UPPSALA UNIVERSITET

Jiahao Lu
Jiahao.Lu.2199@student.uu.se

Markus Sagen
Markus.John.Sagen@gmail.com

Jianbo Li
Jianbo.Li.4196@student.uu.se

Supervisors

Nataša Sladoje
Joakim Lindblad

Project in Computational Science 2020

Estimating Certainty in Deep Learning

Project Goals

- **Implement** several state-of-the-art methods to reach well-calibrated certainty estimates for deep learning based classification task
- **Evaluate** their performances
  - with different **models**
  - on two **datasets**
  - of several **metrics**

Motivation

Deep neural networks tend to be overconfident in their predictions.

Well-calibrated models are essential for trustworthy decision making.

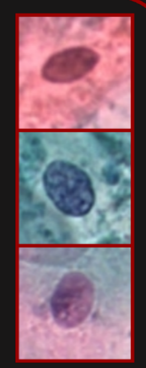
Expressing uncertainty is crucial for high-stakes applications, like oral cancer screening or self-driving cars.

Material and Methods

Data

MNIST
Classes: 10
Size: (28, 28, 1)
Train: 1,200
Val: 300
Test: 10,000

OralCancer[1] (OC)
Classes: 2
Size: (80, 80, 3)
Train: 65,973
Val: 7,330
Test: 55,514



Pre-processing

Hard Label (LS0.0)
OneHot encoding of labels

Label Smoothing[2] (LS0.1)
Soften the targets, e.g., (0, 1)→(0.1, 0.9)

BNN Models

Deterministic Model (Base)
Task output of SoftMax as model's confidence

Monte Carlo Dropout[3] (Drop)
Dropout layers stay active also in test phase

Concrete Dropout[4] (CDrop & LLCDrop)
Auto-tune the Dropout rate with a continuous relaxation of Dropout's discrete masks

Flipout[5] (VI & LLVI)
Decorrelate the gradients by implicitly sampling pseudo-independent weight perturbations for each sample

Post-processing

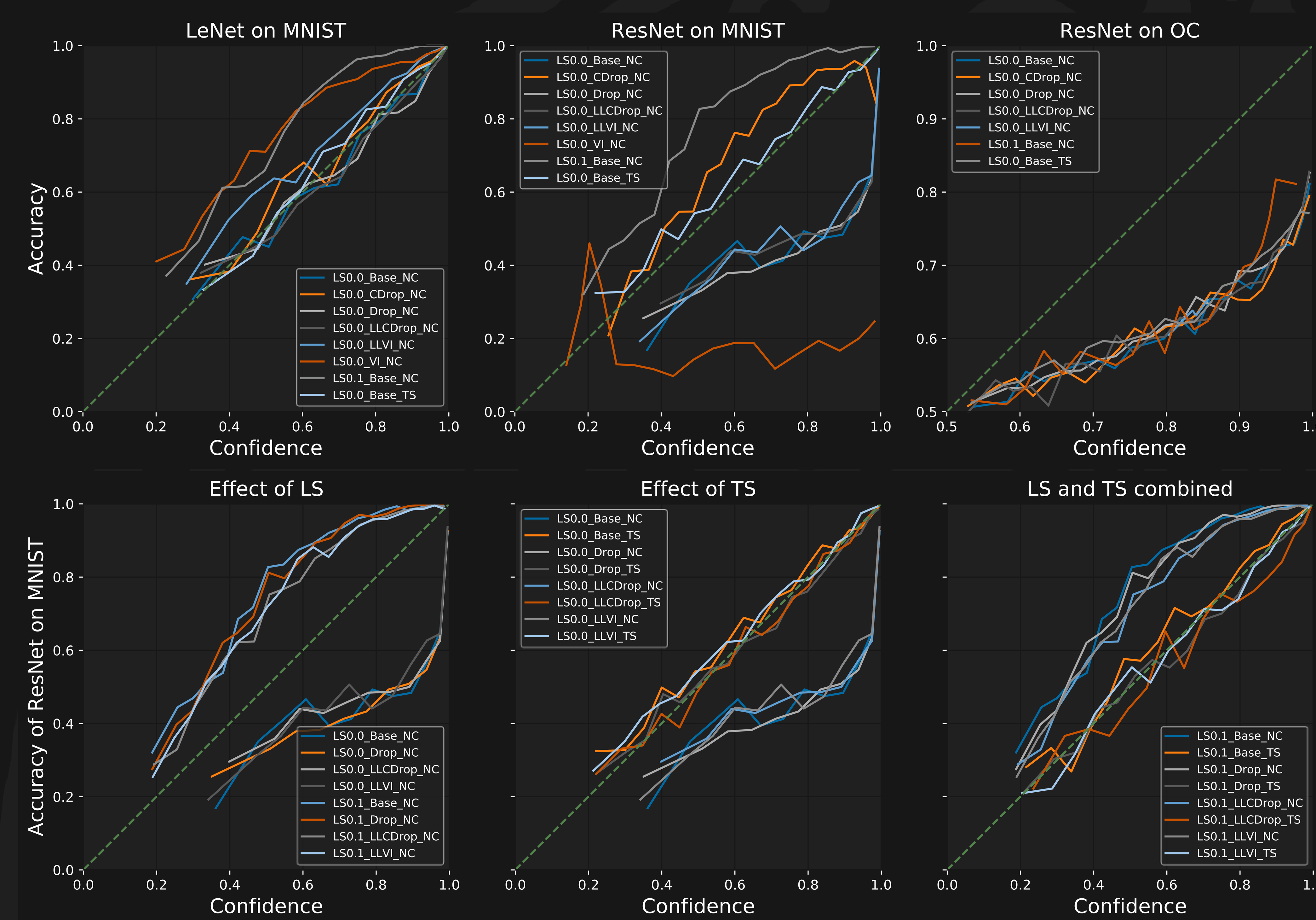
No Calibration (NC)

Temperature Scaling[6] (TS)
Learn a scalar from validation set to rescale input of SoftMax

Results

Label Smoothing	Method	Calibration	Accuracy	AECE[7]	Train time (s/epoch)	Test time (s)
LS0.0	Base	NC	0.801	0.145	**21.61**	**1.14**
LS0.1	Base	NC	**0.841**	0.170	**21.57**	**1.13**
LS0.0	Base	TS	0.809	**0.022**	**21.35**	**1.14**
LS0.0	Drop	NC	0.823	0.133	**20.42**	10.97
LS0.0	CDrop	NC	0.778	0.096	41.73	22.77
LS0.0	LLCDrop	NC	0.821	0.134	21.95	11.72
LS0.0	VI	NC	0.211	0.232	48.86	23.51
LS0.0	LLVI	NC	0.822	0.125	23.05	11.57

Table: Sole method comparison of ResNet on MNIST



Discussion

- BNN methods do not improve calibration significantly.
- LS improves accuracy and mitigates overconfidence.
- TS efficiently assures well-calibrated models.
- LS and TS can be simply combined.
- Applying CDrop/VI only on the last layer reduces train/test time.
- ResNet tends to be worse calibrated than LeNet.

References:
[1] G. Forslid *et al.*, "Deep Convolutional Neural Networks for Detecting Cellular Changes Due to Malignancy," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, 2017, pp. 82–89.
[2] R. Müller *et al.*, "When Does Label Smoothing Help?," *arXiv:1906.02629 [cs, stat]*, Jun. 2019.
[3] Y. Gal *et al.*, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.
[4] Y. Gal *et al.*, "Concrete Dropout," *arXiv:1705.07832 [stat]*, May 2017.
[5] Y. Wen *et al.*, "Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches," *arXiv:1803.04386 [cs, stat]*, Apr. 2018.
[6] C. Guo *et al.*, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1321–1330.
[7]Y. Ding et al., "Evaluation of Neural Network Uncertainty Estimation with Application to Resource-Constrained Platforms," arXiv:1903.02050 [cs, stat], Mar. 2019.