Zifeng Wang web: https://zifengwang.xyz/ email: wangzf18@mails.tsinghua.edu.cn

Research Proposal

of Zifeng Wang (CS DPhil applicant for Fall, 2021)

It was witnessed that machine learning (ML) and deep learning (DL) have been achieving exciting performance on various tasks, including computer vision, natural language processing, and reinforcement learning. Meanwhile, there raises a concern of the reliability and interpretability of existing methods, as models often fall short into making inaccurate predictions when receiving noisy inputs. In this circumstance, exploiting uncertainty in model parameters and predictions appeals to researchers. At present, research in uncertainty-aware ML centers around two aspects: developing methods for estimating uncertainty and utilizing uncertainty metric for various ML tasks.

I. Uncertainty Estimation

Accompanied with the brilliant sucess of deep learning, Bayesian deep learning thrives as the Bayesian counterpart of the frequentist deep learning. Traditional arts advocated to varitional inference (VI) [3]. In the Bayesian perspective of learning, we would like to infer a *posterior* distribution over the model weights $\boldsymbol{\omega}$, i.e., $p(\boldsymbol{\omega}|\mathcal{D})$, when a training set $\mathcal{D} = \{\mathbf{z}_i\}_{i=1}^N$ is observed. However, computing $p(\boldsymbol{\omega}|\mathcal{D})$ is intractable, hence VI proposed to establish a model $q_\theta(\boldsymbol{\omega})$ which is parameterized by θ, e.g., $\theta = (\mu, \Sigma)$ are mean and covariance if $q_\theta(\boldsymbol{\omega})$ is a Gaussian model. With a predefined prior $p(\boldsymbol{\omega})$, we could minimize a variational evidence lower bound (ELBO) as

$$\mathcal{L}(\theta) = -\mathrm{KL}(q_\theta(\boldsymbol{\omega})\|p(\boldsymbol{\omega})) + \mathcal{L}_\mathcal{D}(\theta), \tag{1}$$

where $\mathcal{L}_\mathcal{D}(\theta)$ is likelihood on the dataset. Nonetheless, in order to optimize ELBO, we need to (1) let $q_\theta(\boldsymbol{\omega})$ an isotropic Gaussian or diagonal covariance Gaussian and (2) execute Monte Carlo sampling to compute this objective function. These disadvantages undermine the popularity of VI in DL fields. We need a more light-weighted method to seamlessly adapt the vanilla DL models to the Bayesian version for inferring uncertainties.

To this end, a surge of research proposed to obtain DL uncertainty, including Monte Carlo dropout [4], Deep Ensembles [7], DUQ [10], and DUN [2]. Gal and Ghahramani [4] reinterpreted dropout as a Bayesian approximation of the Gaussian process (GP) [15], thus proving that performing forward passes with turning on dropout is equivalent to Monte Carlo sampling. This method is easy-to-implement on networks with dropout modules but relies on coarse approximations for ensuring scalability, then often results in limited or unreliable estimates. Lakshminarayanan et al. [7] trains multiple networks with different random seeds and regards this process as Monte Carlo sampling, but this method is too expensive in computation. van Amersfoort et al. [10] proposed to use an RBF network [8] to quantify uncertainty from a deterministic network. This method obtains excellent uncertainty in a single forward and maintains competitive performance, however, it requires an RBF module hence not aligned with the most state-of-the-art network architectures. Antorán et al. [2] assumed depth of networks as random variables to exploit the sequential structure of feed-forward networks, which also allows Bayesian inference in a single pass. In DUN, the network depth is assumed following a categorical distribution with trainable parameters. ELBO is then estimated and optimized for updating both weight and depth parameters. This method views each block of the network as a black box, though can yield predictive uncertainty, it cannot explain the exact parameter uncertainty in networks.

II. Applications of Bayesian Deep Learning

As the above mentioned, the uncertainty estimate has remained an open problem, which means there still lies space for me to delve deeper. Besides, I am keen on this line of research because it has many applications in ML such as curriculum learning, active learning, out-of-distribution (OOD) data detection, information-theoretic representation learning, etc.

Curriculum learning. The estimated weight distribution can be used for inferring predictive uncertainty on each sample. In particular, we can quantify the variance of prediction by deriving both

epistemic and aleatoric uncertainties [13]

$$\begin{aligned}&\mathrm{Var}_{q(\mathbf{y}^*)}[\mathbf{y}^*] = \mathbb{E}_{q(\mathbf{y}^*)}[\mathbf{y}^{*\otimes 2}] - \mathbb{E}_{q(\mathbf{y}^*)}[\mathbf{y}^*]^{\otimes 2} \\ &= \underbrace{\int \left\{\mathrm{diag}\{\mathbb{E}_{p(\mathbf{y}^*)}[\mathbf{y}^*]\} - \mathbb{E}_{p(\mathbf{y}^*)}[\mathbf{y}^*]^{\otimes 2}\right\} q_\theta(\boldsymbol{\omega}) d\boldsymbol{\omega}}_{\text{aleatoric}} + \underbrace{\int \left\{\mathbb{E}_{p(\mathbf{y}^*)}[\mathbf{y}^*] - \mathbb{E}_{q(\mathbf{y}^*)}[\mathbf{y}^*]\right\}^{\otimes 2} q_\theta(\boldsymbol{\omega}) d\boldsymbol{\omega}}_{\text{epistemic}},\end{aligned} \tag{2}$$

where $v^{\otimes 2} \triangleq vv^\top$, and $\mathrm{diag}(v)$ represents a diagonal matrix whose element vector is v. This uncertainty metric can be used for measuring confidence when model deals with each sample, thus rearranging sampling orders of training data for boosting model generalization ability.

Active learning. The definition of acquisition function varies in the literature. One of those is mutual information between predictions and model posterior (BALD) [5] defined as

$$\mathbb{I}[\mathbf{y}; \boldsymbol{\omega} | \mathbf{x}, \mathcal{D}] = \mathbb{H}[\mathbf{y}|\mathbf{x}, \mathcal{D}] - \mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D})}\left[\mathbb{H}[\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}]\right], \tag{3}$$

where $\mathbb{I}[x; y]$ is the mutual information between variable x and y and $\mathbb{H}[x]$ is the entropy of x. Computing these metrics requires the posterior $p(\boldsymbol{\omega}|\mathcal{D})$ which is often replaced by $q_\theta(\boldsymbol{\omega})$. Hence modeling the weight distribution $q_\theta(\boldsymbol{\omega})$ is necessary in this scenario.

Out-of-distribution data detection. OOD data indicate outliers which differ from other training data, e.g., the CIFAR-10 images mixed into SVHN. Empirical observations indicate that models are prone to inferring high predictive uncertainty on those outliers. In this regard, obtaining a good estimate of model uncertainty is helpful for detecting malicious samples.

Information-theoretic representation learning. Recent research in information bottleneck (IB) revealed the intrinsic training process of deep neural networks [9], which inspired a line of work on adapting IB [1] or mutual information [6] to learn informative and robust representation. Moreover, Xu and Raginsky [16] proved a non-vacuous error bound under mutual information perspective:

$$|\mathrm{gen}(p(\boldsymbol{\omega}|\mathcal{D}))| \leq \sqrt{\frac{2\sigma^2}{n}\mathbb{I}[\mathcal{D}; \boldsymbol{\omega}]}, \tag{4}$$

where $\mathrm{gen}(p(\boldsymbol{\omega}|\mathcal{D}))$ is the expected generalization error and $\mathbb{I}[\mathcal{D}; \boldsymbol{\omega}]$ indicates the information of dataset contained in model weights. This work bridges the gap between information theory and representation, while is intractable on deterministic networks.

III. Future Research Plans

In my previous research, we adopted jackknife to quantify model confidence for sample selection [14], then extended this tool to DL models [12]. Besides, I worked on causal representation learning for improving the information richness of representations [11]. Recently, my preliminary research [13] demonstrated the potential use of jackknife for uncertainty estimates. My research vision is to develop the next generation of uncertainty-supported DL techniques that are effective and scalable, then develop models for various applications as the aforementioned. In the sequel, I aim at building a more advanced AI system for causal inference under uncertainty. I am specifically excited about applications of my methods in healthcare and the Internet of Things (IoT), where a tremendous amount of data can be collected, and reliability is especially demanding. To more specific, my Ph.D. research attempts to solve the following sub-objectives:

- To derive an unified uncertainty quantification framework for common DL architectures;
- To adopt this technique in representation learning and ML for healthcare and IoT considering uncertainty;
- To exploit uncertainty in causal representation learning and further explore causal inference with uncertainty estimates.

I strongly believe that the above research directions can advance the machine learning research and applications, and will have impacts in both academic and industry.

References

[1] Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. *arXiv preprint arXiv:2006.08437*, 2020.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[5] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192, 2017.

[6] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, pages 1–5. IEEE, 2015.

[10] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. 2020.

[11] Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan E. Kuruoglu, and Yefeng Zheng. Information theoretic counterfactual learning from missing-not-at-random feedback. In *Neural Information Processing Systems*, 2020.

[12] Zifeng Wang, Rui Wen, Xi Chen, Shao-Lun Huang, Ningyu Zhang, and Yefeng Zheng. Finding influential instances for distantly supervised relation extraction. *arXiv preprint arXiv:2009.09841*, 2020.

[13] Zifeng Wang, Rui Wen, Xi Chen, Ercan Engin Kuruoglu, Shao-Lun Huang, and Yefeng Zheng. Uncertainty-guided curriculum learning via infinitesimal jackknife. Technical report, 2020.

[14] Zifeng Wang, Hong Zhu, Zhenhua Dong, Xiuqiang He, and Shao-Lun Huang. Less is better: Unweighted data subsampling via influence function. In *AAAI Conference on Artificial Intelligence*, pages 6340–6347, 2020.

[15] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[16] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.