Modelling Human Visual Uncertainty using Bayesian Deep Neural Networks

Patrick McClure (patrick.mcclure@mrc-cbu.cam.ac.uk)
Tim C Kietzmann (tim.kietzmann@mrc-cbu.cam.ac.uk)
Johannes Mehrer(johannes.mehrer@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, 15 Chaucer Road
Cambridge, CB2 7EF, UK

Nikolaus Kriegeskorte (nk2765@columbia.edu)
Columbia University, 116th St & Broadway
New York, NY 10027

Abstract

Dealing with sensory uncertainty is necessary for humans to operate in the world. Often, multiple interpretations of an event are possible given the sensory evidence, even if one interpretation is most likely. The exact neurobiological mechanism used for representing uncertainty is unknown, but there is increasing evidence that the human brain could use stochasticity to code for uncertainty. However, the convolutional neural networks (CNNs) currently used to model human vision implement deterministic mappings from input to output. We seek to use stochasticity to improve CNNs as both computer vision models and models of human visual perception. We used Gaussian unit noise and sampling to approximate Bayesian CNNs for Ecoset, a large-scale object recognition dataset. We found that sampling during both training and testing improved a CNN's accuracy and ability to represent its own uncertainty for large-scale object recognition. We also found that sampling during both training and testing improved the ability of linear classifiers trained on internal CNN representations to predict human confidence scores for image classification. These results add to the evidence that Bayesian models predict key aspects of human object categorisation behaviour and that sampling in biological neural networks could be a means of representing uncertainty for visual perception in the human brain.

Keywords: Bayesian neural network, human vision, uncertainty

Introduction

Humans must deal with sensory uncertainty to operate in the world (Vilares & Kording, 2011). Often, multiple interpretations are possible given some sensory input, even if one interpretation is most likely. This requires neural representations to code a distribution of interpretations. It has been hypothesised that humans and animals perform near optimal inference by integrating this probabilistically represented information using Bayesian decision theory (Knill & Pouget, 2004; Griffiths, Kemp, & Tenenbaum, 2008). For vision, it has been shown that humans model their own objective uncertainty (Barthelmé & Mamassian, 2009). Several probabilistic neural coding frameworks have been suggested, such as probabilistic population codes (PPC) (Ma, Beck, Latham, & Pouget, 2006) and neural sampling (Fiser, Berkes, Orbán, & Lengyel, 2010). For visual perception in particular, there is evidence for a sampling-based probabilistic representation (Berkes, Orbán, Lengyel, & Fiser, 2011; Moreno-Bote, Knill, & Pouget, 2011; Orbán, Berkes, Fiser, & Lengyel, 2016).

Despite this, current neural network models of high level vision are deterministic and do not model the uncertainty of their learned representations. Specifically, deterministic deep convolutional neural networks (CNNs) have become prominent models in computational neuroscience for visual perception (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). These CNNs either use the maximum likelihood estimate (MLE) or the maximum a posteriori (MAP) solution for the parameters and do not model a distribution of parameters or representations. Utilizing stochasticity can improve a CNN's accuracy and its ability to represent its own uncertainty (McClure & Kriegeskorte, 2017). This is important in building computer vision systems, but also in building better computational models of the human brain.

In this paper, we evaluate sampling in stochastic deep neural networks (DNNs). We approximate a variational Bayesian CNN using MC Gaussian dropout (McClure & Kriegeskorte, 2017). We investigate how much using sampling affects a CNN's classification accuracy, predicted uncertainty, and the ability to predict human confidence scores for natural image classification.

Table 1: The convolutional neural network (CNN) architecture used for CIFAR-10.

Layer	Kernel Size	# Features	Stride	Non-linearity
Conv-1	3x3	64	1	ReLU
MaxPool-1	2x2	64	2	Max
Conv-2	3x3	128	1	ReLU
MaxPool-2	2x2	128	2	Max
Conv-3	3x3	256	1	ReLU
MaxPool-3	2x2	256	2	Max
Conv-4	3x3	512	1	ReLU
MaxPool-4	2x2	512	2	Max
Conv-5	3x3	512	1	ReLU
MaxPool-5	2x2	512	2	Max
Conv-6	3x3	1024	1	ReLU
MaxPool-6	2x2	1024	2	Max
Conv-7	3x3	1024	1	ReLU
AveragePool-1	3x3	1024	0	Max
FC	1024	$n_{classes}$	-	Softmax

Methods

Approximating Bayesian neural networks using Monte Carlo Gaussian dropout

In machine learning, noise has been traditionally injected into neural networks as a form of regularisation during training followed by using the layerwise expectation during testing, as done using Bernoulli dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) in AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) and VGG-16 (Simonyan & Zisserman, 2014). However, sampling both during training and testing in Bayesian CNNs can lead to better representation of uncertainty. Monte Carlo (MC) sampling during training and testing using multiplicative Gaussian unit noise with a mean of 1 and a variance of $\alpha = (1-p)/p$, where p is the dropout hyperparameter, approximates Bayesian inference in neural networks (McClure & Kriegeskorte, 2017). For the matrix V_m of layer m weight means $v_{i,j}$, unit noise $\varepsilon_i \sim \mathcal{N}(0,1)$, and non-linearity h, each unit, $h(z_{m,i})$, is defined using:

$$z_{m,i} = \sum_{j=1}^{n_{m-1}} h(z_{m-1,j})v_{i,j} + \varepsilon_i \sqrt{\alpha} \sum_{j=1}^{n_{m-1}} h(z_{m-1,j})v_{i,j} \qquad (1)$$

Architecture and datasets

Large-scale object recognition We tested three CNNs: (1) a baseline CNN with no sampling, (2) a CNN with Gaussian unit noise before each ReLU non-linearity only during learning, and (3) a CNN with Gaussian unit noise during training and testing. For all CNNs with sampling during testing, 10 MC samples were used. Each CNN had 8 layers, 7 convolutional and a softmax readout layer, which transforms an activation pattern into a probability distribution (Table 1). CNNs with this architecture were trained on Ecoset (Mehrer, Kietzmann, & Kriegeskorte, 2017) using stochastic gradient descent with momentum and weight normalisation (Salimans & Kingma, 2016). ImageNet (Russakovsky et al., 2015), the most widely used image set used to train CNNs, is a 1,000 class object recognition problem with 1.2 million training images and 150,000 validation images. However, the ImageNet categories are biased towards certain entry-level categories, such as birds and dogs. As a computer vision task, this is reasonable, but from a human visual neuroscience perspective models should be trained on the image distributions that more closely resemble the humans experience. The Ecoset project seeks to create a dataset that more closely matches the human visual diet. This image set is a 565 class object recognition with 569,413 training images, 28,900 validation images, and 28,900 testing images. We evaluated how much using MC sampling during testing, which approximates the expected prediction for an input, affected Ecoset trained CNNs. For all of the CNNs with MC sampling, $p = 0.2$ was found to be the best Gaussian dropout parameter value using validation testing.

Table 2: The accuracies for the Ecoset trained CNNs on the Ecoset test set. For MC sampling, the mean and standard deviation of the accuracies across 5 MC runs, each computed with 10 MC samples.

CNN	Ecoset Accuracy (%)
MAP	49.09
MC Training	53.36
MC Training and Testing	55.06 ± 0.09

Modelling human confidence using decision boundaries The translation from internal representations in the human brain to decisions about object categories has been modelled using linear decision boundaries (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; Ritchie & Carlson, 2016). These models successfully predict human reaction times, a proxy for human confidence, for object recognition. A similar approach can be used to predict human confidence scores from the internal representations of DNNs using the decision scores of linear classifiers (Eberhardt, Cader, & Serre, 2016). Eberhardt et al. (2016) created five non-overlapping sets of 300 grey-scaled randomly sampled ImageNet images, each set containing 150 animal and 150 non-animal images. For each image, 50 participants were asked to classify it as animal or non-animal during a fixation task. Eberhardt et al. (2016) computed a human confidence score for each image by considering the fraction of correct animal/non-animal classifications across the 50 participants shown that image. In order to evaluate the ability of sampling to improve prediction of human uncertainties, we trained logistic regression models to classify animal and non-animal images using the internal representations of the Ecoset trained CNNs as input. These logistic regression models were trained using leave-one-out crossvalidation across the five non-overlapping image sets created by Eberhardt et al. (2016). For the MAP CNN, the internal representation for layer m, z_m, for an image was deterministic, leading to the logistic regression optimising:

$$\max_{W} p(y|z_m) \qquad (2)$$

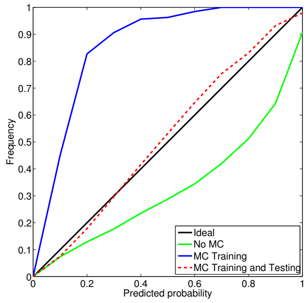
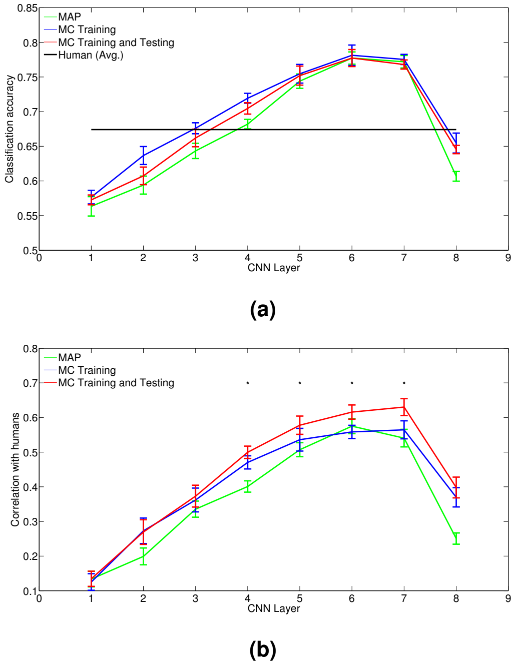


Figure 1: Sampling during training and testing improves CNN calibration for large-scale object recognition. The CNN calibration curves for the Ecoset trained CNN on the Ecoset test set.



(a)

(b)

Figure 2: Sampling during training and testing improves the correlation between CNN-based predicted probabilities and human confidence scores. The mean and standard errors across the five image sets for the (a) accuracies and (b) correlations with human confidence scores for the logistic regression models trained at each CNN. $*$ denotes the "MC Training and Testing" models have significantly higher (p-value$<$0.05, Bonferonni corr.) correlations with human confidence scores than both the "MAP" and "MC Training" models per a non-parametric bootstrap sampling test.

For the stochastic CNN, the internal representation was stochastic, leading to logistic regression optimising:

$$\max_{W} \int p(y|x, z_m) p(z_m|x) dz \tag{3}$$

This integration is approximated using MC sampling with n samples:

$$\max_{W} \frac{1}{n} \sum_{k=1}^{n} p(y|x, z_m^k) \text{ where } z_m^k \sim p(z_m|x) \tag{4}$$

For training, one MC sample was used, as in MC Gaussian dropout. For testing, ten MC samples were used to approximate the predicted probability of class y for input x per:

$$p(y|x) \approx \frac{1}{n} \sum_{k=1}^{n} p(y|x, z_m^k) \text{ where } z_m^k \sim p(z_m|x) \tag{5}$$

Results

Sampling improves accuracy for large-scale object recognition

Using random noise in deep neural networks during learning is often used to reduce overfitting and increase generalisation performance (Srivastava et al., 2014). However, sampling during testing can sometimes lead to accuracy improvements (McClure & Kriegeskorte, 2017). We found that for large-scale object recognition CNN that we trained, MC sampling at test time led to significant accuracy improvements (Table 2) for Ecoset.

Sampling improves the representation of uncertainty for large-scale object recognition

Humans can accurately estimate their objective uncertainty for visual perception (Barthelmé & Mamassian, 2009). AA good computational model of human vision would therefore also need to properly represent its own uncertainty. A model correctly models its own uncertainty (i.e. is well calibrated) if its predicted probabilities closely match the frequency of correctly predicting the true label. To evaluate a network's ability to model its own uncertainty, we calculated the calibration of each of these methods' probabilistic predictions to evaluate the quality of the learned representations. We evaluated how calibrated a prediction was by: (1) Binning test set predictions by predicted probability and then (2) calculating the frequency that predictions in each predicted-probability bin correctly predicted a target label. The larger the difference between these values and the $x = y$ line, the worse the calibration of the model. Sampling during training and testing led to improved calibration of output predictions (Figure 1).

Sampling improves the prediction of human confidence for image classification

We evaluated the how well a logistic regression model predicted human uncertainty by calculating the Pearson correlation coefficient between the human confidence scores for animal/non-animal classification from ImageNet images (Eberhardt et al., 2016) and the output probabilities of a logistic regression model. We hypothesised that MC sampling during training and testing would lead to a significantly higher correlation between model predicted probabilities and human confidence scores. We tested this hypothesis by testing whether the difference between the the correlations was positive for both comparisons using 100,000 bootstrap sampled predictions from the predicted probabilities across layers and input images. MC sampling during both training and testing led to significantly improved correlations to human confidence scores compared to both the MAP and MC sampling only during training (p=1e-5 and p=1e-5, respectively). As found by Eberhardt et al. (2016), the accuracy and correlation with human confidence scores of the linear classifiers generally increases the deeper the layer, except that they decrease at the softmax layer (Figure 2.a). This might have been caused by the increased Ecoset specialisation of the 8th (output) layer features and decreased dimensionality of the 8th layer features. Using a non-parametric bootstrap test for each layer showed that MC sampling during training and testing improved prediction of human confidence scores at deep hidden layers (Figure 2.b). This may be caused by the fact that sampling only during training relies on the assumption of a linear network when approximating the expectation of the output.

This assumption is increasingly violated as we move to deeper layers of the network, which may explain the widening of the gap between the human confidence prediction accuracies of the two models.

Discussion

Biological neural networks are highly stochastic. This variance may code for uncertainty in neural representations. In this work, we evaluated the effect of adding stochasticity (in the form of Gaussian unit noise) and sampling-based testing on large scale CNNs. We tested the effects of sampling during learning and during training and testing on Ecoset trained CNNs. Sampling during training and testing not only improved the accuracy and the representation of uncertainty of CNNs for the Ecoset test set, making the CNNs better computer vision models, but also increased the correlation between the predictions of classifiers trained on the internal representations of CNNs and human confidence scores, making them better models of human visual perception. These improvements were caused by simply injecting random Gaussian noise into the CNNs and integrating over different random noise samples. This mechanism is not only simple, but also neurobiologically plausible. These results add to the evidence that Bayesian models predict human behaviour and that sampling in biological neural networks could be a means of representing uncertainty for visual perception in the human brain.

References

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, *5*(9), e1000504.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*(6013), 83–87.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of cognitive neuroscience*, *26*(1), 132–142.

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In *Advances in neural information processing systems* (pp. 1100–1108).

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). *Statistically optimal perception and learning: from behavior to neural representations* (Vol. 14) (No. 3).

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). *Bayesian models of cognition*. Cambridge University Press.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

McClure, P., & Kriegeskorte, N. (2017). Robustly representing uncertainty in deep neural networks through sampling. *NIPS Bayesian Deep Learning Workshop*.

Mehrer, J., Kietzmann, T. C., & Kriegeskorte, N. (2017). Deep neural networks trained on ecologically relevant categories better explain human it. In *Conference on cognitive computational neuroscience.*

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496.

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, *92*(2), 530–543.

Ritchie, J. B., & Carlson, T. A. (2016). Neural decoding and inner psychophysics: a distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in Neuroscience*, *10*, 190.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of Computer Vision*, *115*(3), 211–252.

Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems* (pp. 901–901).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22–39.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.