Novel uncertainty models for active learning

Daniel Dold
HTWG Konstanz
Institute for Optical Systems

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Content

- What is active learning
- Current Bayesian neural networks
- Different uncertainty measures
- First results with active learning
- Outlook

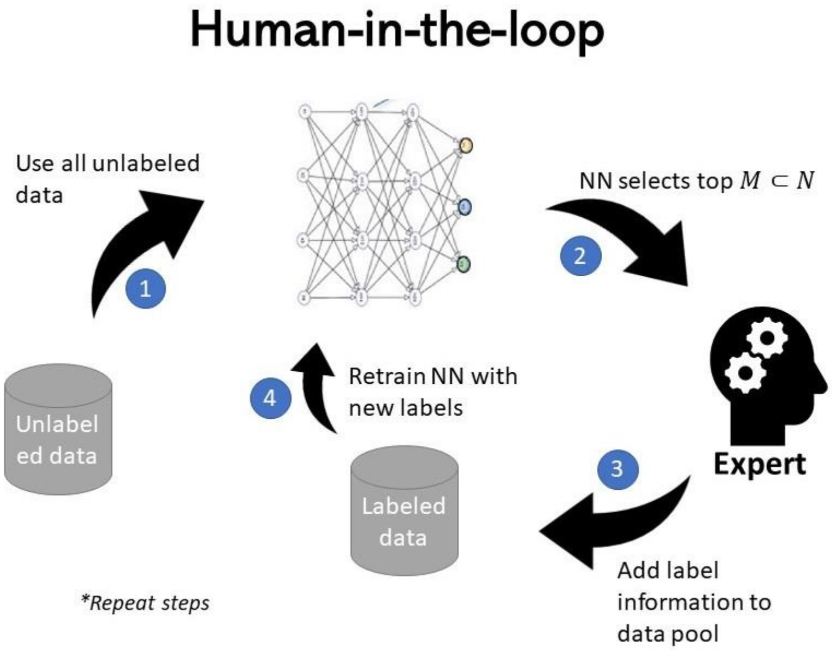
What is active learning

Active learning loop

- Labeling is expensive
  - Need of human experts

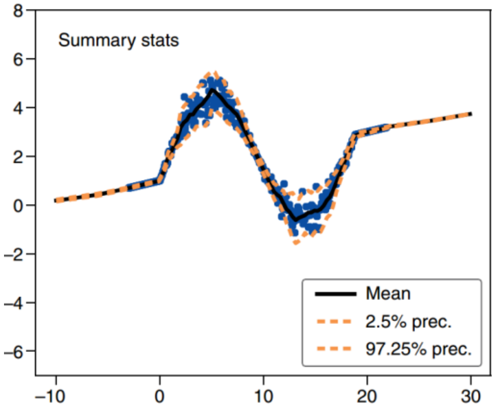
- Aim: Reduce the amount of labeling

- Idea: Use **uncertainty** to propose new candidate

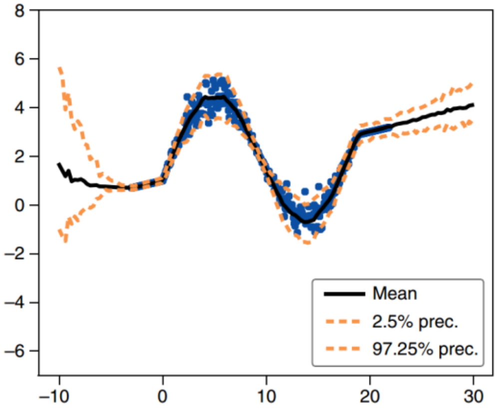


Two kind of uncertainty

Aleatoric uncertainty



Aleatoric + epistemic uncertainty



➔ **We need epistemic uncertainty for AL**

Bayesian neural networks (BNNs)

BNNs include *epistemic uncertainty*

Bayesian model averaging (BMA)

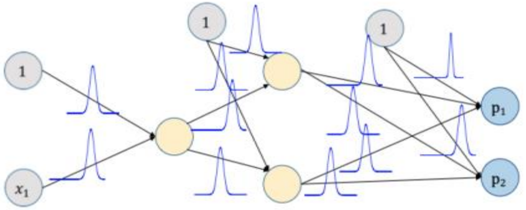
$$p(y|x,D) = \int p(y|x,\theta) \cdot p(\theta|D) d\theta$$

Posterior predictive distribution (ppd)

Using different approximations

- Variational Inference
- MC-Dropout
- Deep Ensembles
- SWAG/ MultiSWAG

Variational inference



$$p(y|x,D) = \int p(y|x,\theta)\cdot p(\theta|D)d\theta \approx \int p(y|x,\theta)\cdot \mathrm{q}_\lambda(\theta)d\theta$$

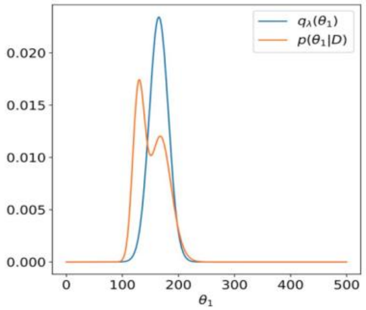
$$\boldsymbol{p(\theta|D) \approx q_\lambda(\theta)}$$

- Aim: Approximate a complicated posterior $p(\theta|D)$ with a simpler one $\mathrm{q}_\lambda(\theta)$.
- Challenge: Tune λ until variational distribution is as close as possible to the real posterior distribution

Minimize reverse KL-divergence

$$KL[q_\lambda(\theta)||p(\theta|D)] = \int q_\lambda(\theta)\log\frac{q_\lambda(\theta)}{p(\theta|D)}d\theta$$

$$\lambda^* = \mathrm{argmin}\{KL[q_\lambda(\theta)||p(\theta)] - \mathbb{E}_{\theta\sim q_\lambda}[\log p(D|\theta)]\}$$



MC-Dropout

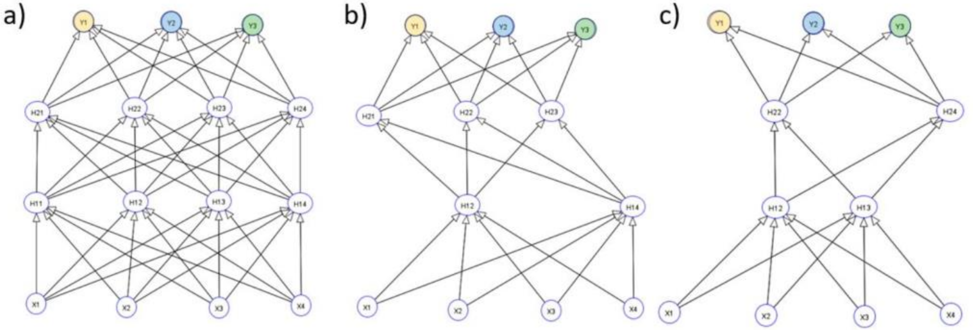
(Gal et al., 2016)*

Use Dropout during training and **inference**

$$p(y|x,D) = \int p(y|x,\theta) \cdot p(\theta|D) d\theta$$

$$\approx \int p(y|x,\theta) \cdot q_{\vartheta}^{*}(\theta) d\theta$$

$$\approx \frac{1}{T} \sum_{t=1}^{T} p(y|x,\theta_t)$$



*Gal, Y., & Uk, Z. A. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani. PMLR. http://yarin.co.

DeepEnsembles

(Lakshminarayanan et al., 2017)*

$$p(y|x,D) = \int p(y|x,\theta) \cdot p(\theta|D) d\theta \approx \frac{1}{M} \sum_{m=1}^{M} p(y|x,\theta_m)$$

- Initialize M models independently
- Train each model independent
- Training with adversarial examples (not necessary)

*Lakshminarayanan, B., Pritzel, A., & Deepmind, C. B. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.

SWAG

(Maddox et al., 2019)*1

- Extension of Stochastic Weight Averaging (SWA) (Izmailov et al., 2018)*2
- Model weights with gaussian distribution
- Algorithm:
  - Pretrain model
  - Retrain model and compute statistics after each epoch:
    - SWA ➔ compute $\bar{\theta}_{SWA}$
    - SWAG ➔ compute $\bar{\theta}_{SWA}, \Sigma_{diag}$ or $\bar{\theta}_{SWA}, \Sigma_{low-rank}$

$$p(y|x,D) = \int p(y|x,\theta) \cdot p(\theta|D) d\theta \approx \frac{1}{T}\sum_{t=1}^{T} p(y|x,\theta_t), \qquad \theta_t \sim N(\bar{\theta}_{SWA}, \Sigma)$$

- MultiSWAG: Combine SWAG with DeepEnsembles (Wilson et al., 2020)*3

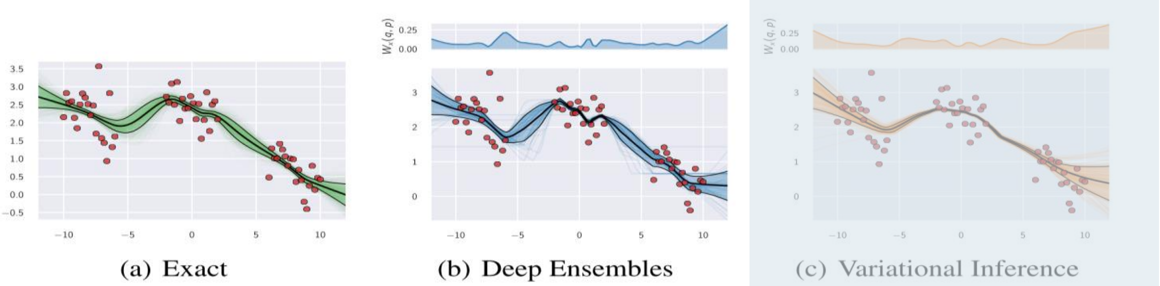
*1 Maddox, W., Garipov, T., Izmailov, P., Vetrov, D., & Wilson, A. G. (2019). A Simple Baseline for Bayesian Uncertainty in Deep Learning.

*2 Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging Weights Leads to Wider Optima and Better Generalization. 34th Conference on Uncertainty in Artificial Intelligence 2018

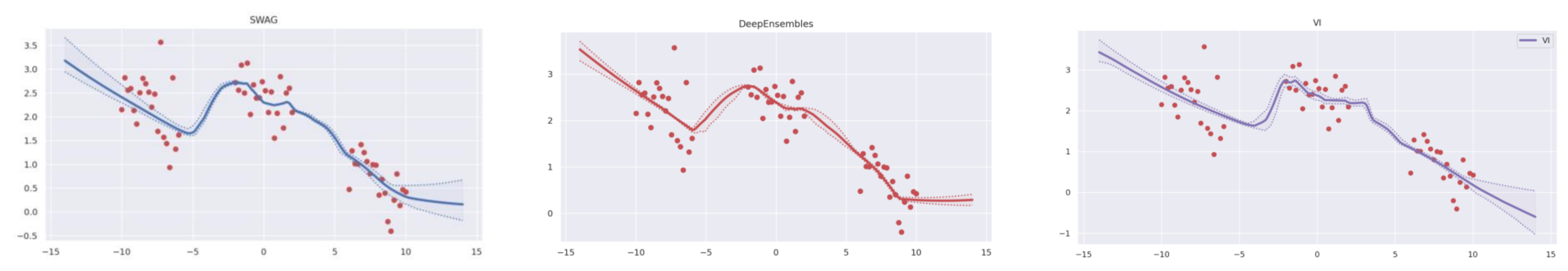
*3 Wilson, A. G., & Izmailov, P. (2020). Bayesian Deep Learning and a Probabilistic Perspective of Generalization.

First experiments on regression data

Reproduce Wilson's results*

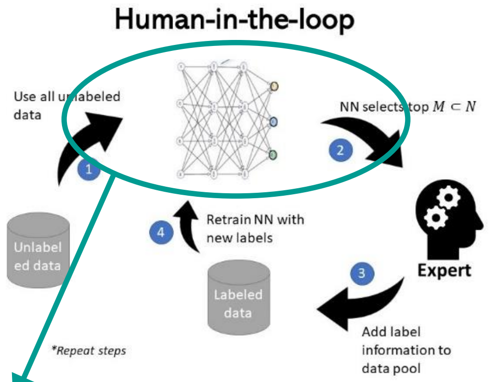


Epistemic uncertainty comparison between HMC, Deep Ensembles and VI. Image taken from Wilson*



*Wilson, A. G., & Izmailov, P. (2020). Bayesian Deep Learning and a Probabilistic Perspective of Generalization.

Parts of active learning



Data-pool ⇒ Model ⇒ **Samples** of predictive posterior distribution (ppd) ⇒ Acquisition function. Using **samples** of ppd ⇒ uncertainties ⇒ Select datapoints according uncertainty

$$ppd_{samples} \sim \mathrm{bnn}(x)$$

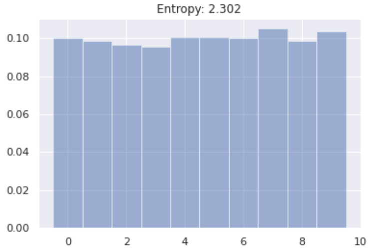
$$uncertainty = \mathrm{f_{af}}\left(ppd_{samples}\right)$$

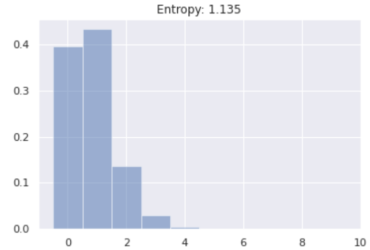
Active Learning with classification

Entropy as a measure of uncertainty

In general:

$$H = -\sum_{c=1}^{C} p_c \log p_c$$





Entropy from ppd samples of data example x

$$H(x) \approx -\sum_{c} \left(\frac{1}{T} \sum_{t=1}^{T} p(y_c | x, \theta_t) \right) \log \left(\frac{1}{T} \sum_{t=1}^{T} p(y_c | x, \theta_t) \right)$$

BALD another uncertainty measure
Bayesian Active Learning by Disagreement

- Uncertainty decomposition
  - BALD (Houlsby et al., 2011) can be interpreted as epistemic uncertainty (Depeweg et al., 2017)

$$BALD = \underbrace{\mathrm{H}[y|x,D]}_{} - \underbrace{\mathbb{E}_{\theta\sim p(\theta|D)}\left[\mathrm{H}[\mathrm{y}|\mathrm{x},\theta]\right]}_{}$$

$$\mathrm{BALD} \approx \underbrace{-\sum_{c}\left(\frac{1}{T}\sum_{t=1}^{T} p(y_c|x,\theta_t)\right)\log\left(\frac{1}{T}\sum_{t=1}^{T} p(y_c|x,\theta_t)\right)}_{\sum uncertainty} + \underbrace{\frac{1}{T}\sum_{t,c} p(y_c|x,\theta_t)\log p(y_c|x,\theta_t)}_{aleatoric\ uncertainty}$$

*Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian Active Learning for Classification and Preference Learning.

Acquisition functions (classification)

$$\overline{p}(y_c|x) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} p(y_c|x,\theta_t)$$

- Entropy

$$H(x) \approx -\sum_c \overline{p}(y_c|x)\log\overline{p}(y_c|x)$$

- BALD (Houlsby et al., 2011)

$$I[y,\theta|x,D] \approx -\sum_c \overline{p}(y_c|x)\log\overline{p}(y_c|x) + \frac{1}{T}\sum_{t,c} p(y_c|x,\theta_t)\log p(y_c|x,\theta_t)$$

- Variation-ratio (Freeman, 1965)

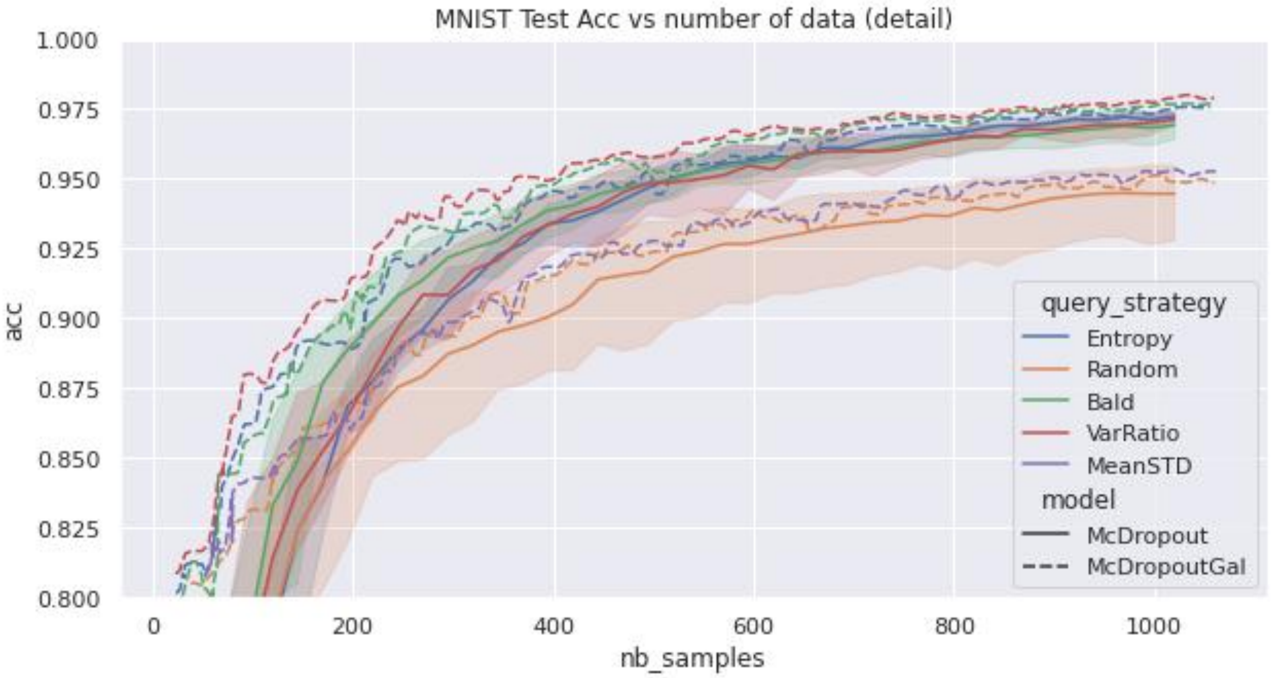
$$\mathrm{Var}Ratio(x) = 1 - \max_y \overline{p}(y_c|x)$$

- Random

Primary results with AL

On MNIST dataset

Compare with Gal*



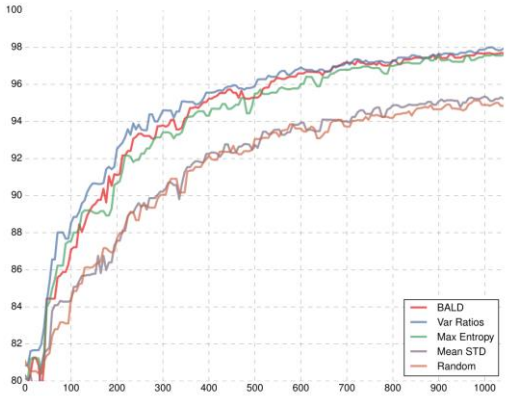
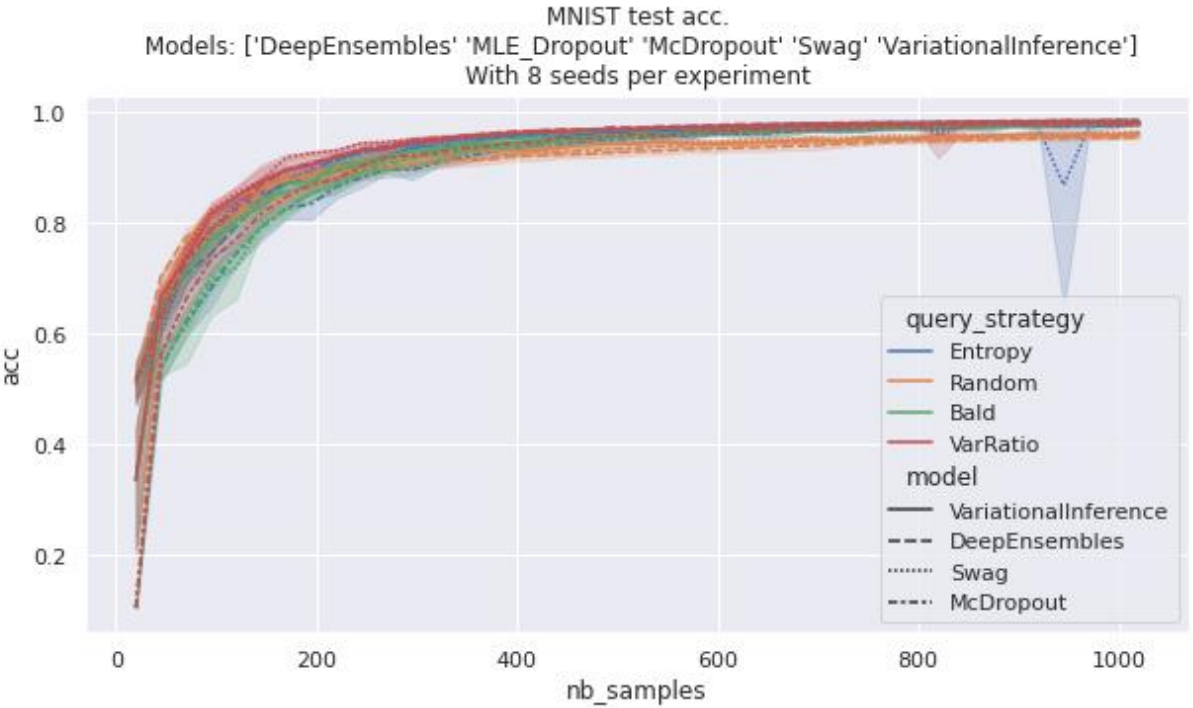


Image from Gal*

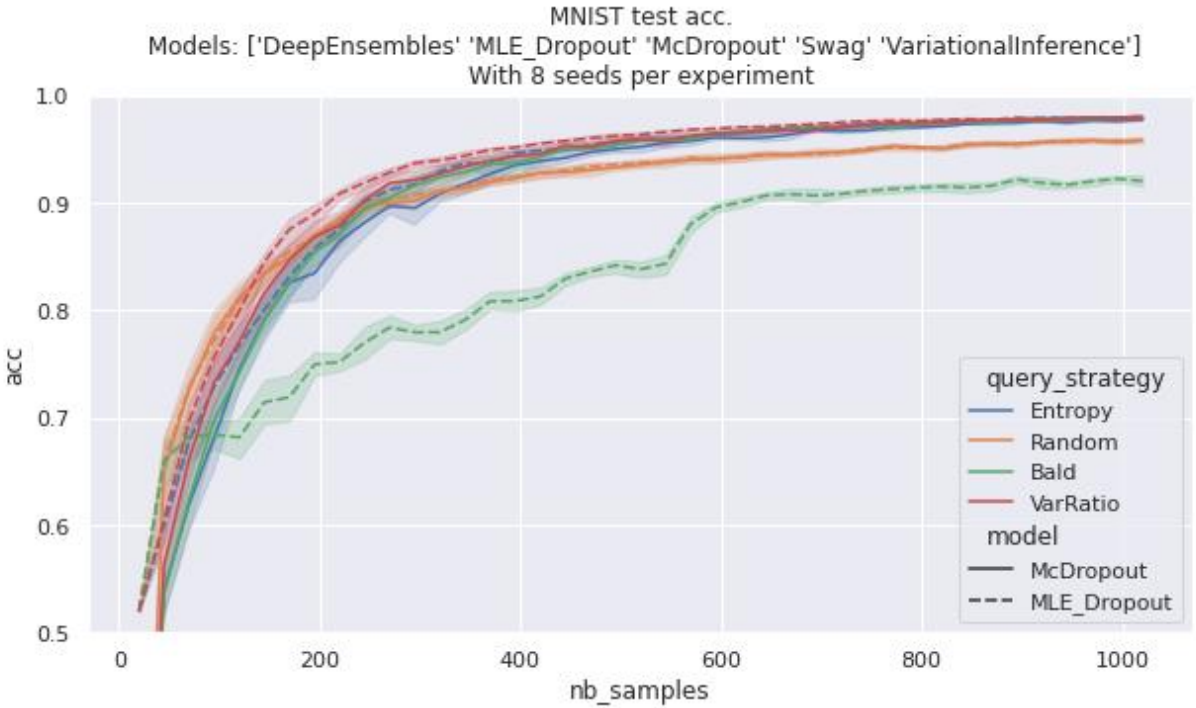
*Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep Bayesian Active Learning with Image Data. 34th International Conference on Machine Learning, ICML 2017

Compare ACC



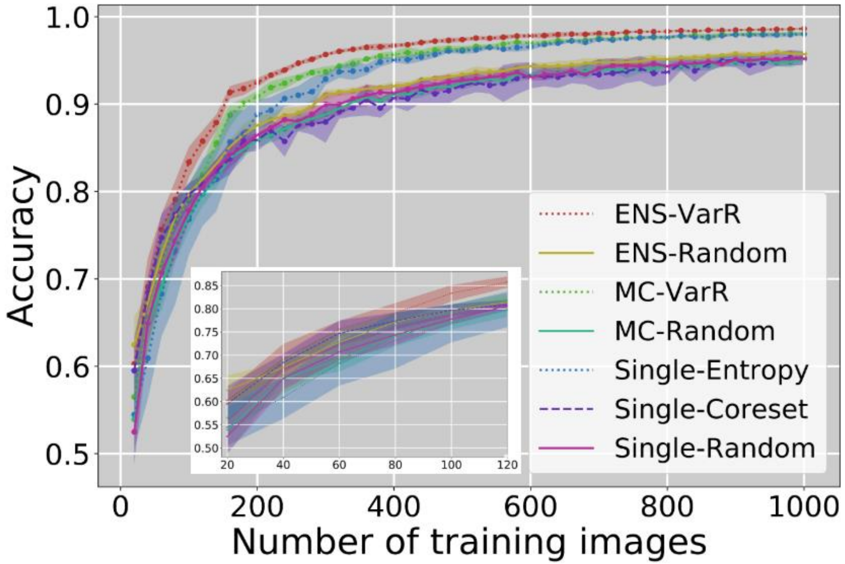
Model	Query strategy	$\overline{AUC}$	Standard error	N
DeepEnsembles	Bald	0,908	3,12E-03	4
DeepEnsembles	Entropy	0,913	2,26E-03	4
DeepEnsembles	Random	0,888	3,43E-03	4
DeepEnsembles	VarRatio	**0,918**	3,40E-03	4
McDropout	Bald	0,885	1,43E-03	8
McDropout	Entropy	0,880	2,77E-03	8
McDropout	Random	0,882	1,57E-03	8
McDropout	VarRatio	0,892	2,12E-03	8
Swag	Bald	0,888	2,80E-03	8
Swag	Entropy	0,907	3,60E-03	8
Swag	Random	0,897	1,76E-03	8
Swag	VarRatio	**0,918**	1,74E-03	8
VariationalInference	Bald	0,902	2,31E-03	5
VariationalInference	Entropy	0,907	1,85E-03	5
VariationalInference	Random	0,895	1,80E-03	5
VariationalInference	VarRatio	0,916	9,57E-04	5

Compare MLE solution with BNNs

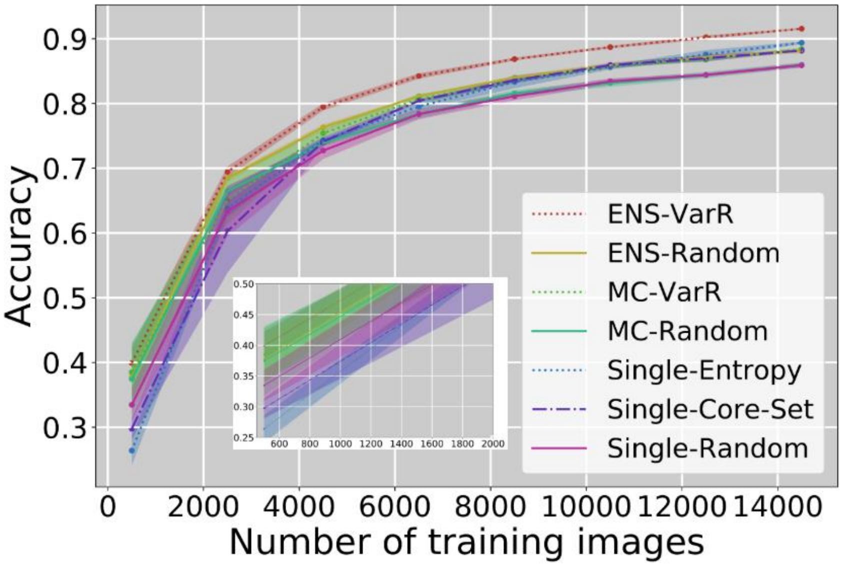


Model	Query strategy	$\overline{AUC}$	Standard error	N
McDropout	Bald	0,885	1,43E-03	8
McDropout	Entropy	0,880	2,77E-03	8
McDropout	Random	0,882	1,57E-03	8
McDropout	VarRatio	0,892	2,12E-03	8
MLE_Dropout	Bald	0,815	2,60E-03	8
MLE_Dropout	Entropy	0,897	1,20E-03	8
MLE_Dropout	Random	0,888	1,45E-03	8
MLE_Dropout	VarRatio	**0,907**	7,70E-04	8

The power of ensembles for active learning in image classification (Beluch et al., 2018)*



(a) MNIST on S-CNN



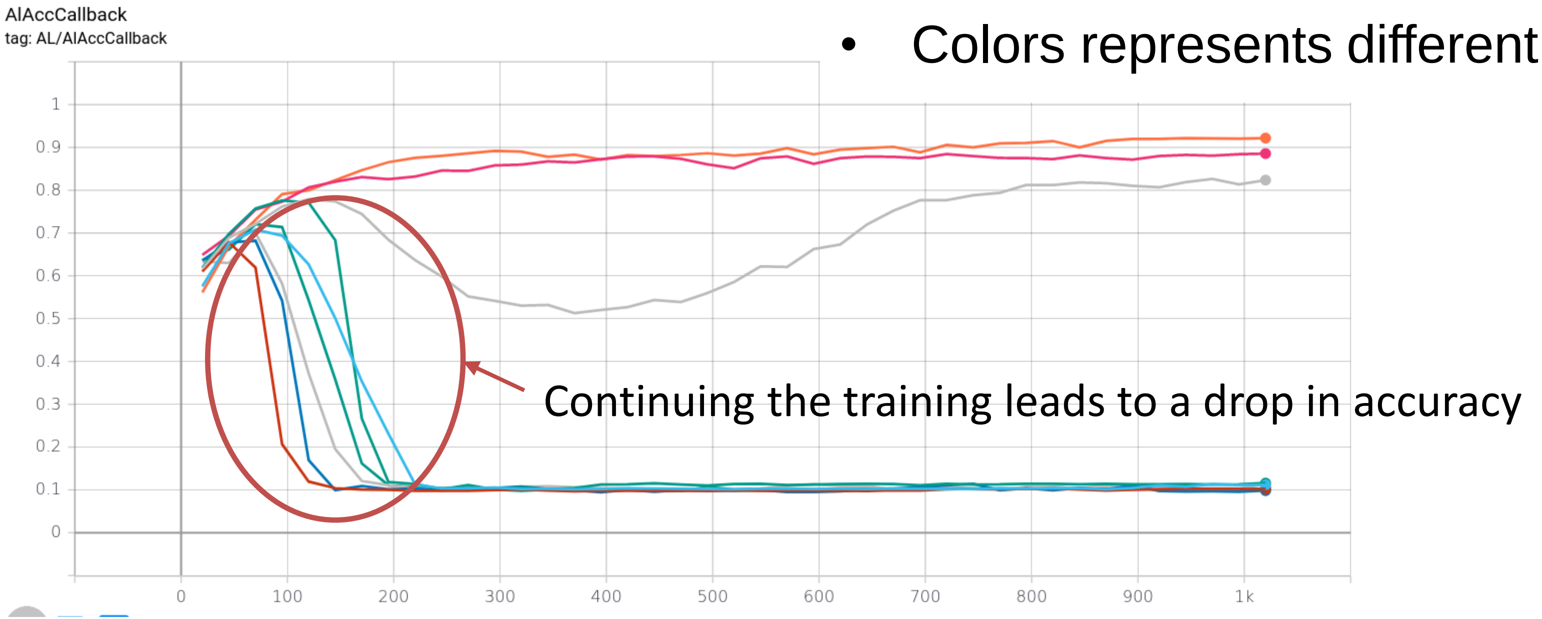
(b) CIFAR-10 on DenseNet

*Beluch Bcai, W. H., Nürnberger, A., & Bcai, J. M. K. (2018). The power of ensembles for active learning in image classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Don’t cut corners

Continue Training

- VI with Random acquisition function
- Colors represents different seeds



First results on MNIST

- MNIST dataset is too simple
  - No statement about models or acquisition functions
  - No advantage due to epistemic uncertainty
- Using acquisition functions motivated by uncertainty performs better than random.
- AL requires retraining from scratch.

Thanks for your attention