Simple Regularisation for Uncertainty-Aware Knowledge Distillation

Martin Ferianc[1] and Miguel Rodrigues[1]
[1]Department of Electronic and Electrical Engineering, University College London, London, UK
martin.ferianc.19@ucl.ac.uk, @MartinFerianc, martinferianc, martinferianc.github.io/
Work in progress

Summary

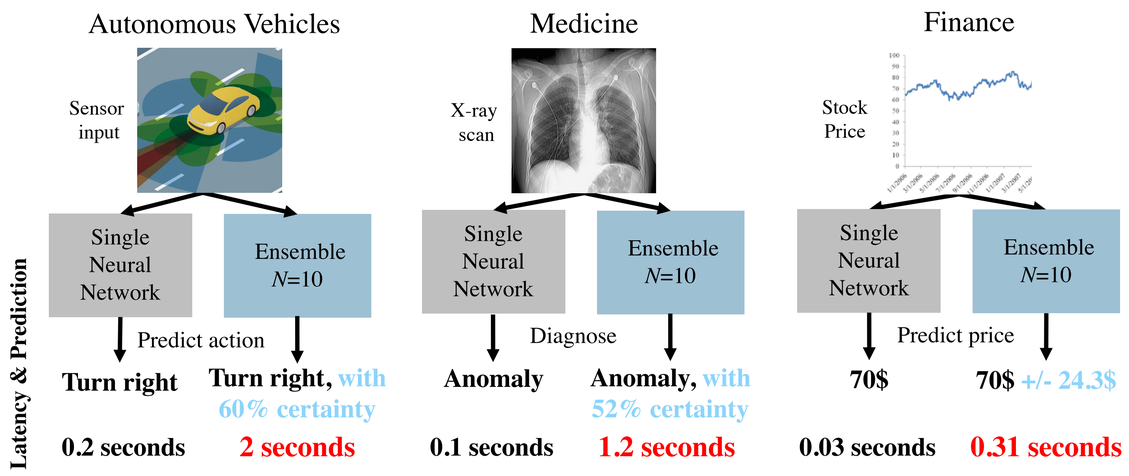
- **What is this about?** Weight regularisation of multi-headed neural nets during knowledge distillation for capturing ensemble's uncertainty and generalisation.
- **What is the problem?** Ensembles, can quantify their uncertainty, but they are $\mathcal{O}(N)$ slower than a single net. N is the ensemble size.
- **How is it addressed?** Distill relevant knowledge from ensemble to a student while preserving generalisation and uncertainty estimation of the ensemble.
- **Contribution: Weight regularisation promoting diversity towards better uncertainty estimation in knowledge distillation for multi-headed neural net architectures.**

Introduction

Uncertainty quantification (UQ) in machine learning is important for understanding what a model *does not know* and to build trust with users. **Ensembles of neural nets** [1] are simple and **good at a) uncertainty estimation and b) generalisation**.

Neural nets are not able to estimate their uncertainty *by default*. However, we can **train** N nets with the same architecture and use the nets to give N predictions to **empirically estimate both their aleatoric and epistemic uncertainty** [1, 2].

Application: UQ is essential for **safety-critical and regulated real-world applications**, where observing a prediction made by a network is not enough.



Challenge: Ensembles of neural nets are $\mathcal{O}(N)$ **slower and more resource expensive to run**, than a single neural network, where N is ensemble size.

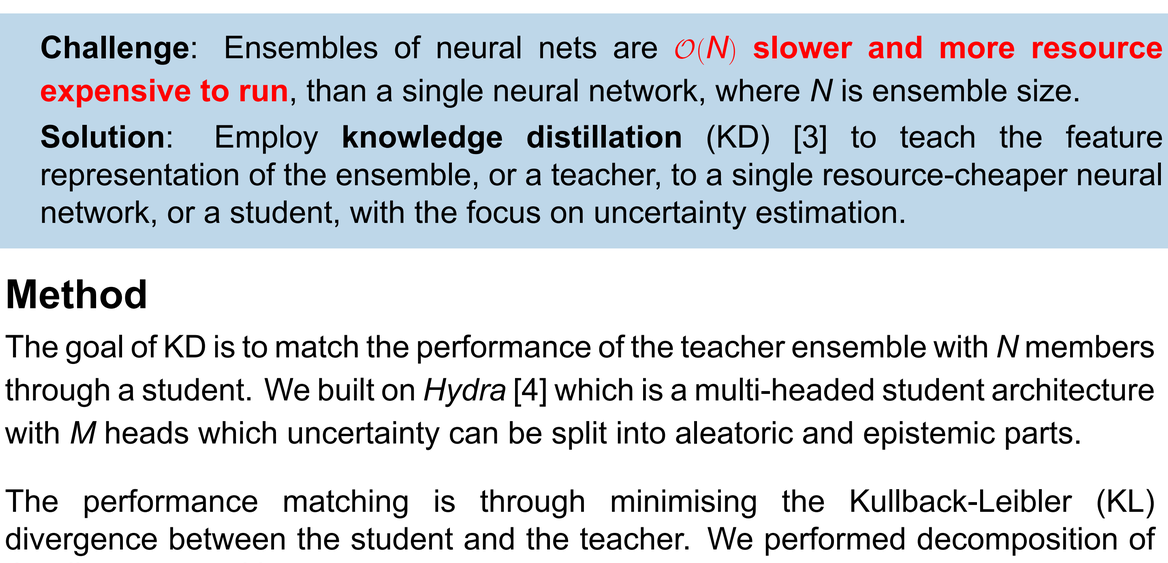
Solution: Employ **knowledge distillation** (KD) [3] to teach the feature representation of the ensemble, or a teacher, to a single resource-cheaper neural network, or a student, with the focus on uncertainty estimation.

Method

The goal of KD is to match the performance of the teacher ensemble with N members through a student. We built on *Hydra* [4] which is a multi-headed student architecture with M heads which uncertainty can be split into aleatoric and epistemic parts.

The performance matching is through minimising the Kullback-Leibler (KL) divergence between the student and the teacher. We performed decomposition of the divergence with respect to:

L_1 **correctness,** L_2 **aggregation** [3], L_3 **individuality** [4] and added L_4 **diversity**.



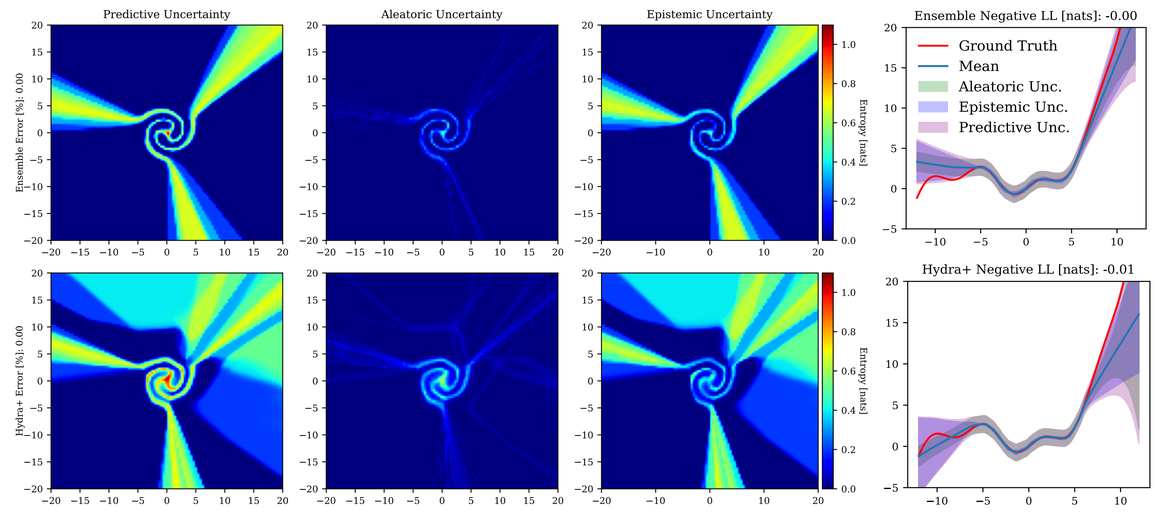
Student with a core and $M = 2$ heads is matched with ensemble consisting of $N = 3$ members. Training is performed with respect to the 4 loss components: L_1 correctness, L_2 aggregation, L_3 individuality and L_4 diversity.

Diversity: Define the mean head weight at a level as $\boldsymbol{\theta}_S^{mean} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\theta}_S^m$.

$$L_4 = \sum_{l=1}^{L}\sum_{m=1}^{M} \frac{1 + \cos(\boldsymbol{\theta}_S^{l,mean}, \boldsymbol{\theta}_S^{l,m})}{2}$$

Force the weights to be less similar to each other to create disagreement and diverse feature representations. Results in better calibration and uncertainty estimation.

Experiments

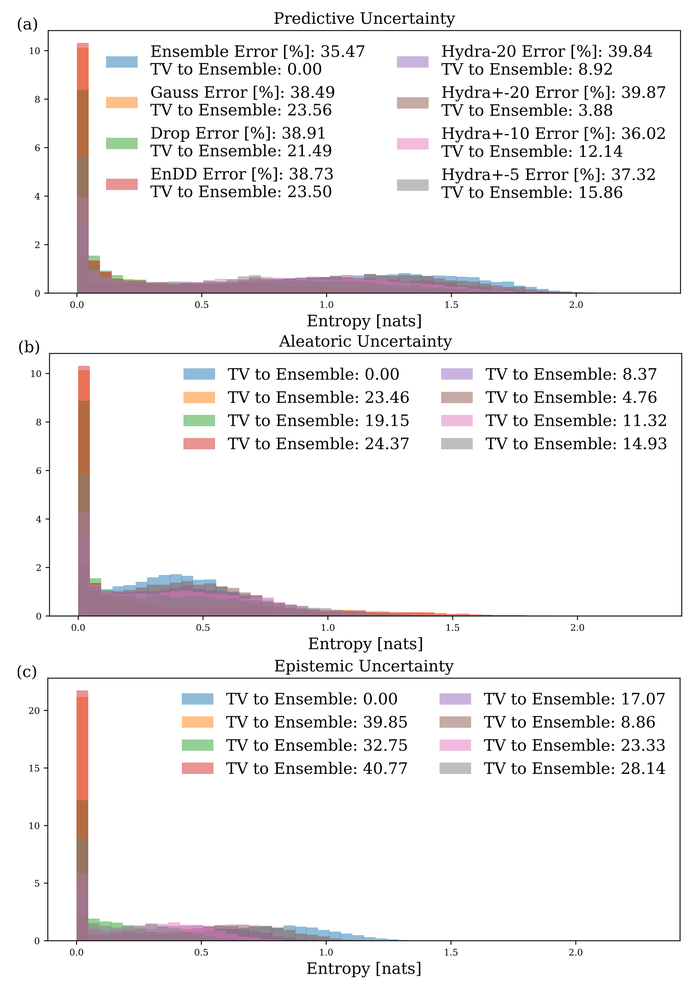


Toy classification (left) and regression (right) comparing ensemble performance and the discussed Hydra+ in uncertainty decomposition and generalisation.

Method	Error [%]	ECE [%]	#FLOPS [M]	#Params [M]
Ensemble-20	7.20±0.05	3.81±0.11	117.75	3.04
Gauss	8.53±0.21	6.08±0.13	13.48	0.13
Drop	8.70±0.14	5.17±0.08	15.07	0.12
EnDD	8.91±0.08	6.45±0.11	13.48	0.12
Hydra-20	7.49±0.06	3.10±0.10	15.07	1.71
Hydra+-20	7.56±0.06	3.08±0.03	15.07	1.71
Hydra+-10	7.59±0.04	3.14±0.01	14.23	0.88
Hydra+-5	7.67±0.08	3.69±0.08	13.81	0.46

Method	Error [%]	ECE [%]	#FLOPS [G]	#Params [M]
Ensemble-20	10.92±0.10	4.62±0.20	2.81	55.95
Gauss	11.70±0.10	9.53±0.09	0.80	4.55
Drop	11.61±0.19	8.84±0.14	0.80	4.55
EnDD	11.80±0.29	9.70±0.34	0.80	4.55
Hydra-20	11.44±0.28	4.83±0.14	1.04	19.23
Hydra+-20	11.54±0.16	3.74±0.10	1.04	19.23
Hydra+-10	11.37±0.05	4.98±0.07	0.92	11.51
Hydra+-5	11.29±0.07	6.47±0.11	0.85	7.64

Comparison on the test dataset of SVHN (left) and CIFAR-10 (right). The number after method name denotes N or M. ECE is expected calibration error #FLOPS, #Param are the number of FLOPS and parameters.



Uncertainty decomposition for augmentations applied to the CIFAR-10 test set and the compared methods for one seed/experiment. To create an out of distribution dataset we applied a 30% vertical shift to the test part of the dataset. (a) is the total uncertainty, (b) is the aleatoric uncertainty (c) is the epistemic uncertainty. TV denotes total variation to the ensemble histogram and Error denotes the error on the augmented test set.

Hydra+ **was able to better mimic the uncertainty estimation** than the compared methods with only a small loss in the generalisation performance.

Key-Takeaway

It is possible to distil uncertainty estimation along with the generalisation performance of the ensemble without having to observe additional data through regularisation.

Room for Improvement and Future Work

Manually defined student architecture, manually defined the number of heads M, if N or the ensemble size is small it is very difficult to distil any uncertainty.

Paper

Code

Acknowledgements: Martin Ferianc was sponsored through a scholarship from the Institute of Communications and Connected Systems at UCL and through the PhD Enrichment scheme at The Alan Turing Institute. Lastly, we thank DFUQ'22 reviewers for feedback and encouragement.

References

[1] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[2] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] L. Tran, B. S. Veeling, K. Roth, J. Swiatkowski, J. V. Dillon, J. Snoek, S. Mandt, T. Salimans, S. Nowozin, and R. Jenatton, "Hydra: Preserving ensemble diversity for model distillation," *arXiv preprint arXiv:2001.04694*, 2020.