Uncertainty Estimation with Deep Networks

Anirudh Seth
aniset@kth.se

September 2020

1 Introduction

Neural Networks have proven to be immensely powerful and achieved state-of-the-art performance on wide range of domains like natural language processing [21], computer vision [15] , weather forecasting [2] and medical diagnosis [19]. Despite achieving impressive scores on benchmark tasks ,NN's are prone to produce overconfident predictions [8],[6]. These overconfident predictions are a result of poor assessment and quantification of predictive uncertainty especially when dealing with data scarcity.

The uncertainty in predictions can arise due to the lack of knowledge (or understanding) of the model a.k.a distributional uncertainty [13] or epistemic uncertainty [3]. Aleatoric uncertainty on the other hand is irreducible and is representative of the unknowns that differ on each run of the experiment. Majority of the real world problems also suffer from the problem of dataset shift [17] ,a mismatch between the distributions of your training and test dataset.With the increased application of NN's in safety critical tasks like perception systems for autonomous vehicles [16], medical diagnosis [19] etc. estimating and evaluating the quality of predictive uncertainty of the model has become a crucial task.

This essay summaries the research on Bayesian approaches - Bayes by Backprop [1] as well as Non-Bayesian approaches like Ensembles with random-initialization [11]. and Ensemble Distribution Distillation (EnD2) [14] to tackle the aforementioned task.Bayes by Backprop builds upon the prior application of variational inference on the stochastic hidden units of an autoencoder [10],[18],[18] . The authors use the same approach but on the weights of the network and emperically show an improvement in predictive performance especially when dealing with limited data. Lakshminarayanan et al., 2017 [11] describe a simple yet scalable method for estimating predictive uncertainty using (i) ensembles and (ii) adversarial training [4] and also propose a series of tasks for evaluating the quality of the estimates, in terms of calibration and generalization. The ensembles are constructed using a non bayesian approach by training on the same data with different random seeds.Ensemble Distribution Distillation [14] adopts a bayesian viewpoint,the author make use of a class of models, known as Prior Networks [13] to model the distribution of the ensemble predictions rather than just the mean.The key idea is to preserve both the distributional information and improved classification performance of an ensemble within a single neural network.

2 Methods

2.1 Weight Uncertainty in Neural Networks [1]

Bayesian inference on NN, uses the posterior distribution of the weights,$P(\mathbf{w} \mid \mathcal{D})$,to evaluate the predictive distribution of an unknown label $\hat{y}$ with input $\hat{x}$ using

$$P(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}) = E_{P(\mathbf{w}\mid\mathcal{D})}[P(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}, \mathbf{w})] \tag{1}$$

$$= \int p\left(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}, \mathbf{w}\right) p(\mathbf{w} \mid \mathcal{D}) d\mathbf{w} \tag{2}$$

Taking the above expectation requires an integral which is intractable.Blundell, et al. [1] introduced Bayes by Backprop , a solution to the intracatbility of the above equation.Based on the work of Hinton and Van Camp [7] and Graves [5] the author propose the use of a variational approximation $q(\mathbf{w} \mid \theta)$ that minimizes the Kullback-Leibler (KL) divergence to the true posterior distribution $P(\mathbf{w} \mid \theta)$.

$$\theta^{\star} = \arg\min_{\theta} \mathrm{KL}[q(\mathbf{w} \mid \theta) \| P(\mathbf{w} \mid \mathcal{D})] \tag{3}$$

$$= \arg\min_{\theta} \int q(\mathbf{w} \mid \theta) \log \frac{q(\mathbf{w} \mid \theta)}{P(\mathbf{w}) P(\mathcal{D} \mid \mathbf{w})} \mathrm{d}\mathbf{w} \tag{4}$$

$$= \arg\min_{\theta} \mathrm{KL}[q(\mathbf{w} \mid \theta) \| P(\mathbf{w})] - E_{q(\mathbf{w} \mid \theta)}[\log P(\mathcal{D} \mid \mathbf{w})] \tag{5}$$

The second step is to sample from the simpler distribution and approximate the exact cost (3) as

$$\mathcal{F}(\mathcal{D}, \theta) \approx \textstyle\sum_{i=1}^{n} \log q\left(\mathbf{w}^{(i)} \mid \theta\right) - \log P\left(\mathbf{w}^{(i)}\right) - \log P\left(\mathcal{D} \mid \mathbf{w}^{(i)}\right) \tag{6}$$

where $w(i)$ denotes the i^{th} Monte Carlo sample drawn from the variational posterior $q(\mathbf{w} \mid \theta)$ .In order to calculate derivatives of the variational distribution parameters , mean and standard deviation (for a Gaussian distribution), the local reparameterisation trick [9] is utilized which 'moves' these parameters out of the distribution function for any weight w. Once the distribution of weights is learned, efectively an ensemble of many neural networks, the authors combine the outputs for a better prediction.

2.2 Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles [11]

Lakshminarayanan et al., report a simple and a scalable method to assess model uncertainty using NN's by using

- **Scoring rules as training criterion :** The rule assigns a score to the predictive distribution , the higher the score the better. A proper scoring function has the property $S\left(p_{\theta}, (y, x)\right) \leq S(q_{\theta}, (y, x))$, where p_{θ} is the predictive distribution , q_{θ} is the true distribution and θ are the parameters of the NN. The author demonstrate how commonly used NN loss functions like maximum likelihood , softmax cross entropy loss in multi-class classification problems , MSE between predictive probability of a label and one-hot encoding of the correct label (a.k.a Brier score) are valid scoring rules. For regression models, that usually output a single value $\mu(\mathrm{x})$, the authors propose a modification to the final layer to output two values - the predictive mean $\mu(\mathrm{x})$ and variance $\sigma^2(\mathrm{x})$.A single output is treated as a sample of a gaussian distribution. With the predicted mean and variance, the model is trained to minimize the below negative log likelihood as follows.

$$-\log p_{\theta}\left(y_n \mid \mathbf{x}_n\right) = \frac{\log \sigma_{\theta}^2(\mathbf{x})}{2} + \frac{\left(y - \mu_{\theta}(\mathbf{x})\right)^2}{2\sigma_{\theta}^2(\mathbf{x})} + constant \tag{7}$$

- **Adversarial training to smooth predictive distributions :** Given an input x , target y , loss $\ell(\theta, \mathbf{x}, y)$ (e.g $\cdot - \log p_{\theta}(y \mid \mathbf{x})$), the fast gradient method generates a adversarial example,$\mathrm{x}' = \mathrm{x} + \epsilon sign\left(\nabla_{\mathrm{x}} \ell(\theta, \mathrm{x}, y)\right)$ which is 'close' to the original input. If ϵ is small , (x', y) can be treated as additional training data (data augmentation). Intuitively , the model is trained using adversarial training methods to smoothen the predictive distribution.

- **Training Ensemble of NN's:** Ensembles of M networks trained independently on the entire dataset using random initialization are adopted that de-corelates the predictions from individual models. This approach has its advantage where the ensemble members can be trained in parallel with the entire training dataset. The ensemble is treated as a uniformly weighted mixture model and the predictions are combined as $p(y \mid \mathbf{x}) = M^{-1} \sum_{m=1}^{M} p_{\theta_m}\left(y \mid \mathbf{x}, \theta_m\right)$. Due to computational restrictions , the ensemble prediction is approximated as a Gaussian mixture of the form $M^{-1} \sum \mathcal{N}\left(\mu_{\theta_m}(\mathbf{x}), \sigma_{\theta_m}^2(\mathbf{x})\right)$ where the $\mu_*(\mathrm{x}) = M^{-1} \sum_m \mu_{\theta_m}(\mathrm{x})$ and $\sigma_*^2(\mathrm{x}) = M^{-1} \sum_m \left(\sigma_{\theta_m}^2(\mathrm{x}) + \mu_{\theta_m}^2(\mathrm{x})\right) - \mu_*^2(\mathrm{x})$ respectively.

2.3 Ensemble Distribution Distillation (EnD^2) [14]

Malinin et al. further build upon the works of Lakshminarayanan et al. [11] with the goal to not only capture the mean of the ensemble but also the diversity. Intuitively,an ensemble of models can be seen as samples from a distribution.They utilise a Prior Network [13] that parameterizes a conditional distribution over output distributions $\mathrm{p}\left(\pi \mid x^*, \mathcal{D}\right)$ with parameter θ and input x^*.

$$\left\{\mathrm{P}\left(y \mid x^*, \theta^{(m)}\right)\right\}_{m=1}^{M} \rightarrow \left\{\mathrm{P}\left(y \mid \pi^{(m)}\right)\right\}_{m=1}^{M}, \quad \pi^{(m)} \sim \mathrm{p}\left(\pi \mid x^*, \mathcal{D}\right) \tag{8}$$

EnD2 is an applications of ML estimation to Prior networks.The training is done as follows -

1. Transfer Dataset Creation : A dataset $\mathcal{D}_{ens} = \left\{x_i, \pi_i^{(1:M)}\right\}_{i=1}^{N} = \hat{\mathbf{p}}(x, \pi)$ is created ,where x_i is input from training set $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ and categorical distribution $\left\{\pi_i^{(1:M)}\right\}_{i=1}^{N}$ from the ensemble for each input.
2. The model $\mathrm{p}(\pi \mid x; \theta)$ is then trained by minimizing the NLL for each categorical distribution as : $\mathcal{L}\left(\theta, \mathcal{D}_{\text{ens}}\right) = -E_{\hat{\mathrm{p}}(x)}\left[E_{\hat{\mathrm{p}}(\pi \mid x)}[\ln \mathrm{p}(\pi \mid x; \theta)]\right]$
3. Temperature Annealing: [12] is used during optimization of the KL diveregence for faster convergence and to increase the common support of model and target empirical distribution.The empirical distribution is first heated with a temperature T during training, followed by scheduled annealing to return it to its initial state .
4. The predictive distribution can be evaluated as (for Dirichlet Prior)

$$\mathrm{P}\left(y \mid x^*; \hat{\theta}\right) = E_{\mathrm{p}\left(\pi \mid x^*; \hat{\theta}\right)}[\mathrm{P}(y \mid \pi)] = \hat{\pi} \tag{9}$$

3 Comparison

The Bayesian approach to model uncertainty using VI in [1] relies on several approximations. The quality of these predictions severely depends upon

- The degree of approximations applied due to computational and implementations constraints for eg. mean-field approximation.
- The choice of the prior and variational posterior for analytical solutions.

These constraints raise the need for a general purpose solution. Ensemble models as used in [11] and [14] have been empirically shown to not only provide robust measures of uncertainty but are also capable of distinguishing different forms of uncertainty - knowledge uncertainty and data uncertainty as shown in [14] . EnD [11] and EnD^2 [14] perform better in comparision to single NN [1] on C10,C100 and TIM in terms of Mean Classification error [14].EnD^2 outperforms or matches EnD in terms of Prediction Rejection Ratio (PRR) on all datasets with or without auxiallry data [14]. At the same time , a fundamental limitation of ensembles is that the computational cost of training and inference is many times greater than that of a single NN.Especially in [11] , to train an ensemble with M members, each network has to train on the entire training set. The biggest advantage of EnD [11] is the assumption of random initialization , essentially enabling you to train all the ensemble members in parallel. EnD^2 [14] is able to preserve the information about the diversity of the members networks and therefore yield better results than standard Ensemble distillation. This significantly improves the performance on Out of Distribution (OOD) detection.

4 Weaknesses

- **Weight Uncertainty in Neural Networks [1]**
  - Variational posterior with mean field approximation have been shown to provide inaccurate uncertainty predictions especially when dealing with time-series data [20]. The approach cannot be generalized and the performance is greatly influenced by the data it is applied on.
  - The choice of approximate distribution was influenced for an easier analytic solution , usually a conjugate .
- **Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles [11]**
  - The choice of explicitly de-co relating networks for a scalable solution may result in loss of uncertainty.The use of implicit ensembles where the ensembles share parameters should be investigated to draw concrete conclusions/justify the choice.
  - The authors claim the model requires little/no hyper parameter tuning and they were randomly initialized. However,the justification to why these parameters shouldn't effect the results is not provided.

- **Ensemble Distribution Distillation (EnD^2) [14]**
  - Assumption of Dirichlet prior due to tractable analytic properties is restrictive and limits the diversity of the output distribution.
  - A prior network may not give accurate predictions for inputs it has not trained on (as demonstrated by Malinin and Gales [14].

References

[1] Charles Blundell et al. “Weight Uncertainty in Neural Networks”. In: (May 2015).

[2] D. N. Fente and D. Kumar Singh. “Weather Forecasting Using Artificial Neural Network”. In: *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. 2018, pp. 1757–1761.

[3] Yarin Gal. *Uncertainty in Deep Learning. PhD thesis, University of Cambridge.* 2016.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples.* 2014. arXiv: `1412.6572 [stat.ML]`.

[5] Alex Graves. “Practical Variational Inference for Neural Networks”. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems.* NIPS’11. Granada, Spain: Curran Associates Inc., 2011, pp. 2348–2356. ISBN: 9781618395993.

[6] Simon Hecker, Dengxin Dai, and Luc Van Gool. *Failure Prediction for Autonomous Driving.* 2018. arXiv: `1805.01811 [cs.CV]`.

[7] Geoffrey E. Hinton and Drew van Camp. “Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights”. In: *Proceedings of the Sixth Annual Conference on Computational Learning Theory.* COLT ’93. Santa Cruz, California, USA: Association for Computing Machinery, 1993, pp. 5–13. ISBN: 0897916115. DOI: `10.1145/168304.168306`. URL: `https://doi.org/10.1145/168304.168306`.

[8] Heinrich Jiang et al. *To Trust Or Not To Trust A Classifier.* 2018. arXiv: `1805.11783 [stat.ML]`.

[9] Diederik P. Kingma, Tim Salimans, and Max Welling. “Variational Dropout and the Local Reparameterization Trick”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2.* NIPS’15. Montreal, Canada: MIT Press, 2015, pp. 2575–2583.

[10] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes.* 2013. arXiv: `1312.6114 [stat.ML]`.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: (Dec. 2016).

[12] Yann LeCun, Y. Bengio, and Geoffrey Hinton. “Deep Learning”. In: *Nature* 521 (May 2015), pp. 436–44. DOI: `10.1038/nature14539`.

[13] Andrey Malinin and Mark Gales. *Predictive Uncertainty Estimation via Prior Networks.* 2018. arXiv: `1802.10501 [stat.ML]`.

[14] Andrey Malinin, Bruno Mlodozeniec, and M.J.F. Gales. “Ensemble Distribution Distillation”. In: (Apr. 2019).

[15] E. Nishani and B. Çiço. “Computer vision approaches based on deep learning and neural networks: Deep neural networks for video analysis of human pose estimation”. In: *2017 6th Mediterranean Conference on Embedded Computing (MECO).* 2017, pp. 1–4.

[16] Molly O’Brien et al. *Dependable Neural Networks for Safety Critical Tasks.* 2019. arXiv: `1912.09902 [cs.LG]`.

[17] Joaquin Quionero-Candela et al. *Dataset Shift in Machine Learning.* The MIT Press, 2009. ISBN: 0262170051.

[18] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. *Stochastic Backpropagation and Approximate Inference in Deep Generative Models.* 2014. arXiv: `1401.4082 [stat.ML]`.

[19] Qeethara Al-Shayea. “Artificial Neural Networks in Medical Diagnosis”. In: *Int J Comput Sci Issues* 8 (Feb. 2011), pp. 150–154.

[20] Richard Turner et al. “Two problems with variational expectation maximisation for time-series models”. In: *Bayesian Time Series Models* (Jan. 2011), pp. 109–130.

[21] Tom Young et al. *Recent Trends in Deep Learning Based Natural Language Processing.* 2017. arXiv: `1708.02709 [cs.CL]`.

Self assessment for bonus

I attempt to summarize the content of the three research papers thoroughly with a brief yet properly referenced literature review. The strengths and weakness of each model is examined in a original and coherent way. I tried to point out some weakness of each paper by reviewing additional papers not in the list (improvements by others in recent publications).