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

Arsenii Ashukha* Alexander Lyzhov* Dmitry Molchanov* Dmitry Vetrov

ICLR'20
Machine learning impacts critical decisions

Uncertainty Estimation

- Reliable metrics
- Strong baselines
- Wide comparison of existing techniques
Ensembles of DNNs

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

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

The metrics can give a great method a low score
Log-likelihood = \sum_{(x,y) \in D} \log p_{\text{ens}}(y \mid x)

\text{softmax}(z)_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}

z \leftarrow \log p_{\text{ens}}(y \mid x)

Use \textit{calibrated log-likelihood} instead of \textit{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

Multimodal methods:
- Deep ensembles
- Snapshot ensembles
- Cyclical SGLD
- ...

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

Loss landscape
Deep ensemble equivalent score (DEE)
Different optima
Single optimum, flexible
Single optimum

CIFAR100

Deep ensemble equivalent (DEE)

# samples (ensemble size)

Deep ensemble
cSGLD
SSE
FGE
SWAG
FFG VI
K-FAC Laplace
Dropout
Single model
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 (%): 76, 77, 78, 79, 80

Calibrated log-likelihood (CLL): -0.95, -0.90, -0.85, -0.80, -0.75

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.