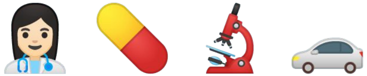
Pitfalls of In-Domain Uncertainty Estimation & Ensembling in Deep Learning

Arsenii Ashukha* Alexander Lyzhov* Dmitry Molchanov* Dmitry Vetrov

ICLR'20

SAMSUNG AI Center - Moscow
Higher School of Economics
p(B|A)yesgroup.ru
Skoltech
Skolkovo Institute of Science and Technology

Machine learning impacts critical decisions



Uncertainty Estimation

Reliable metrics

Strong baselines

Wide comparison of existing techniques

Ensembles of DNNs

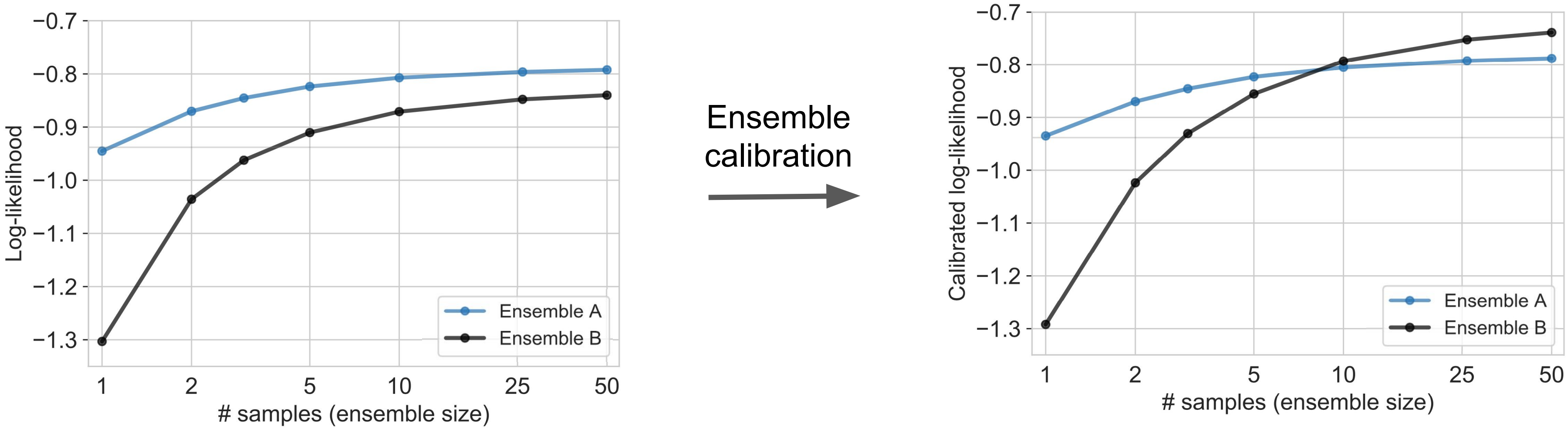
$$p_{\text{ens}}(y_i \mid x_i) = \frac{1}{K} \sum_{k=1}^{K} p(y_i \mid x_i, \omega_k)$$

- Log-likelihood
- Brier score
- Calibration errors (e.g. ECE, TACE)
- Misclassification detection performance (AUCs)

The metrics can give a great method a low score

$$\text{Log-likelihood} = \sum_{(x,y)\in D} \log p_{\text{ens}}(y \mid x)$$

$$softmax(z)_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \qquad z \leftarrow \log p_{\text{ens}}(y \mid x)$$



Use *calibrated log-likelihood* instead of *log-likelihood*

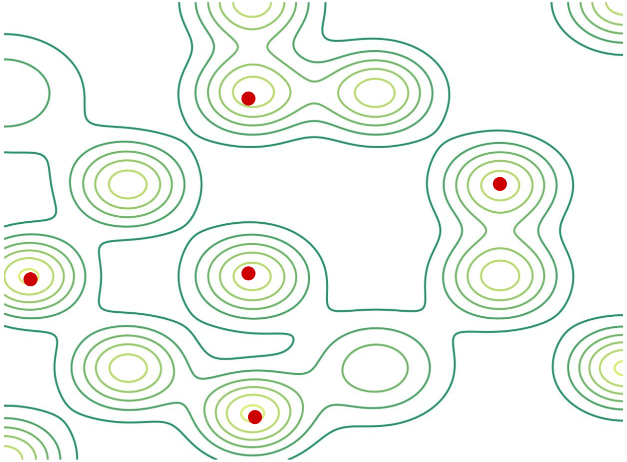
- **Brier score** (like **log-likelihood**) needs calibration ✅

- **Calibration errors** ❌
  - have model-specific biases
  - fail to provide consistent ranking depending on hyperparameters

- **Misclassification detection performance** results in incompatible values for different models ❌

Ensembles of DNNs

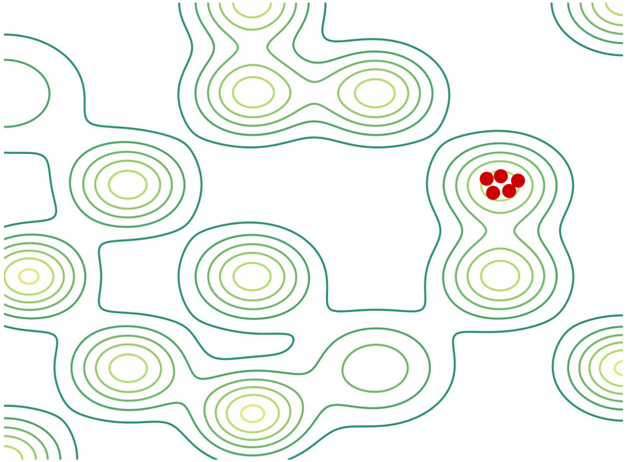
Loss landscape



Multimodal methods:

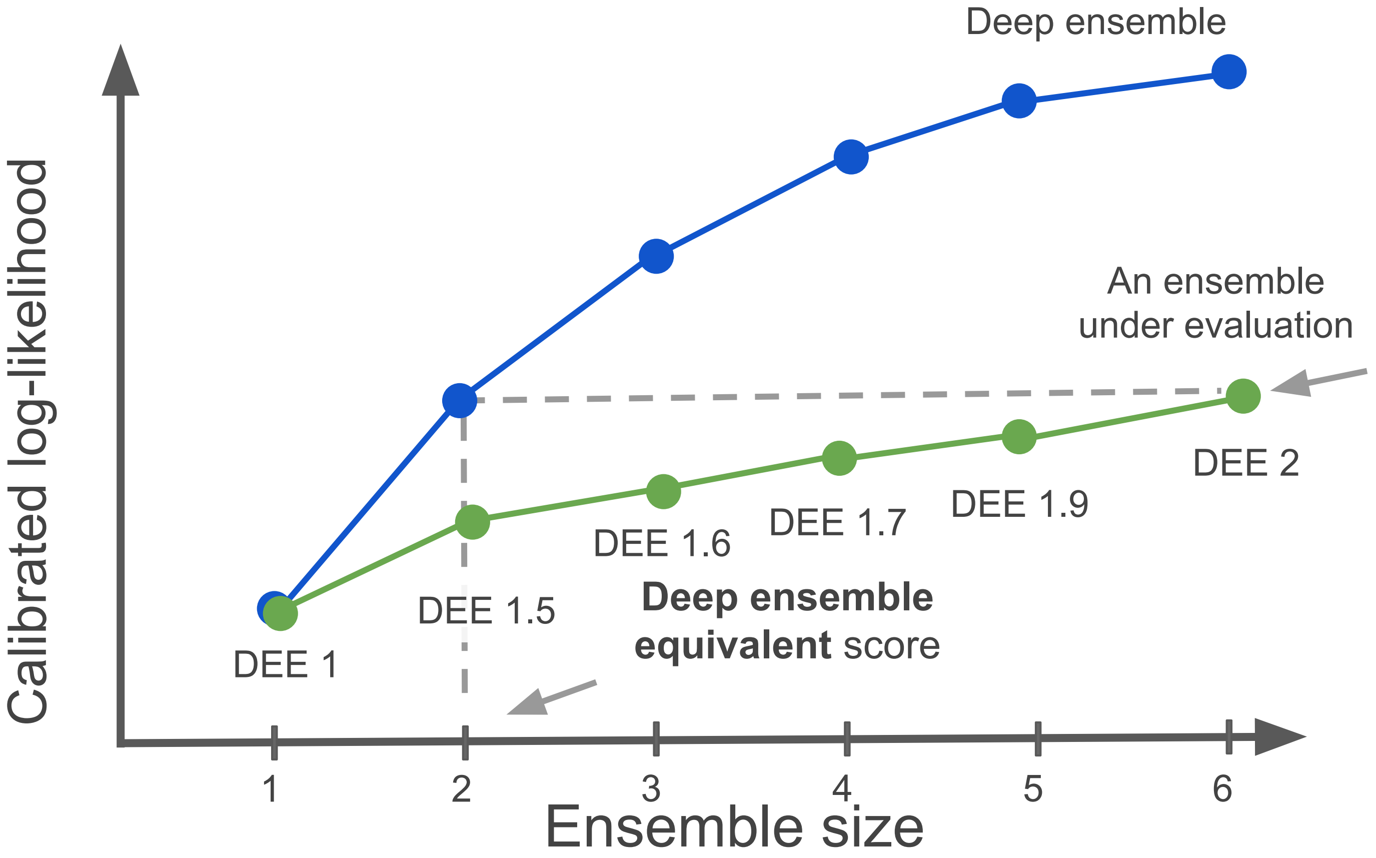
- Deep ensembles
- Snapshot ensembles
- Cyclical SGLD
- ...

Loss landscape

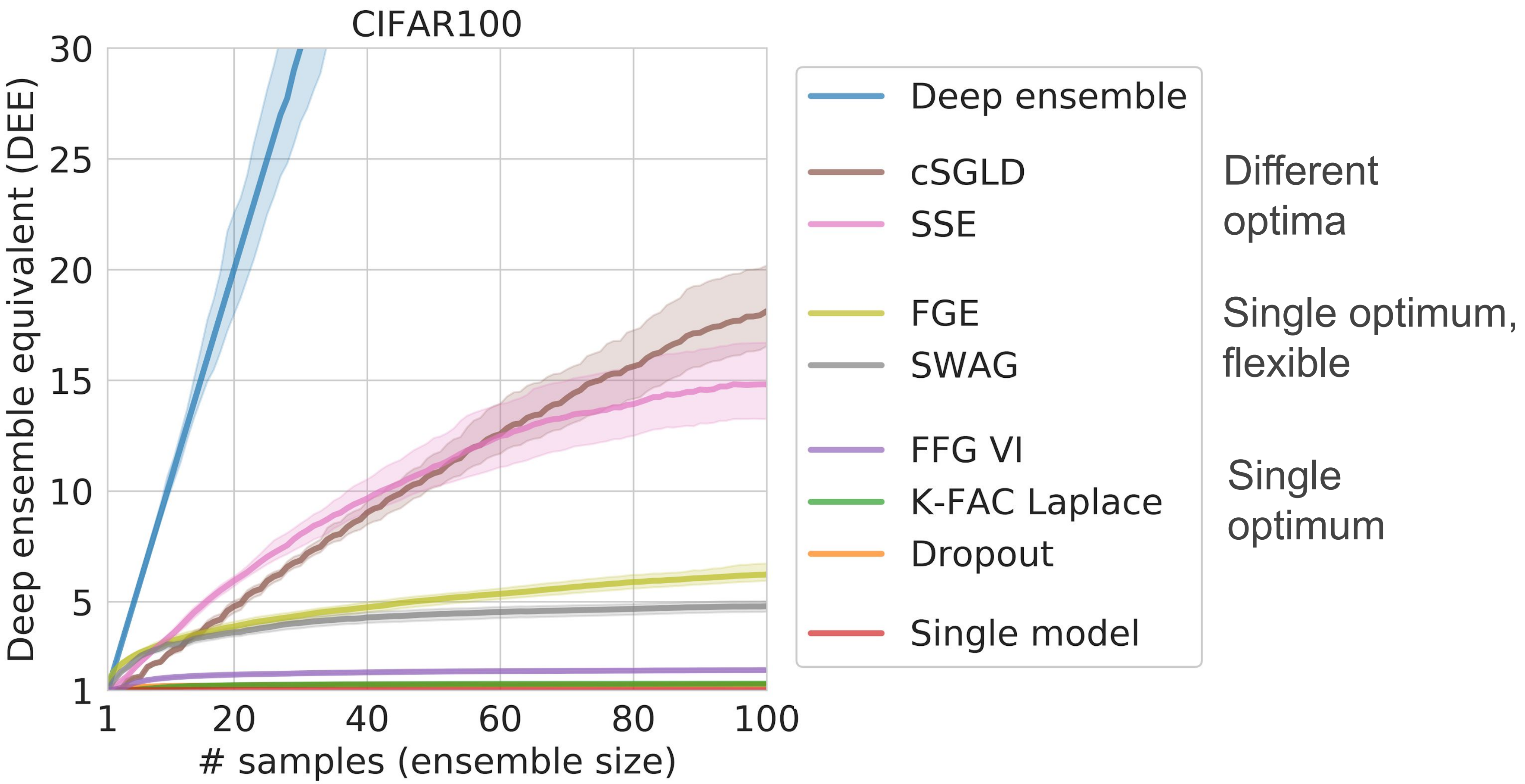


Local methods:

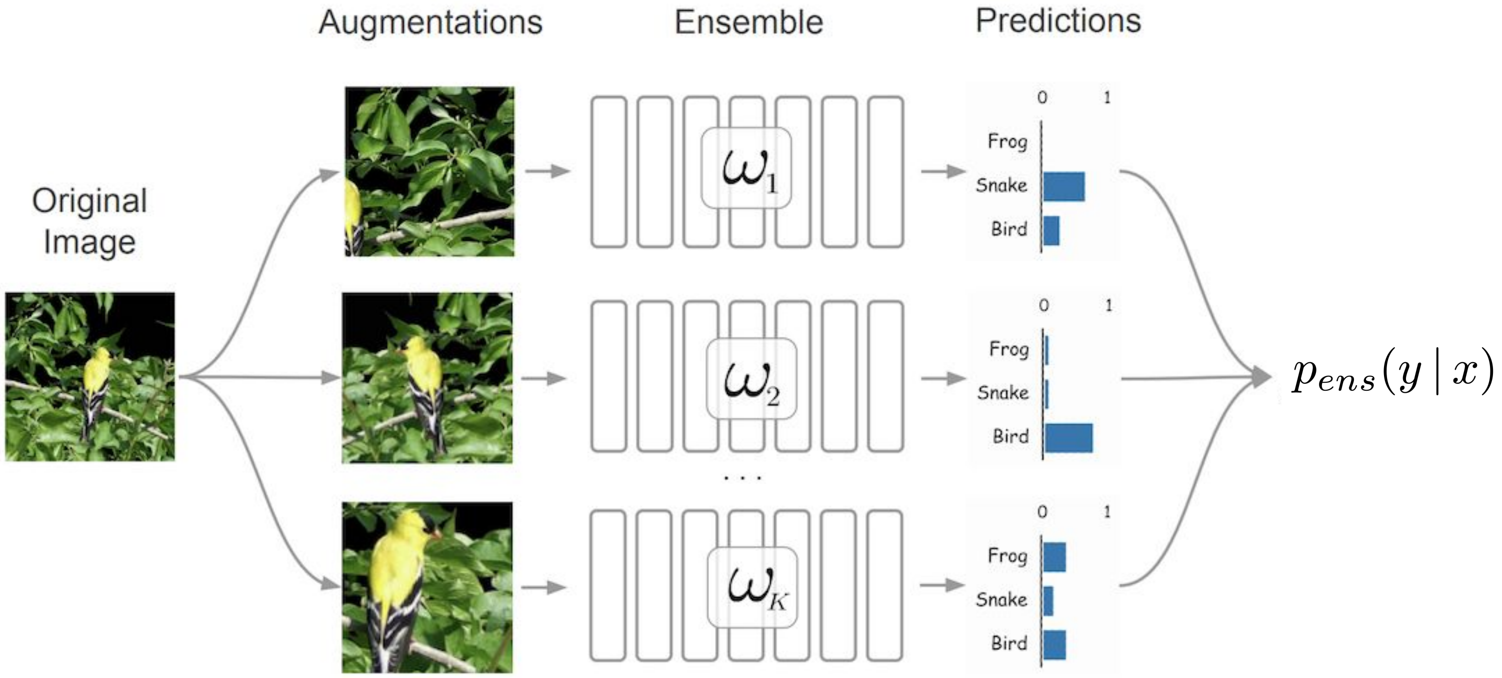
- MC-dropout
- Variational inference
- K-FAC Laplace
- Fast geometric ensembling
- SWA-Gaussian
- ...



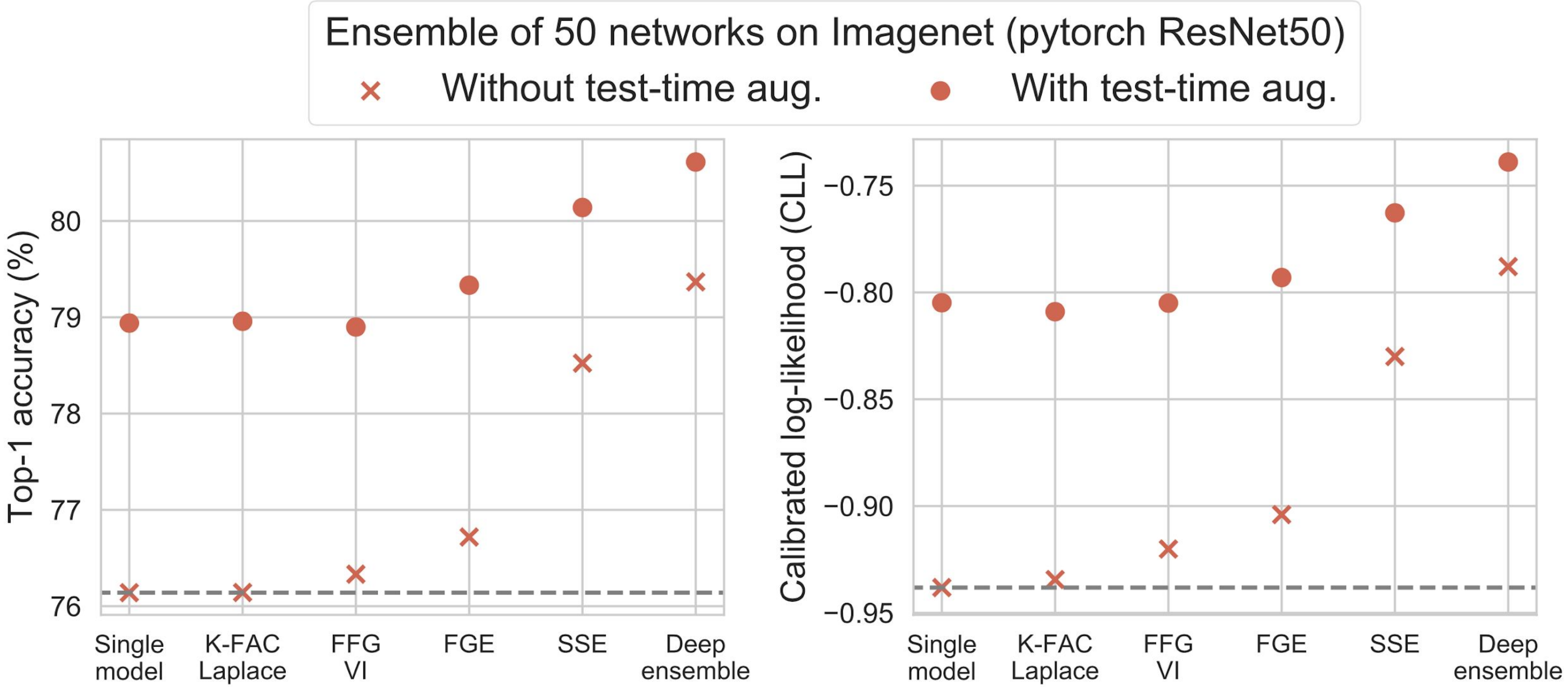
Deep ensemble equivalent score (DEE)



CIFAR100
Deep ensemble equivalent (DEE)
samples (ensemble size)
30
25
20
15
10
5
1
1
20
40
60
80
100
Deep ensemble
cSGLD
SSE
FGE
SWAG
FFG VI
K-FAC Laplace
Dropout
Single model
Different optima
Single optimum, flexible
Single optimum



Test-time data-augmentation improves ensembles for free


Ensemble of 50 networks on Imagenet (pytorch ResNet50)
Without test-time aug.
With test-time aug.
Top-1 accuracy (%)
80
79
78
77
76
Single model
K-FAC Laplace
FFG VI
FGE
SSE
Deep ensemble
Calibrated log-likelihood (CLL)
−0.75
−0.80
−0.85
−0.90
−0.95

Pitfalls of In-Domain Uncertainty Estimation & Ensembling in Deep Learning

- Metrics of in-domain uncertainty, e.g. log-likelihood, are unreliable, use *calibrated log-likelihood* instead
- Most ensembles are equivalent to a very small deep ensemble
- Test-time data augmentation improves ensembles for free

Open Review

GitHub

Forum PDF   bayesgroup/pytorch-ensembles