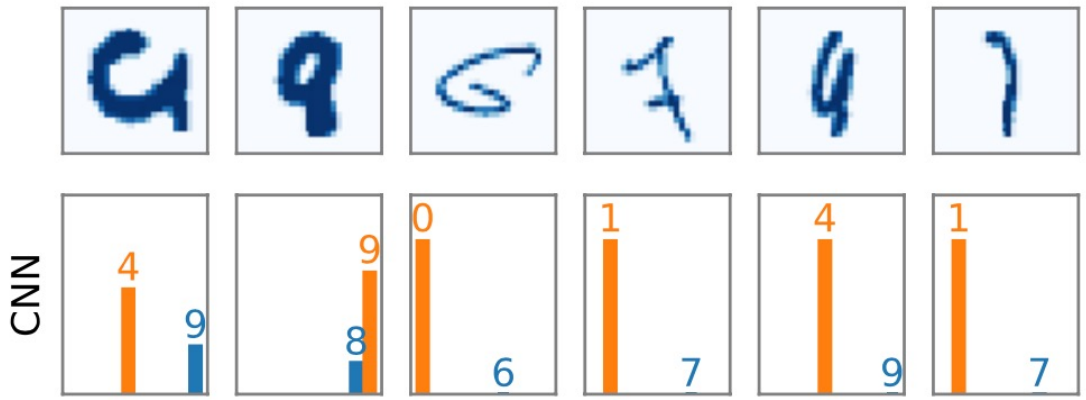


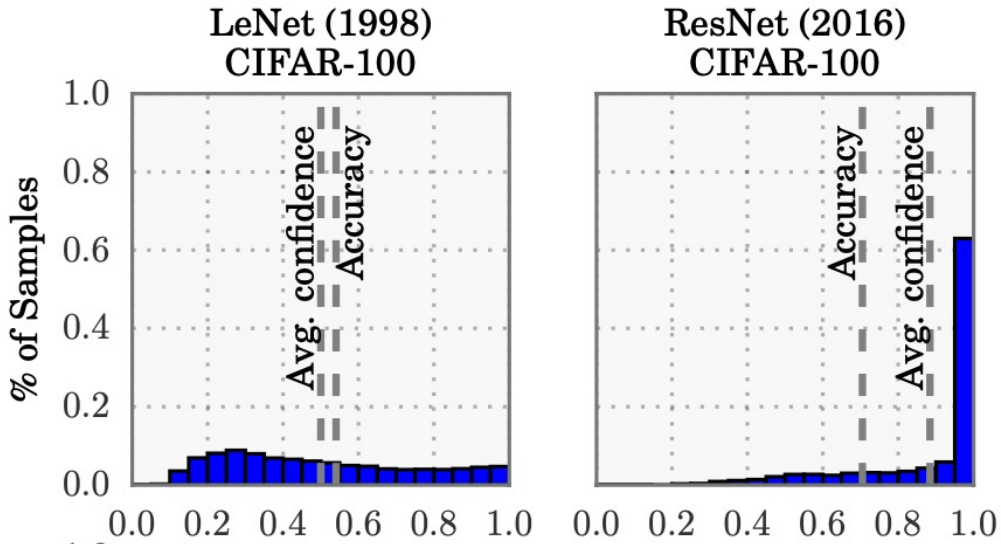
These images are classified with >99.6% confidence as the shown class by a Convolutional Network.

Goodfellow, Shlens, Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015

Overconfidence of DL

- Neural network predictions are often overconfident
- Are such predictions useful for downstream tasks?

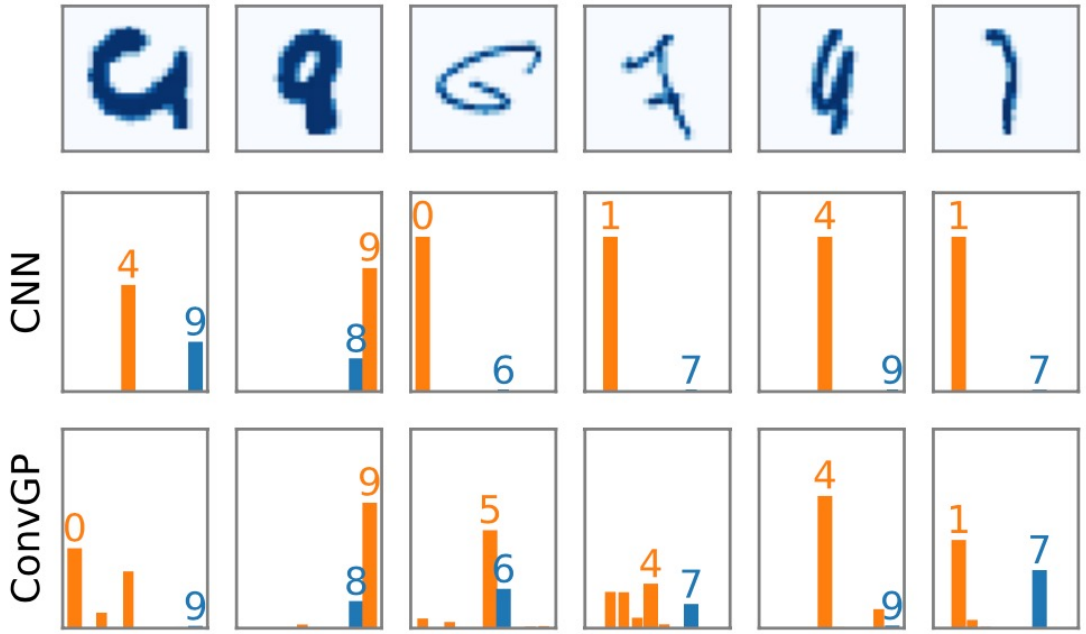




Guo, Pleiss, Sun, Weinberger. On Calibration of Modern Neural Networks, ICML 2017

Deep learning calibration

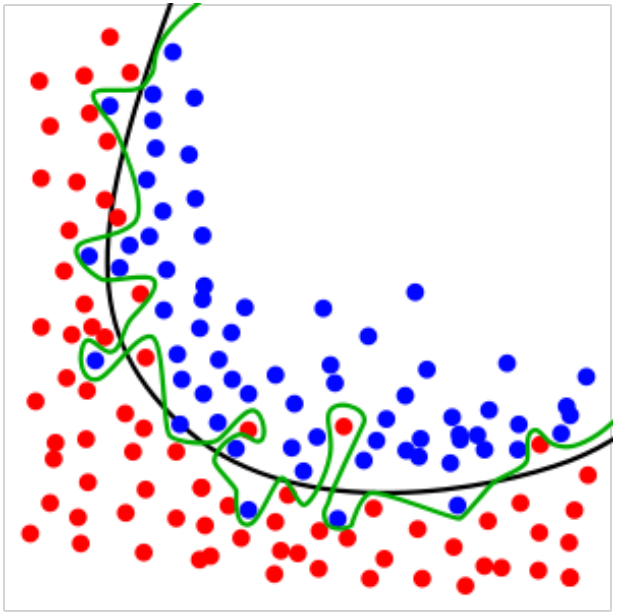
- Multiple calibration techniques
- Bayesian models are often less susceptible



Dutordoir, Wilk, Artemev, Hensman. Bayesian Image Classification with Deep Convolutional Gaussian Processes, AISTATS 2020

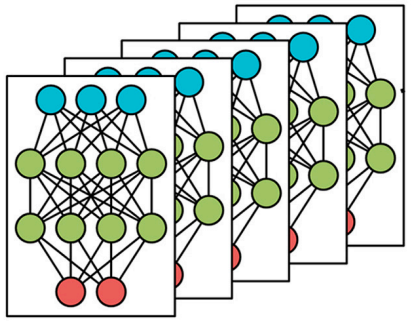
Remedies

- Towards better models
  - Augmentations improve generalisation
  - Ensembling improves uncertainty
  - Dropout increases robustness
  - Skip connections help optimisation

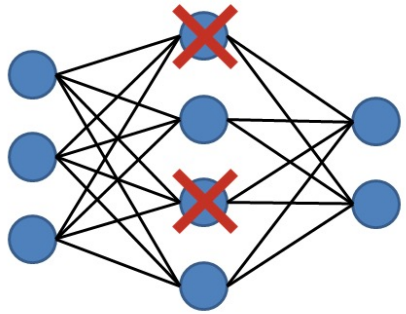


Augmentations

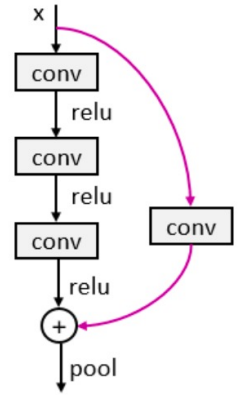
Ensembles



Dropout

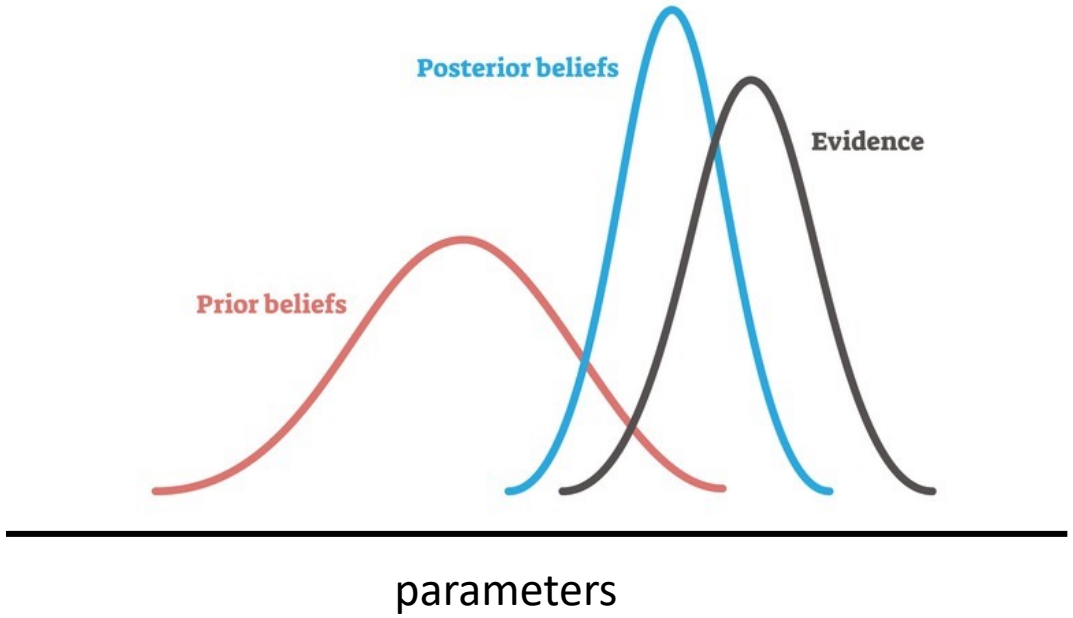


Skip connection



Bayesian learning

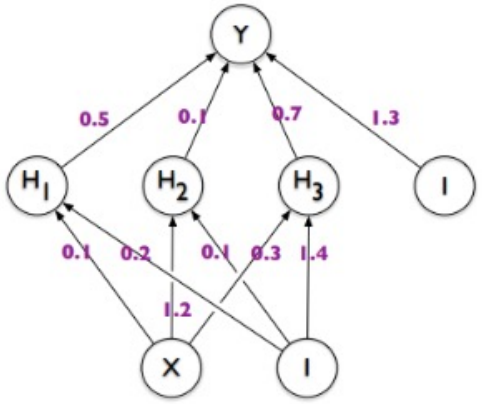
- Assign prior *p(params)* and likelihood *p(data|params)*
- Infer posterior *p(params|data)*
- We obtain distributions of neural networks
  - Captures uncertainty
- Bayes is expensive: can we apply it to NNs?

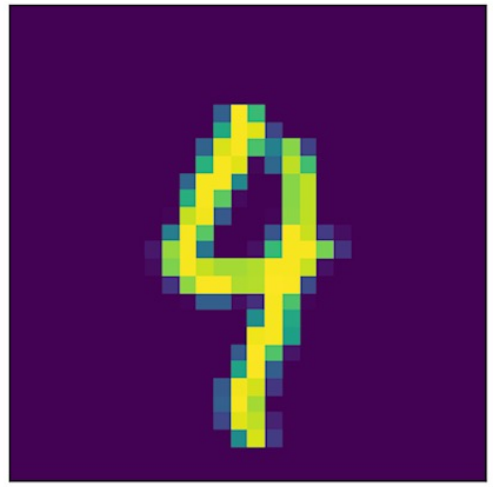


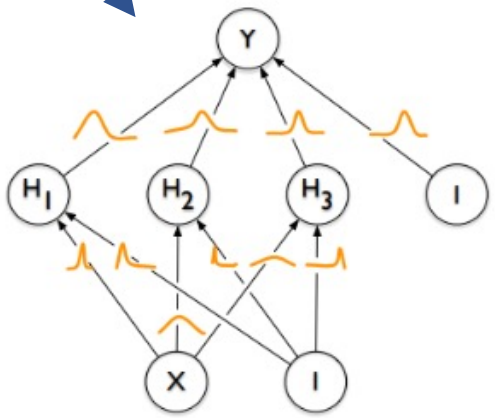
Thomas Bayes

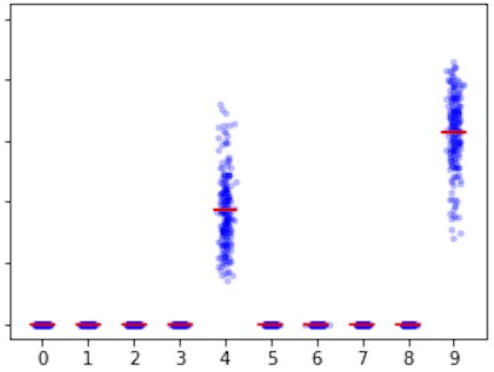
Bayesian neural networks (BNN)

- Bayesian deep learning
  - Probabilistic models with parameter uncertainty
  - Define what we know and what we don't know
  - Solid statistical foundations
  - Improved robustness
  - Millions of parameters





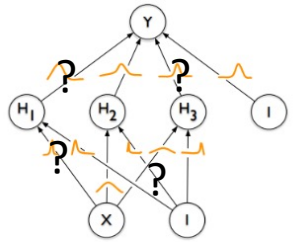




$$\text{parameter posterior:} \quad p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

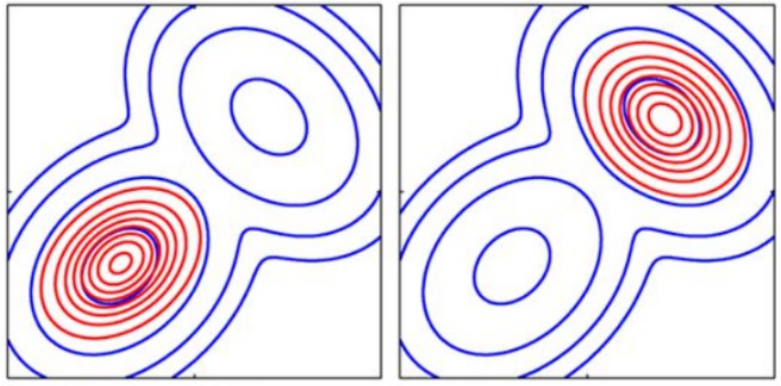
$$\text{predictive posterior:} \quad p(y|\mathbf{x},\mathcal{D}) = \int p(y|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

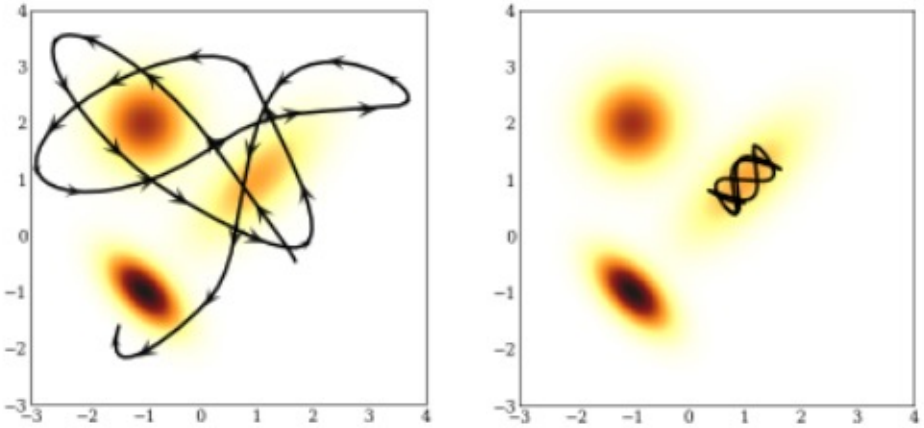
Standard solutions: variational inference and MCMC



- Variational inference
  - capture one mode
- MCMC
  - travel parameter space

$$\min_{\gamma} \quad KL[q_\gamma(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})]$$

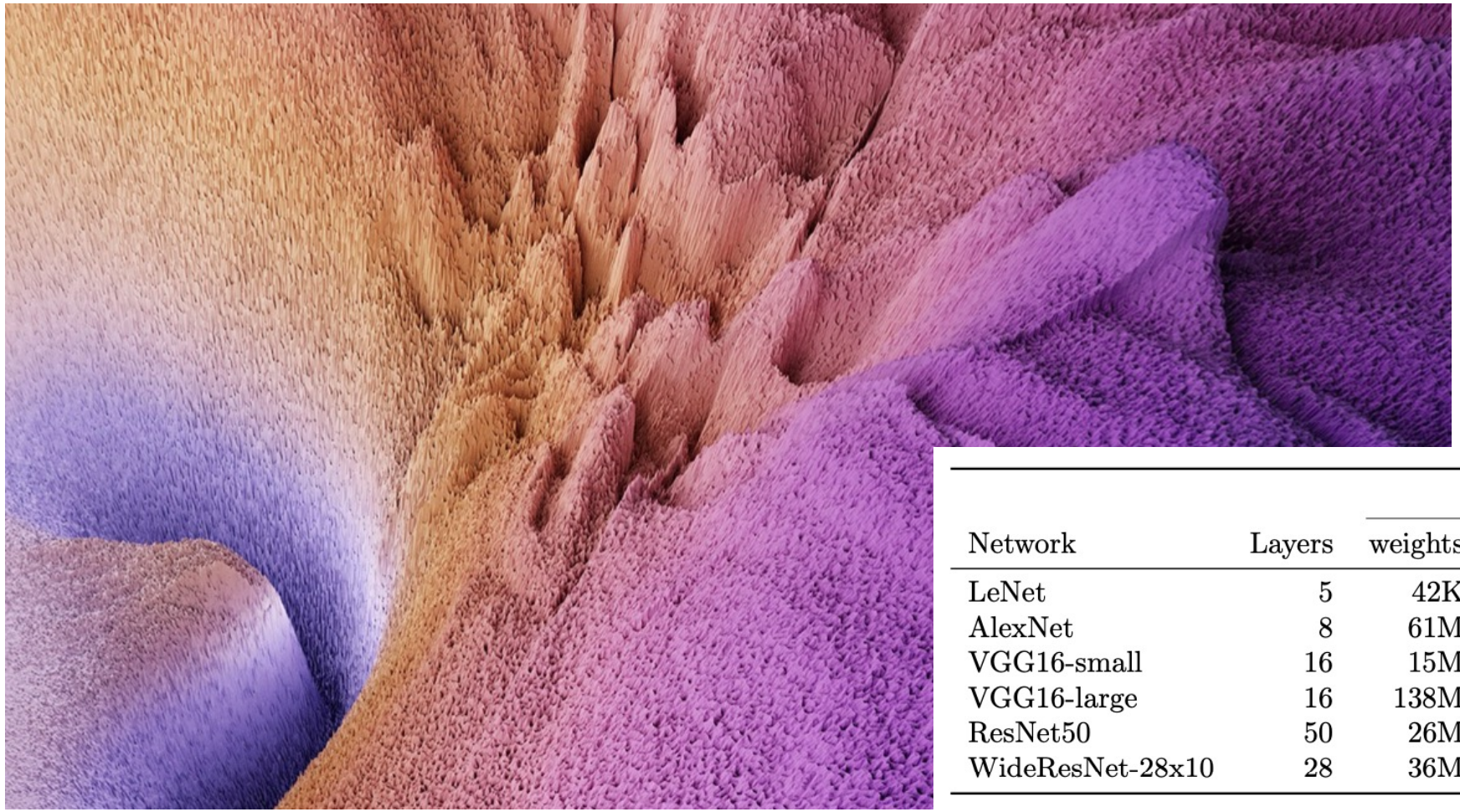




Blundell et al. Weight Uncertainty in Neural Networks. ICML'15

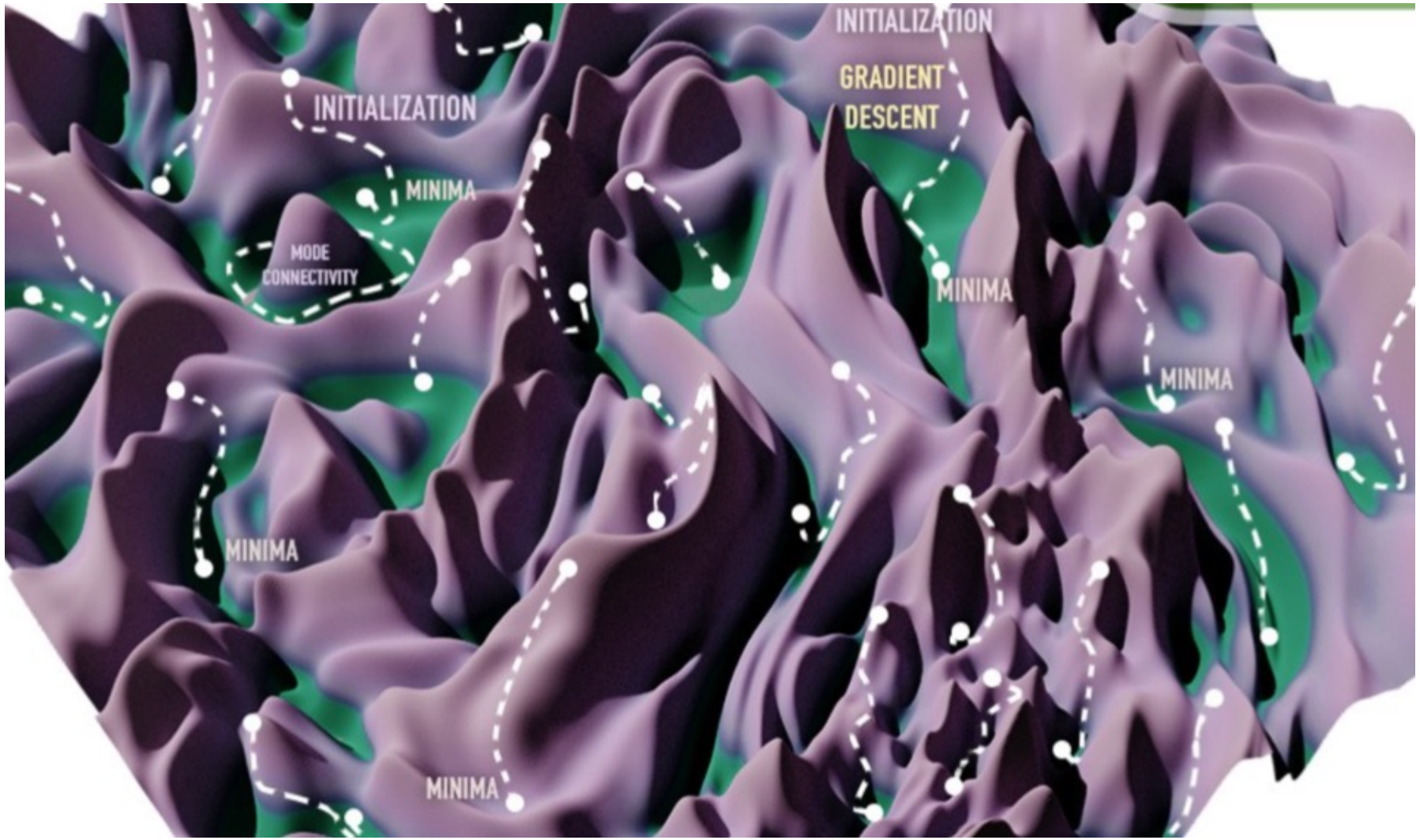
Izmailov et al. What Are Bayesian Neural Network Posteriors Really Like? ICML'21

The posterior is absurdly large



		Parameters		
Network	Layers	weights	nodes	w/n ratio
LeNet	5	42K	23	1800x
AlexNet	8	61M	18,307	3300x
VGG16-small	16	15M	5,251	2900x
VGG16-large	16	138M	36,995	3700x
ResNet50	50	26M	24,579	1000x
WideResNet-28x10	28	36M	9,475	3800x

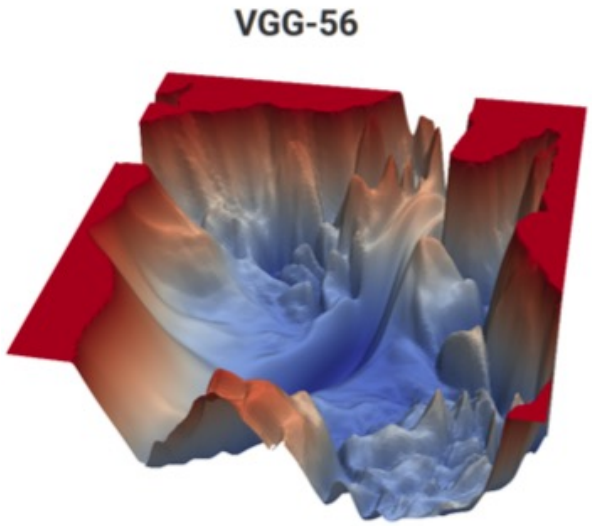
https://losslandscape.com/

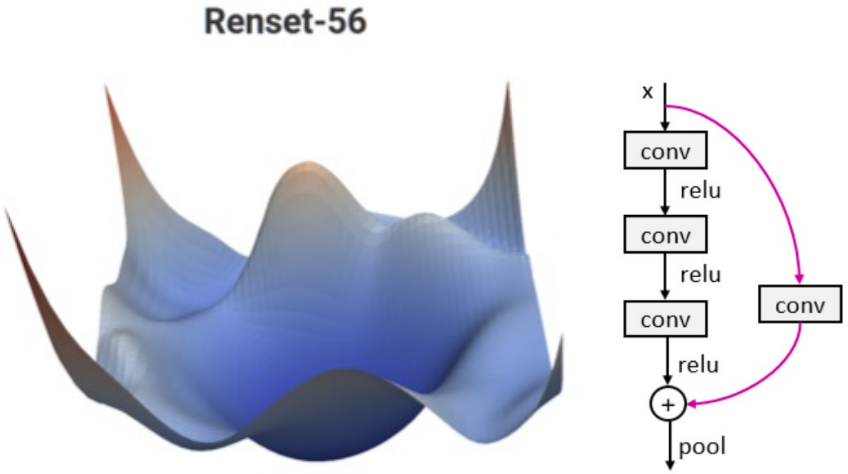


https://losslandscape.com/

Architecture matters

- Architectures have different loss landscapes





- => We can have simpler posteriors by smart design choices

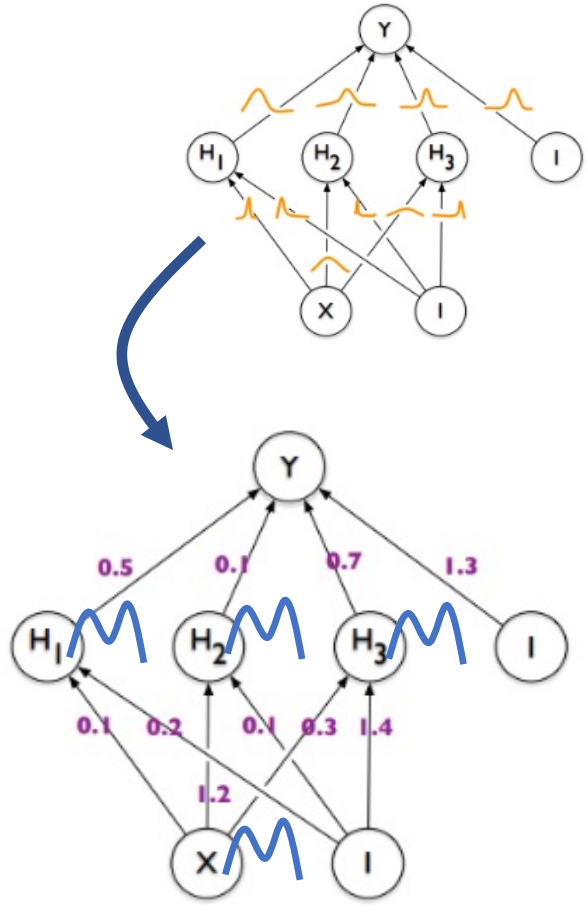
Hao et al. Visualising the loss landscape of neural networks, NIPS'18

Low-rank BNNs

- Our hypothesis: can we define order of magnitude smaller parameterisation?
- Node-based BNN

$$\mathbf{f}^0 = \mathbf{x}$$
$$\mathbf{f}^\ell = \sigma\left(\mathbf{W}^\ell_{\text{opt}}(\mathbf{f}^{\ell-1} \circ \mathbf{z}^\ell)\right)$$
$$\mathbf{z}^\ell \sim p(\mathbf{z}), \qquad dim(\mathbf{z}^\ell) = dim(\mathbf{f}^\ell)$$

  - Weights are point-optimised (re-use pretrained models)
  - Node variables $\mathbf{z}$ inject stochasticity
  - We use mixture-VI inference

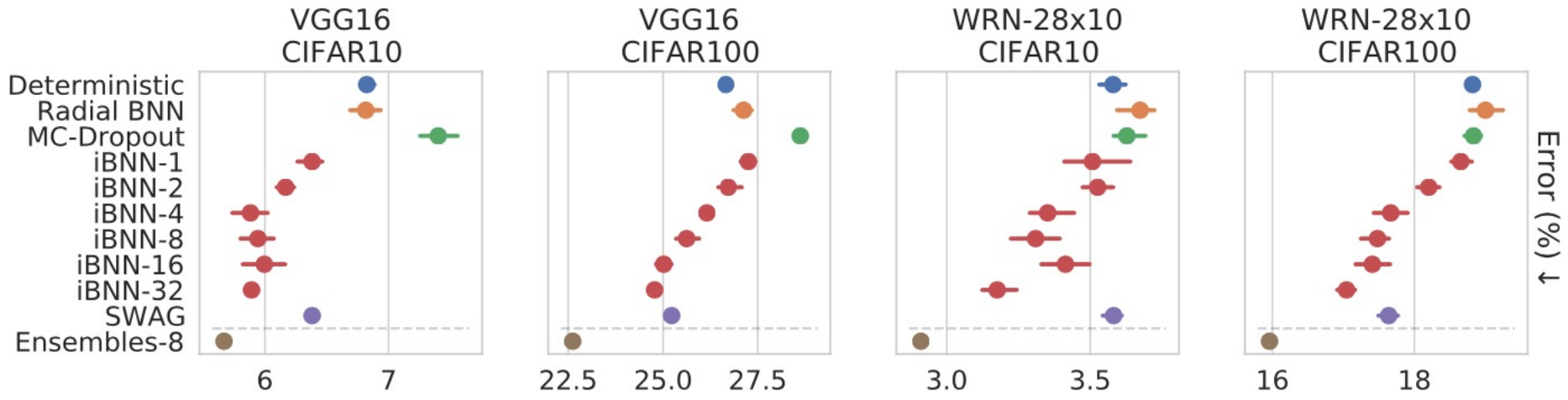


Network	Layers	Parameters		
		weights	nodes	w/n ratio
LeNet	5	42K	23	1800x
AlexNet	8	61M	18,307	3300x
VGG16-small	16	15M	5,251	2900x
VGG16-large	16	138M	36,995	3700x
ResNet50	50	26M	24,579	1000x
WideResNet-28x10	28	36M	9,475	3800x

Dusenberry et al. Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors. ICML'20

Node BNN performance

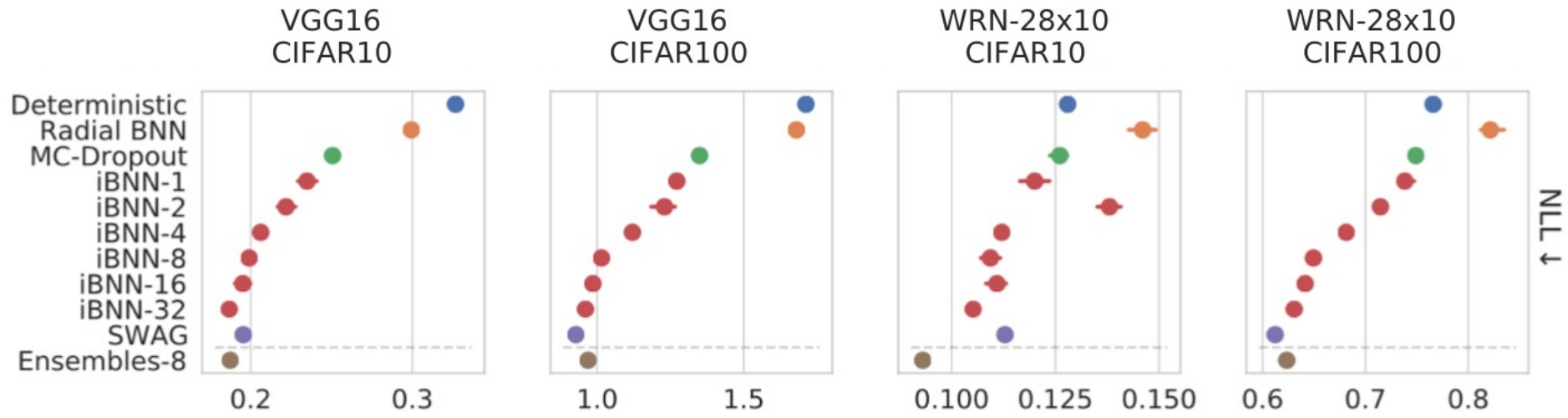
- Prior and posterior distributions for weights
  - Improved accuracy and calibration



Trung, Kaski, Heinonen. Scalable Bayesian neural networks by layer-wise input augmentation, 2021

Node BNN performance

- Prior and posterior distributions for weights
  - Improved accuracy and calibration



Trung, Kaski, Heinonen. Scalable Bayesian neural networks by layer-wise input augmentation, 2021

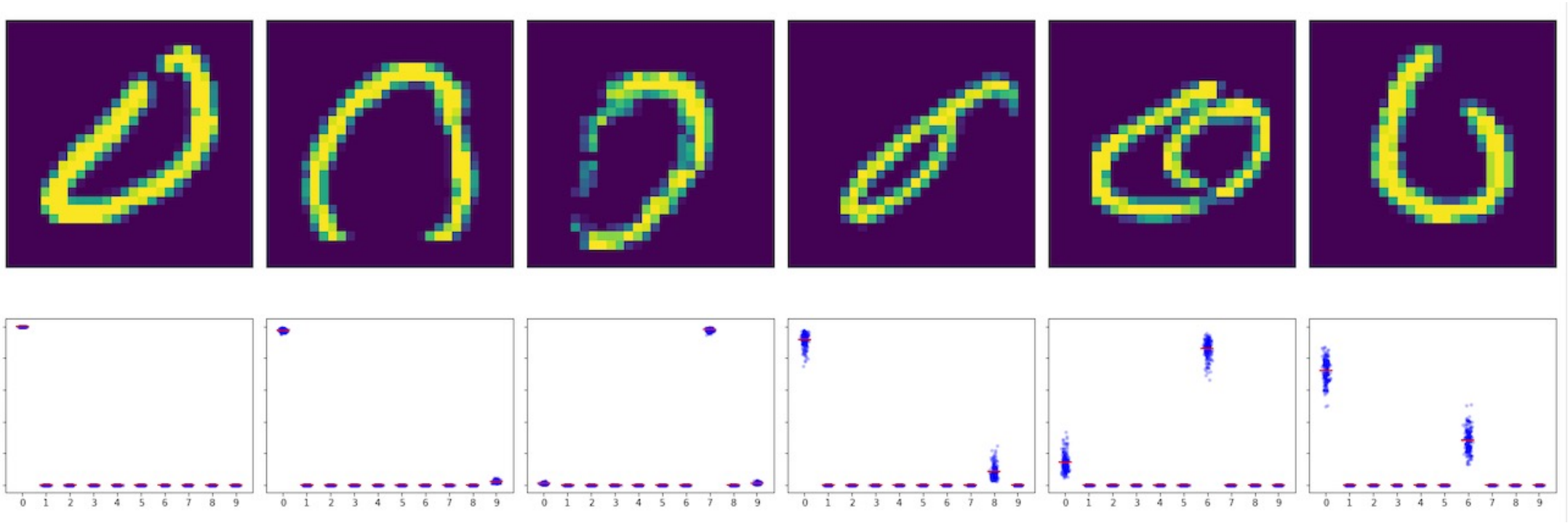
Turning pretrained models into Bayesian ones

- We start from pretrained weights, and infer node variables **z**
- Cheap performance boost
- Dramatic calibration improvement

Table 2. Results for ResNet-50 on the validation set of IMAGENET using pretrained parameters from `torchvision.models`. Each experiment is run once.

	Error (%) (↓)	NLL (↓)	ECE (↓)
SGD	23.87	0.962	0.037
iBNN-4	23.19	0.927	0.030
iBNN-8	23.10	0.919	0.026
iBNN-16	23.12	0.912	0.019
iBNN-32	**23.02**	**0.905**	**0.015**

We obtain reasonable uncertainties



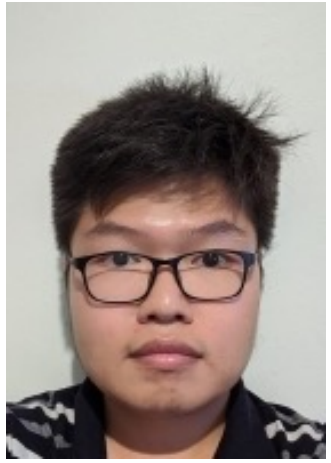
Joint work at ~Helsinki, Finland

Trinh, Kaski, Heinonen.
Scalable Bayesian neural networks by layer-wise input augmentation
Technical report, 2020

Trinh, Heinonen, Acerbi, Kaski
Tackling covariate shift with node-based Bayesian neural networks
Submitted 2022

Prof. Samuel Kaski
Aalto University

Prof. Luigi Acerbi
University of Helsinki

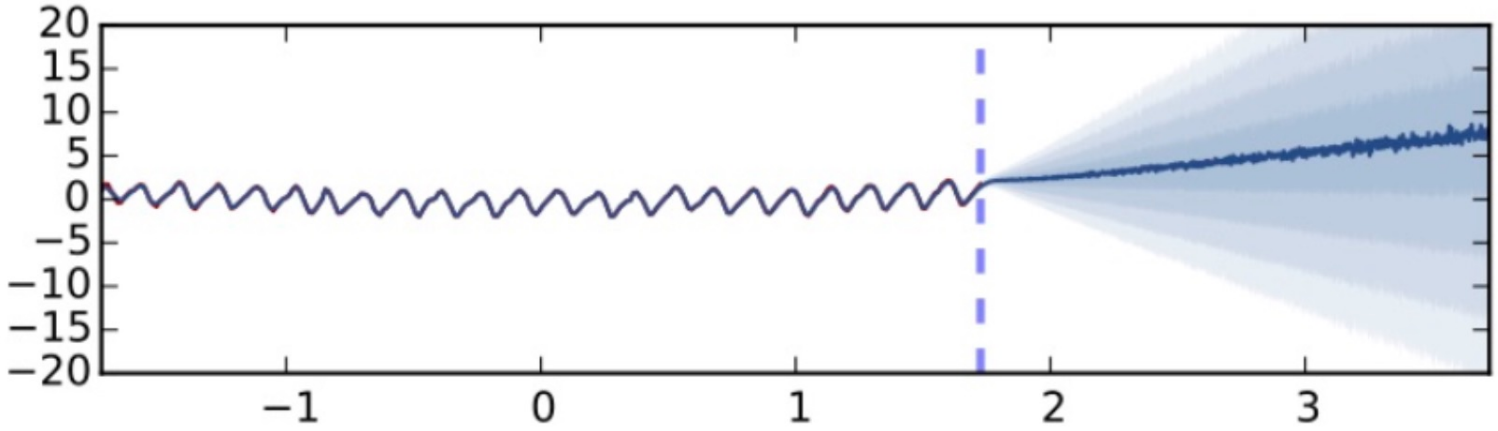


Trung Tringh
Aalto University

Markus Heinonen
Aalto University

It’s all about model and data

- Is our model appropriate for the underlying problem?
- Do we have enough data given the model?



Conclusion

- BNNs have better ‘out-of-the-box’ performance than CNNs over calibration, robustness and generalisation
- (Regular) weight-based BNN inference is a challenging task
- Node-based BNNs provide state-of-the-art performance cheaply
- More analysis of BNNs wrt adversariality, pattern sensitivity and generalisation are needed

Thanks!