Deep Ensemble Bayesian Active Learning

Remus Pop
The University of Edinburgh
s0908127@sms.ed.ac.uk

Patric Fulop
The University of Edinburgh
patric.fulop@ed.ac.uk

1 Introduction

There have been tremendous successes in using deep neural networks and in particular CNNs in critical domains such as skin cancer detection (Haenssle et al., 2018), retinal disease detection (De Fauw et al., 2018) or brain tumour survival prediction (Lao et al., 2017). However, there are still a number of technical challenges that need to be addressed. Among the most important ones are accurate uncertainty representation and the need of huge amounts of annotated data that is often costly to acquire (Hoi et al., 2006; Smith et al., 2018).

Active Learning (AL) is a sound framework (Cohn et al., 1996) that aims to reduce the amount of labelled data needed for a specific task by acquiring and labelling new unlabelled data points based on their uncertainty or representativeness, using an *acquisition function*. One of the most promising uncertainty based methods is Deep Bayesian Active Learning (DBAL) (Gal et al., 2017), which uses Monte-Carlo dropout (MC-dropout) as a framework to obtain uncertainty estimates. The approach presented by Beluch et al. (2018) uses ensemble models to obtain better uncertainty estimates compared to DBAL methods. Finally, density-based approaches try to identify the samples that are most representative of the entire unlabelled set, albeit at a computational cost (Sener and Savarese, 2017).

Our hypothesis is that overconfident predictions for DBAL methods are an outcome of the mode collapse phenomenon in variational inference methods (Blei et al., 2017; Srivastava et al., 2017), and that by combining the expressive power of ensemble methods with MC-dropout we can obtain better uncertainties without trading representativeness. In our work we find evidence for the mode collapse phenomenon and link it to over-confident classifications. Finally, we present Deep Ensemble Bayesian Active Learning (DEBAL) which confirms our intuition about using ensemble models to address mode collapse and enhance the MC-Dropout technique.

Deep Bayesian Active Learning. A probabilistic neural network is given by a model $f(\boldsymbol{x};\boldsymbol{\theta})$ with a prior $p(\boldsymbol{\theta})$ (usually Gaussian) over the parameter space $\boldsymbol{\theta}$ and a likelihood $p(y = c|\boldsymbol{x},\boldsymbol{\theta})$. By obtaining the posterior distribution over $\boldsymbol{\theta}$, one can make predictions y^* about new data points $\boldsymbol{x}^*$:

$$p(y^*|\mathbf{x}^*, \mathbb{X}, \mathbb{Y}) = \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathbb{X},\mathbb{Y})}[f(\boldsymbol{x};\boldsymbol{\theta})]$$

Bayesian uncertainty is obtained via MC-dropout, a scalable method that replaces $p(\boldsymbol{\theta}|\mathbb{X},\mathbb{Y})$ with the dropout distribution $\hat{q}(\boldsymbol{\theta})$ and has been shown to be equivalent to approximate variational inference (Gal, 2016). Two examples of AL acquisition functions used in our work are:

- *MaxEntropy* (Shannon, 2001). $H[y|\boldsymbol{x},\boldsymbol{\theta}] = -\sum_c p(y = c|\boldsymbol{x},\boldsymbol{\theta})\mathrm{log}p(y = c|\boldsymbol{x},\boldsymbol{\theta})$
- *Bayesian Active Learning by Disagreement (BALD)* (Houlsby et al., 2011).

$$I(y,\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{\theta}) = H[y|\boldsymbol{x};\mathbb{X},\mathbb{Y}] - \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathbb{X},\mathbb{Y})}\Big[H[\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}]\Big]$$

The first one captures *epistemic* uncertainty only, which is a consequence of insufficient learning of model parameters due to lack of data, whilst the second one also captures the genuine stochasticity in the data (noise), and is known as *aleatoric* uncertainty (Smith and Gal, 2018; Depeweg et al., 2017).

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

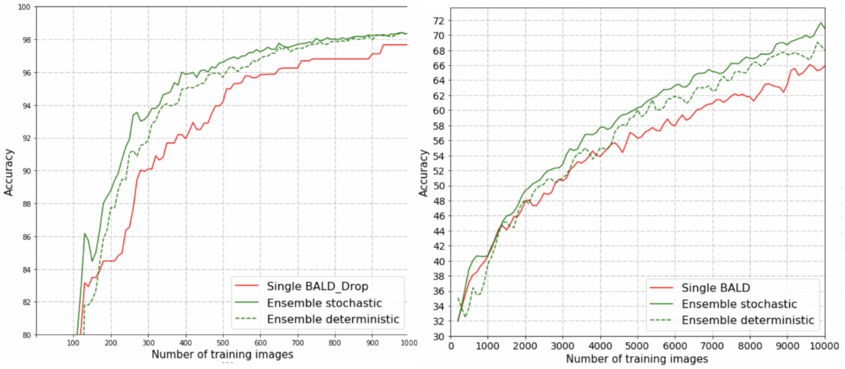

Figure 1: Test accuracy as a function of size of the incremental training set during AL. Effect of using an ensemble of three similar models (stochastic or deterministic) instead of one single MC-Dropout network. **Left:** MNIST. **Right:** CIFAR-10

Using the Bayesian MC-Dropout framework, with class conditional probability $p(y|\boldsymbol{x}, \boldsymbol{\theta})$, we rewrite the stochastic predictive entropy with K MC-Dropout forward passes at test time:

$$H[y|\boldsymbol{x}, \boldsymbol{\theta}] = -\sum_{c} \Big(\frac{1}{K}\sum_{k} p(y = c|\boldsymbol{x}, \boldsymbol{\theta}_k)\Big) \log\Big(\frac{1}{K}\sum_{k} p(y = c|\boldsymbol{x}, \boldsymbol{\theta}_k)\Big)$$

2 DEBAL: Deep Ensemble Bayesian Active Learning

We confirm the performance of Gal (2016) using BALD on MNIST and CIFAR10, but observe a lack of diversity in the data acquired during the AL process. Further experiments debunked the hypothesis that images belonging to some specific classes are more uncertain and difficult to classify, due to resemblance of data from other classes. Additional experiments linked over-sampling to negative effects on classification performance at test time. Our experimental evidence reinforces the argument presented in Smith and Gal (2018), where the authors argue that the MC-Dropout technique suffers from over-confident predictions. This behaviour arises from its inability to capture the full posterior distribution of the data. To address the mode collapse issue, we propose *DEBAL*, a stochastic ensemble of M MC-Dropout models, with $M << K$. Each member of the ensemble is characterized by a different set of weights $\boldsymbol{\theta}_m$ trained independently and uniformly weighted at prediction time. Our predictions are further averaged by a number of MC-Dropout forward passes, giving rise to what we call a *stochastic ensemble*. The predictive entropy for our *stochastic ensemble* becomes:

$$\mathrm{H}[y|\boldsymbol{x}; \mathbb{X}, \mathbb{Y}] = -\sum_{c} \Big(\frac{1}{M}\frac{1}{K}\sum_{m}\sum_{k} p(y = c|\boldsymbol{x}, \boldsymbol{\theta}_{m,k})\Big) \log\Big(\frac{1}{M}\frac{1}{K}\sum_{m}\sum_{k} p(y|\boldsymbol{x}, \boldsymbol{\theta}_{m,k})\Big)$$

, where $\boldsymbol{\theta}_{m,k}$ denotes the model parameters for ensemble model member m in the k^{th} forward pass.

2.1 Results

For both datasets, DEBAL shows significant improvements in classification accuracy. Although we illustrate only BALD in Figure 1, similar results were obtained for other acquisition functions. We hypothesize that the supreme performance of the stochastic ensemble method is a result of higher quality uncertainty estimates obtained during AL.

To validate our claims, we observe how DEBAL and single MC-Dropout behave on both seen and unseen datasets (distributions), using the NotMNIST dataset of letters A-J from different fonts (Bulatov, 2011) as the unseen distribution. For the known distribution (Figure 2, left) both methods produce low uncertainty for the majority of the test samples, as expected. However, for the single MC-Dropout network the distribution is characterized by fatter tails (both extremely confident and extremely uncertain about a significant number of images). The ensemble method, however, results in a clustered distribution of uncertainty. This further illustrates that stochastic ensembles learn a more representative part of the input space.

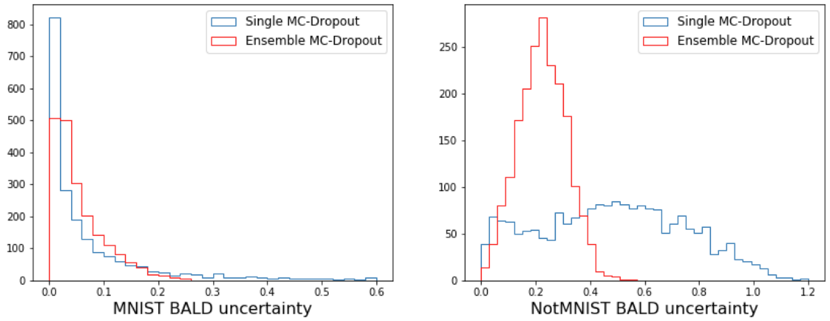

Figure 2: Histogram of BALD uncertainty of **MNIST** (left) and **NotMNIST** (right) images (2,000 random but balanced test set). Uncertainty obtained from single MC-Dropout and ensemble MC-Dropout methods at the end of the AL process.

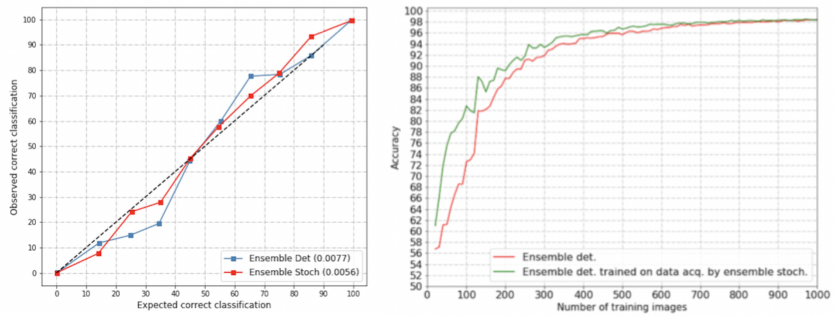

Figure 3: **Left**: MNIST uncertainty calibration. Expected fraction and observed fraction. Ideal output is the dashed black line. MSE reported in paranthesis. Calibration averaged over 3 different runs. **Right**: Deterministic ensemble trained on data acquired by stochastic ensemble.

On the unseen distribution (Figure 2, right) the broad uniform distribution of uncertainty from the single network shows the existence of images that the classifier is both extremely certain and uncertain about. This implies that the features learned previously are only partially transferable to the unseen dataset. For the ensemble, on the other hand, the uncertainty is much smaller and centered on a few values. Features learned during the initial training on MNIST are more general, thus transferable. This behaviour is a desirable and realistic one when evaluating a similar but new dataset.

Deterministic vs stochastic ensemble. Further experiments on both seen and unseen distributions show that the additional improvement in DEBAL over its deterministic counterpart is due to the stochastic method's ability to better capture uncertainty. Moreover, we compare the two methods in terms of uncertainty calibration by looking whether the expected fraction of correct classifications matches the observed proportion (Figure 3, left). Results show that the stochastic ensemble method leads to a better calibrated uncertainty. Finally, training the deterministic ensemble with data acquired by the stochastic one leads to an overall increase in performance, further reinforcing our hypothesis that DEBAL has a better quality of uncertainty (Figure 3, right).

3 Conclusion and Future Work

In this work, we focused on the mode collapse problem for DBAL in the context of image classification. We improved on state of the art by leveraging the expressive power and statistical properties of model ensembles. We linked the performance improvement to a better representation of data uncertainty resulting from our method. We are currently testing our method on a medical imaging dataset, with promising initial results. For future work, DEBAL's superior uncertainty representation could be used to address one of the major issues of deep networks in safety-critical applications: adversarial examples.

Acknowledgments

The work of Patric Fulop was supported by Microsoft Research through its PhD Scolarship Programme and the University of Edinburgh.

References

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bulatov, Y. (2011). Notmnist dataset. *Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html*.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., and Udluft, S. (2017). Uncertainty decomposition in bayesian neural networks with latent variables. *arXiv preprint arXiv:1706.08495*.

Gal, Y. (2016). Uncertainty in deep learning. *University of Cambridge*.

Gal, Y., Islam, R., and Ghahramani, Z. (2017). Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*.

Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A., Thomas, L., Enk, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*.

Hoi, S. C., Jin, R., Zhu, J., and Lyu, M. R. (2006). Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., and Zhai, G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):10353.

Sener, O. and Savarese, S. (2017). Active learning for convolutional neural networks: Acore-set approach. *stat*, 1050:27.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. (2018). Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24):241733.

Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*.

Srivastava, A., Valkoz, L., Russell, C., Gutmann, M. U., and Sutton, C. (2017). Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318.