Discriminative Bayesian neural networks know what they do not know

Christian Leibig, Siegfried Wahl
ZEISS Vision Science Lab
University of Tübingen
72076 Tübingen, Germany
{christian.leibig, siegfried.wahl}@uni-tuebingen.de

1 Introduction

Discriminative deep neural networks (NNs) that are trained in a supervised manner constitute a very successful approach to accurate, categorical classification of high-dimensional data. A major downside however is their limited perceptual horizon, given by the number of classes. Instances of unkown classes are erroneously, but often confidently associated with a known class. The actively studied integration of Bayesian modelling and neural networks [1–11] enables a new perspective: predictions can be associated with uncertainties, where the latter ideally would be high in face of data from classes not trained to recognize.

A particularly useful theoretical connection between dropout networks [12, 13] and approximate but efficient Bayesian inference was identified by Gal and Ghahramani [9, 14, 10]. Using this approach, informative and interpretable uncertainty measures were recently derived from deep Bayesian convolutional networks trained for disease detection from high-dimensional medical images [15]. Analysing the causes of uncertainty, it was found for several settings [15] that it is in particular the difficult predictions that carry a high uncertainty. Images far from the training data could only be detected if they resided in the vicinity of the decision boundary, rendering the detection of abnormal images [16] such as *rubbish class examples* [17] infeasible.

The aim of this work is to show that an almost trivial modification of the standard uncertainty readout (Fig. 1b & 2a) is useful for the detection of data from unknown categories without any label for the latter. Our inspiration comes from the fact that classification networks can be formulated such that they differ from regression networks merely by an additional nonlinearity (a sigmoid or softmax). Bayesian regression networks in turn can exhibit high uncertainty for predictions far from the training data [9, 18]. In the following we provide evidence that the uncertainty about the softmax input is high in regions of the data space that has not been sampled during training.

2 Results

For visualization purposes, we first trained a small Bayesian NN (3 hidden layers with 100 units each) with dropout layers interleaved ($p_{drop} = 0.5$) on a 2D toy classification problem (Fig. 1a). We approximated the predictive posterior distribution over the softmax output (or correspondingly the input) by collecting $T = 100$ Monte Carlo samples with dropout turned on at test time. We quantified uncertainties by the predictive standard deviation σ_{pred} of the approximate predictive posterior over the softmax output (Fig. 1b) and input (Fig. 1c) respectively. Uncertainty about the softmax output is high in the vicinity of the decision boundary, whereas uncertainty about the softmax input is high in regions that are both far from the training data and not captured by the softmax output uncertainty. One might wonder why a large uncertainty about the softmax input (Fig. 1c) does not necessarily result in an overall softmax prediction around 0.5 or at least in a high uncertainty about the softmax output. This is because the approximate predictive posterior over the softmax input may assume values that all reside in the saturating regime of the softmax nonlinearity. We have indeed observed

this (data not shown) in particular for data points with large enough distances from the separating hyperplane. We conclude that far from the training data and the separating hyperplane the prediction uncertainty is low (i.e. the class membership probability as indicated by the softmax output is overly confident), while the model uncertainty (as captured by the softmax input spread) is high.

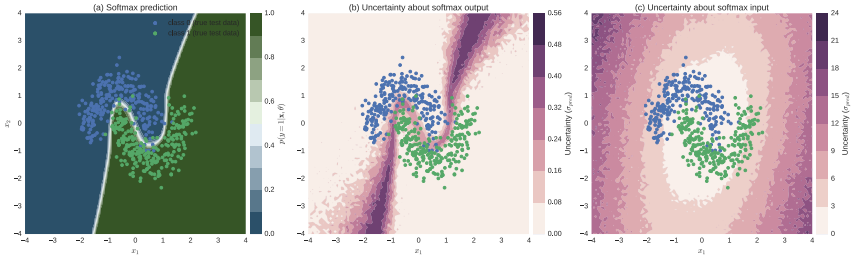

Figure 1: **Illustration of uncertainty for a 2D binary classification problem. (a)** Conventional softmax output obtained by turning off dropout at test time. **(b)** Uncertainty about the softmax output tends to be high for regions in input space that reside in the vicinity of the (extrapolated) decision boundary. **(c)** Uncertainty about the softmax input tends to increase with distance from the training data. Color-coded dots in all subplots correspond to test data the network has not seen during training.

Next we applied the same procedure to a high-dimensional image (512x512 pixels) scenario. For this purpose we took a Bayesian convolutional network (14 weight layers, with 0.2 dropout applied to conv. layers) that had been trained previously for binary disease detection from fundus images [15]. Figure 2 (a) shows the distribution of uncertainty values for 53576 healthy and diseased fundus [19] images (blue) of the softmax output. The space of images with content unknown to the disease detection network was sampled by performing predictions with associated uncertainties (green in Fig. 2) on the 2012 ImageNet [20] validation set (49101 colour images from 1000 different categories). Whereas the uncertainty about the softmax output is not predictive about the presence of an unknown concept (similar distributions in fig. 2a), uncertainty about the softmax input is shifted towards higher values for confounding (ImageNet) vs. known (Kaggle DR) images (Fig. 2b).

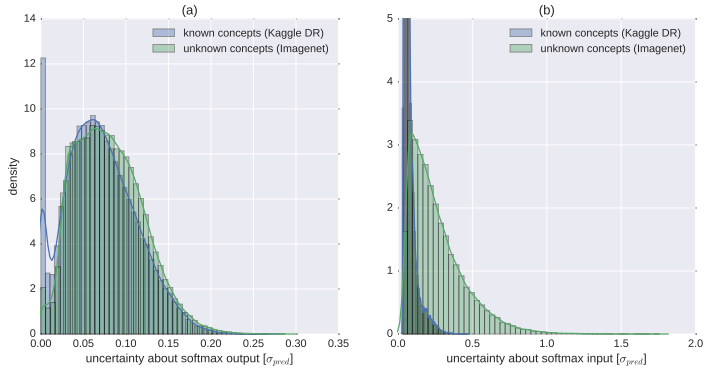

Figure 2: **Anomaly detection using the uncertainty of a discriminative Bayesian convolutional NN.** The network has a limited world view of two categories (*healthy* and *diseased*) because it had been trained for diabetic retinopathy detection from Kaggle DR[19] fundus images. **(a)** Distribution of uncertainties about the softmax output for Kaggle DR[19] test data (53576 fundus images) and Imagenet 2012 validation data [20] (49101 images showing objects from 1000 different categories). **(b)** The same as (a) but for uncertainties about the softmax input.

Conclusion

Extending on previous evidence [15] that the width of the approximate predictive posterior over the softmax output is sensitive to difficult decisions, we have provided evidence that uncertainty about

the softmax input can be useful for detecting images with unknown semantic content. Taken both aspects together, deep discriminative Bayesian neural networks - even though trained on a limited world view - know a lot about what they do not know.

References

[1] N Tishby, E Levin, and S A Solla. Consistent inference of probabilities in layered networks: predictions and generalizations. In *International joint Conference on Neural Networks*, 1989. doi: 10.1109/IJCNN.1989.118274.

[2] David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448.

[3] Geoffrey E. Hinton, Geoffrey E. Hinton, Drew van Camp, and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. *Proceedings of the sixth annual conference on Computational learning theory - COLT '93*, 1993. doi: 10.1145/168304.168306.

[4] Radford M. Neal. Bayesian learning for neural networks. *Lecture notes in statistics*, 1996.

[5] Alex Graves. Practical Variational Inference for Neural Networks. *NIPS*, 2011.

[6] Diederik P Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. *NIPS*, 2015. ISSN 10495258.

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *ICML*, 2015.

[8] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680*, 2015.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv: 1506.02142*, 2015.

[10] Yarin Gal and Zoubin Ghahramani. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *arXiv: 1506.02158*, 2015. doi: 10.1029/2003JD002581.

[11] Christos Louizos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. *ICML*, 2016.

[12] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv: 1207.0580*, 2012.

[13] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958, 2014. ISSN 15337928. doi: 10.1214/12-AOS1000.

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Appendix. *arXiv: 1506.02157*, 2015.

[15] Christian Leibig, Vaneeda Allken, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *bioRxiv*, 2016. doi: http://dx.doi.org/10.1101/084210.

[16] V Chandola, A Banerjee, and V Kumar. Anomaly Detection: A Survey. *ACM computing surveys (CSUR)*, 2009.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *ICLR*, 2015.

[18] Alex Kendall and Roberto Cipolla. Modelling Uncertainty in Deep Learning for Camera Relocalization. *arXiv:1509.05909v2*, 2016.

[19] Kaggle competition on Diabetic Retinopathy Detection, 2015. URL https://www.kaggle.com/c/diabetic-retinopathy-detection.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y.