Last Layer Empirical Bayes

Valentin Villecroze, Yixin Wang, Gabriel Loaiza-Ganem

Layer 6 AI, Toronto
University of Michigan

Abstract

- We introduce **LLEB**, a well-motivated method for uncertainty quantification (**UQ**) in neural networks.
- LLEB is based on an interpretation of ensembles as **empirical Bayes** [1].
- LLEB performs on par, but does not outperforms, existing UQ approaches.

Background on UQ

- For a classifier $p(y|x,\theta)$, UQ methods provide a **distribution** $q^*(\theta)$ **over weights θ**, rather than a single θ^*, e.g.:
  - MC dropout [2]: $q^*(\theta)$ is obtained by keeping dropout on at test time.
  - Bayesian neural networks use variational inference to optimize the ELBO,

$$\mathbb{E}_{q^*(\theta)}[\log p(\mathcal{D}|\theta)] - \mathbb{KL}(q(\theta)\|\pi(\theta))$$

  - Deep ensembles [3]: M models are independently trained and $q^*(\theta)$ is given by equally weighting each model, i.e.

$$q^*(\theta) = \sum_{m=1}^{M} \delta_{\theta_m^*}(\theta)$$

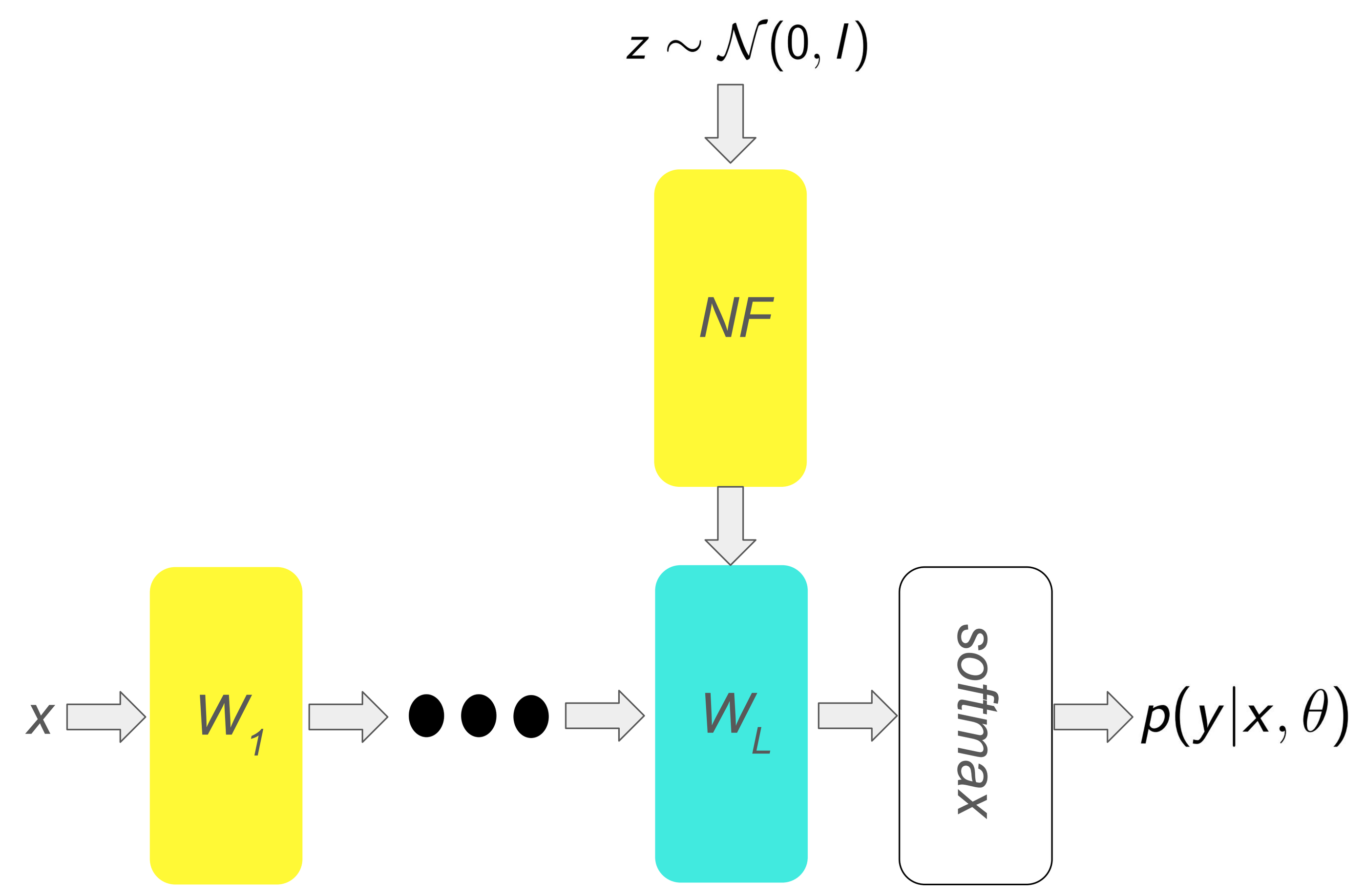
    - Ensembles are expensive to train but are often considered the gold standard for UQ.
- Averaging over $q^*(\theta)$, i.e. $p(y|x) = \mathbb{E}_{q^*(\theta)}[p(y|x,\theta)]$, is used to make predictions.
- The variability of predictions over $\theta \sim q^*(\theta)$ quantifies uncertainty, e.g. $\sum_y \mathrm{var}_{q^*(\theta)}[p(y|x,\theta)]$
- **LLEB is a way to obtain a new** $q^*(\theta)$ **using normalizing flows.**

Last Layer Empirical Bayes

- We train $q^*(\theta)$ as a normalizing flow.
  - The flow can be trained along with the classifier or using a pre-trained and fixed classifier.
  - The training objective is given by:

$$\max_q \mathbb{E}_{q(\theta)}[\log p(\mathcal{D}|\theta)]$$

  - For tractability, we only use the flow on the last layer.
    - Defining $q^*(\theta)$ only on the last layer has worked elsewhere in UQ [4].



Why is LLEB sensible?

- Intuitively, the flow prevents collapse onto a point mass and allows to find a distribution over optimal parameter configurations.
- The prior $\pi(\theta)$ in the ELBO is usually fixed, learning it is called **empirical Bayes**.
  - If the ELBO is also optimized over the prior $\pi(\theta)$, then $\pi^*(\theta)$ and $q^*(\theta)$ are optimal if and only if:
    - $\pi^*(\theta) = q^*(\theta)$
    - $q^*(\theta)$ places all its mass on the set of likelihood-maximizing values of θ
- This is implicitly what ensembles do: they maximize the ELBO with an infinitely flexible prior [1].
- **LLEB attempts to optimize this objective with a more adequately strong prior.**

Results

	Train/Test: MNIST, OOD: Fashion-MNIST			Train/Test: Fashion-MNIST, OOD: MNIST		
Method	Acc. (↑)	ECE (↓)	AUC (↑)	Acc. (↑)	ECE (↓)	AUC (↑)
Default	98.02 ± 0.05	$\mathbf{0.00 \pm 0.00}$	-	88.02 ± 0.10	$\mathbf{0.01 \pm 0.00}$	-
LLL	98.02 ± 0.05	0.75 ± 0.00	$\mathbf{0.96 \pm 0.00}$	88.02 ± 0.10	0.66 ± 0.00	$\mathbf{0.82 \pm 0.01}$
MCD	$\mathbf{98.50 \pm 0.05}$	0.01 ± 0.00	0.91 ± 0.00	$\mathbf{88.47 \pm 0.05}$	0.02 ± 0.00	0.75 ± 0.03
LLEB (ours)	97.74 ± 0.24	$\mathbf{0.00 \pm 0.00}$	0.95 ± 0.01	87.83 ± 0.37	$\mathbf{0.01 \pm 0.00}$	0.72 ± 0.03
Default ($M = 5$)	98.26 ± 0.02	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.97 \pm 0.00}$	88.71 ± 0.09	$\mathbf{0.02 \pm 0.00}$	0.84 ± 0.01
LLL ($M = 5$)	98.26 ± 0.02	0.76 ± 0.00	0.96 ± 0.00	88.71 ± 0.09	0.67 ± 0.00	0.87 ± 0.01
MCD ($M = 5$)	$\mathbf{98.69} \pm 0.02$	0.02 ± 0.00	0.95 ± 0.00	89.34 ± 0.08	0.04 ± 0.00	$\mathbf{0.89 \pm 0.00}$
LLEB ($M = 5$, ours)	98.30 ± 0.08	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.97 \pm 0.00}$	$\mathbf{89.44 \pm 0.16}$	0.03 ± 0.00	$\mathbf{0.89 \pm 0.01}$

	Train/Test: CIFAR-10, OOD: SVHN			Train/Test: SVHN, OOD: CIFAR-10		
Method	Acc. (↑)	ECE (↓)	AUC (↑)	Acc. (↑)	ECE (↓)	AUC (↑)
Default	92.82 ± 0.09	$\mathbf{0.05 \pm 0.00}$	-	$\mathbf{95.26 \pm 0.03}$	0.03 ± 0.00	-
LLL	92.82 ± 0.09	0.70 ± 0.00	$\mathbf{0.94 \pm 0.01}$	$\mathbf{95.26 \pm 0.03}$	0.73 ± 0.00	$\mathbf{0.92 \pm 0.00}$
MCD	92.29 ± 0.09	0.10 ± 0.01	0.89 ± 0.02	95.11 ± 0.05	0.09 ± 0.01	0.89 ± 0.00
LLEB (ours)	$\mathbf{92.85 \pm 0.09}$	0.06 ± 0.00	$\mathbf{0.94 \pm 0.01}$	95.23 ± 0.03	$\mathbf{0.02 \pm 0.01}$	0.86 ± 0.01
Default ($M = 5$)	$\mathbf{94.82 \pm 0.01}$	$\mathbf{0.01 \pm 0.00}$	0.91 ± 0.01	$\mathbf{96.55 \pm 0.03}$	$\mathbf{0.01 \pm 0.00}$	0.97 ± 0.00
LLL ($M = 5$)	$\mathbf{94.82 \pm 0.01}$	0.73 ± 0.00	0.90 ± 0.01	$\mathbf{96.55 \pm 0.03}$	0.74 ± 0.00	0.97 ± 0.00
MCD ($M = 5$)	94.72 ± 0.04	0.12 ± 0.00	0.93 ± 0.01	96.54 ± 0.02	0.11 ± 0.00	$\mathbf{0.98 \pm 0.00}$
LLEB ($M = 5$, ours)	94.78 ± 0.01	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.95 \pm 0.01}$	96.52 ± 0.03	$\mathbf{0.01 \pm 0.00}$	$\mathbf{0.98 \pm 0.00}$

References

[1] *Deep Ensembles Secretly Perform Empirical Bayes*, Loaiza-Ganem et al., 2025
[2] *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, Gal and Ghahramani, ICLR 2016
[3] *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*, Lakshminarayanan et al., NeurIPS 2017
[4] *Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLu networks*, Kristiadi et al., ICLR 2020