Uncertainty in Bayesian Convolutional Neural Networks

Ethan Goan, Dimitri Perrin, Kerrie Mengersen, Clinton Fookes

ej.goan@qut.edu.au @ethangoan

Abstract

Convolutional Neural Networks (CNNs) have provided state-of-the-art results for many machine learning problems, though are typically implemented within a frequentist approach. This poster details how variational inference can be applied to perform approximate Bayesian inference for image classification, and how we can visualise uncertainty in the output.

Bayesian CNN

We can define a CNN model for classification as,

$$\begin{aligned}\mathbf{\Phi}_1 &= a(\mathbf{X} \circledast \mathbf{W}_1)\\ \mathbf{\Phi}_i &= a(\mathbf{\Phi}_{i-1} \circledast \mathbf{W}_i)\\ \phi_{N-1} &= \text{vect}(\mathbf{\Phi}_{N-1})\\ \mathbf{f}^{\boldsymbol{\omega}}(\mathbf{X}) = \mathbf{y}^* &= \text{softmax}(\phi_{N-1}^T \mathbf{W}_N)\end{aligned}$$

where,

- $\mathbf{X}$ is our input
- $\mathbf{W}_i$ is an array of parameters at later i
- N is the number of layers
- $a(\cdot)$ is a element-wise non-linear function
- $\mathbf{y}^*$ is the model output
- $\boldsymbol{\omega}$ is the set of all parameters $\{\mathbf{W}\}^N$

We can use this model to define our likelihood and form our posterior over model parameters $\mathbf{W}$,

$$p(\boldsymbol{\omega}|\mathcal{D}) = \frac{\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{X})p(\boldsymbol{\omega})}{p(\mathcal{D})}$$

where $\mathcal{D}$ is the set of inputs $\mathbf{X}$ and true output labels $\mathbf{y}$ in our training data.
Since exact inference is intractable, we assume a simplified form for our posterior, in that of a factorised Gaussian with variational parameters $\boldsymbol{\theta}$,

$$q_\theta(\boldsymbol{\omega}) = \prod_{j=1}^{M} \mathcal{N}(w_j|\mu_j, \sigma_j^2)$$

where M is the number of all scalar parameters in our model and $\boldsymbol{\theta} = \{\mu, \sigma\}^M$.
We want our assumed posterior to be a good approximation of the true posterior. To achieve this, we can apply the KL divergence between the simplified and the true posterior.

$$\text{KL}\Big(q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||p(\boldsymbol{\omega}|\mathcal{D})\Big) = \; -\mathcal{F}[q_\theta] + \log p(\mathcal{D})$$

$\mathcal{F}[q_\theta]$ represents the Evidence Lower Bound (ELBO), which we can maximise w.r.t. $\boldsymbol{\theta}$ to perform approximate inference. This is illustrated graphically in Figure 1, and with derivation of the ELBO in the appendix.

References

[1] D. Barber and C. M. Bishop, "Ensemble learning in bayesian neural networks," *Nato ASI Series F Computer and Systems Sciences*, vol. 168, pp. 215–238, 1998.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[3] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *ICML*, 2015.

Evidence Lower Bound (ELBO)

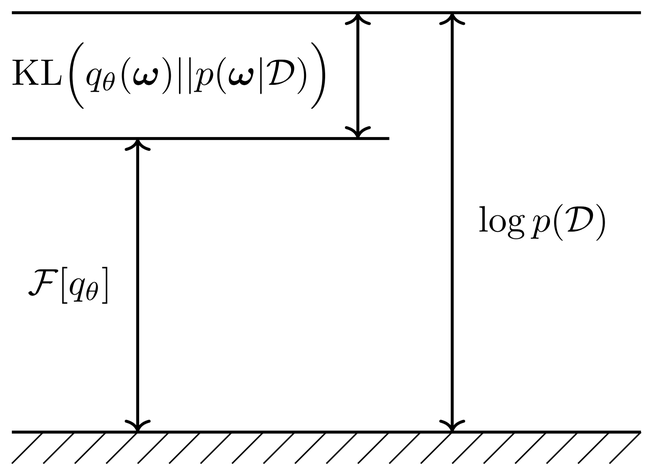

Figure 1: Graphical illustration of how the minimisation of the KL divergence between the approximate and true posterior maximises the lower bound on the evidence. As the KL Divergence between our approximate and true posterior is minimised, the ELBO $\mathcal{F}[q_\theta]$ tightens to the log-evidence. Therefore maximising the ELBO is equivalent to minimising the KL divergence between the approximate and true posterior. Replicated with permission from [1].

Experimental Results

Figure 2 illustrates the output of some difficult to classify images from the MNIST data set using the LeNet-5 network architecture [2]. A Bayesian approach allows us to reason about uncertainty in our predictions, whilst also maintaining a predictive accuracy of 98.7% on 10k test images.

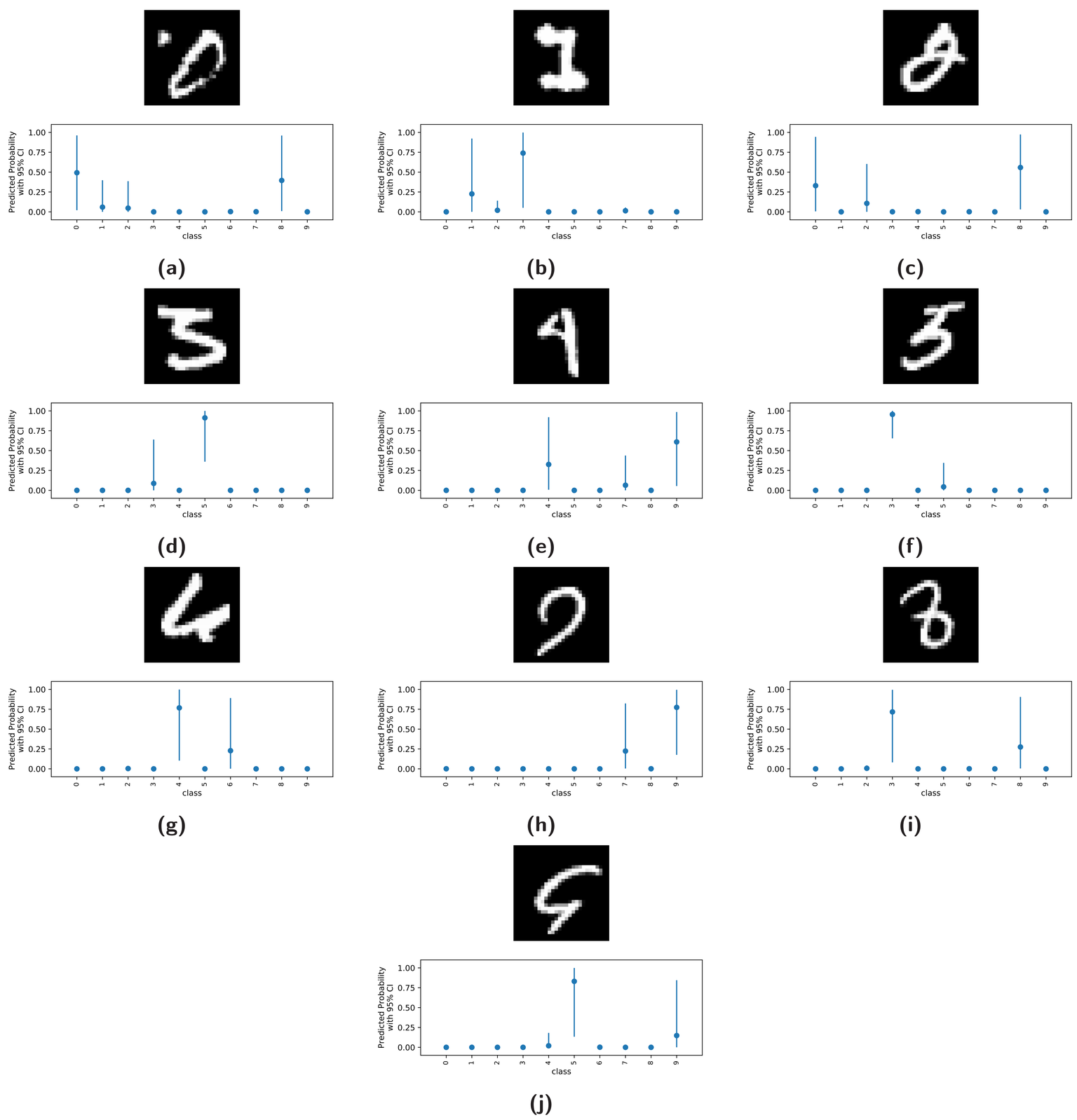

Figure 2: Examples of difficult to classify images from each class in MNIST along with the 95% credible interval. True class for each image is 0-9 arranged in alphabetical order. Trained using the Bayes by Backprop algorithm [3].

Appendix: Derivation of the ELBO

$$\begin{aligned}\text{KL}\Big(q_{\boldsymbol{\theta}}(\boldsymbol{\omega})||p(\boldsymbol{\omega}|\mathcal{D})\Big) &= \mathbb{E}_q\big[\log \frac{q_{\boldsymbol{\theta}}(\boldsymbol{\omega})}{p(\boldsymbol{\omega})} - \log p(\mathcal{D}|\boldsymbol{\omega})\big] + \log p(\mathcal{D})\\ &= \; -\mathcal{F}[q_\theta] + \log p(\mathcal{D})\end{aligned}$$