DEEP MODELS CALIBRATION WITH BAYESIAN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We apply Bayesian Neural Networks to improve calibration of state-of-the-art deep neural networks. We show that, even with the most basic amortized approximate posterior distribution, and fast fully connected neural network for the likelihood, the Bayesian framework clearly outperforms other simple maximum likelihood based solutions that have recently shown very good performance, as temperature scaling (Guo et al., 2017). As an example, we reduce the Expected Calibration Error (ECE) from 0.52 to 0.24 on CIFAR-10 and from 4.28 to 2.456 on CIFAR-100 on two Wide ResNet with 96.13% and 80.39% accuracy respectively, which are among the best results published for this task. We demonstrate our robustness and performance with experiments on a wide set of state-of-the-art computer vision models. Moreover, our approach acts off-line, and thus can be applied to any probabilistic model regardless of the limitations that the model may present during training. This make it suitable to calibrate systems that make use of pre-trained deep neural networks that are expensive to train for a specific task, or to directly train a calibrated deep convolutional model with Monte Carlo Dropout (Gal & Ghahramani, 2015; Kingma et al., 2015) approximations, among others. However, our method is still complementary with any Bayesian Neural Network for further improvement.

1 INTRODUCTION

Deep Neural Networks (DNNs) have achieved state of art performance in many task such as Image Recognition (Huang et al., 2017; Szegedy et al., 2017; Zagoruyko & Komodakis, 2016), language modeling (Mikolov et al., 2013a;b), machine translation (Vaswani et al., 2017) or speech (Hinton et al., 2012). For that reason, neural networks are now used in many applications. However, there are many task in which the probabilistic information given by the model is necessary to have a good performance, but only if those probabilities are *reliable*. For example, a probabilistic classifier can be incorporated into a more complex model considering multiple sources of information, by the use of e.g., probabilistic graphical models (Koller & Friedman, 2009), or by combining neural networks with languages models in natural language processing tasks (Gulcehre et al., 2017). In addition, probabilistic outputs of classifiers have proven to be useful in many areas apart from classical machine learning tasks, such as language recognition (Brümmer & van Leeuwen, 2006), speech recognition (Tüske et al., 2018) or medical diagnosis (Caruana et al., 2015). In all those areas, reliable probabilities address tasks where it is very different to decide towards a class with a very high, extreme probability; than with a very moderate probability, much closer to the threshold given by the decision rule at hand.

However, all this state of art performance is measured in terms of how discriminative a model can be, and this information is directly given by the accuracy. This means that, as long as we get a sample correctly classified, we do not care if the confidence of our prediction is well calibrated, which in a sens measures how well does our model manage uncertainty.

In Bayesian statistics, the reliability of probabilities is measured by their calibration. As a consequence, the machine learning community has been exploring methods to calibrate the output of classifiers, and the many beneficial properties of well-calibrated probabilities (Zadrozny & Elkan, 2002a; Cohen & Goldszmidt, 2004; Niculescu-Mizil & Caruana, 2005). Moreover, in modern deep

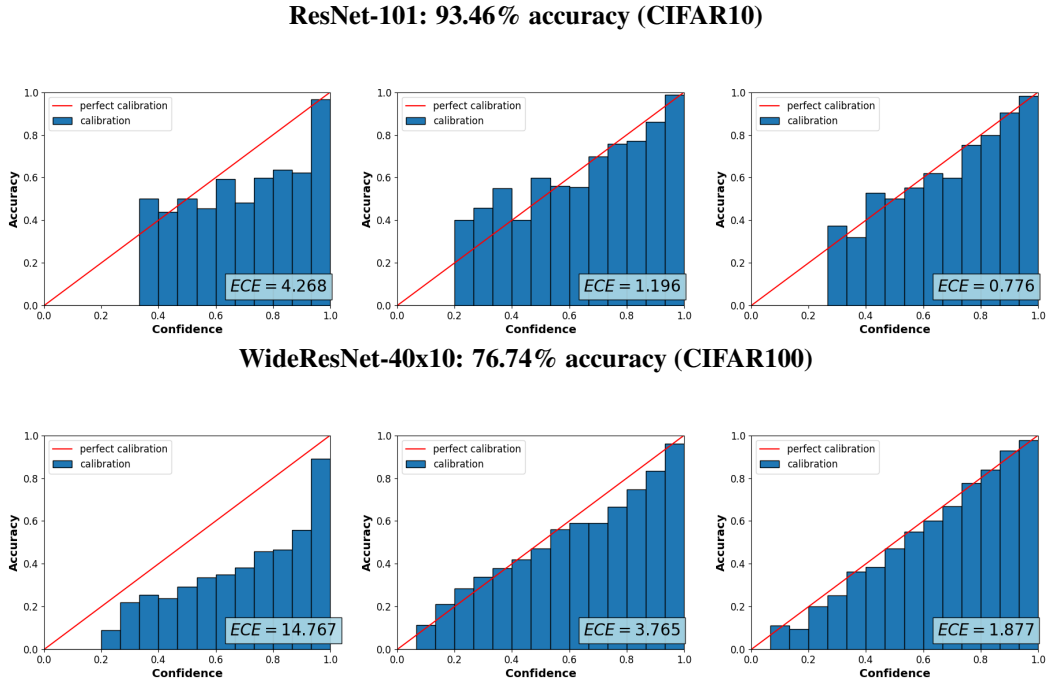

Figure 1: Reliability diagrams (Guo et al., 2017) for two networks trained on CIFAR-10 and CIFAR-100. From left to right we have the model uncalibrated, calibrated with temp-scal and calibrated using a Bayesian Neural Network. The red line represents perfect calibration. We also plot the Expected Calibration Error for 15 bins (see section 4 for a description). The lower the better.

learning architectures, improvements must be applied in an efficient way avoiding, for instance, expensive Markov Chain Monte Carlo algorithms or deep convolutional Monte Carlo Dropout regularized neural networks, among others.

In the past, neural networks trained with a cross-entropy criterion tended to present relatively good calibration. However, a relevant recent work in Guo et al. (2017) has evidenced that modern state-of-the-art neural networks are badly calibrated in general. Moreover, the same work shows that calibration can be dramatically improved by very simple maximum-likelihood parametric techniques, among which Temperature Scaling (temp-scal from now on) is highlighted as the preferred choice, due to its extreme simplicity, very good behavior in general and computation efficiency. In fact, temp-scal outperforms more complex techniques in most cases, leading to the conclusion that good calibration can be better achieved with simpler techniques. This conclusion follows the hypothesis that the space configured by the outputs of a deep model is relatively simple, and therefore good performance, measured as expected calibration error (ECE), can be achieved with very simple models. In fact, temp-scal is a technique that performs nicely in complex, multiclass tasks (Guo et al., 2017).

However, in this work we demonstrate an alternative hypothesis: the calibration can be further improved by *expressive models*, as long as the *uncertainty* in the problem is properly addressed. According to this, we show experiments where a Bayesian Neural Network (BNN from now on) outperforms the calibration of temp-scal in a wide range of tasks. The key point of our proposal is using Bayesian inference, which effectively includes the uncertainty of the problem (Bishop, 2006), rather than maximum-likelihood solutions, that can be sensitive to sources of non-considered uncertainty. In this sense, we consider that calibration is a task that is not achieved optimally with simpler techniques such as temp-scal. Thus, more expressive models are possible to cope with the complexity of the calibration task, but being more robust and insensitive to the uncertainty inherently present in any classification task. As example, figure 1 shows the reliability diagram and ECE comparing calibration of a raw state-of-the-art neural network, calibrated with temp-scal and calibrated with our method, ie, BNNs.

Up to our knowledge, nobody has applied BNNs to improve the calibration of classifiers in and off-line setting, which is the second main contribution of this work. We understand *off-line* as set-up where previously trained deep models can be calibrated with a further stage, without the need of re-training them. Contrary to other Bayesian methods such as Monte Carlo Dropout (Gal & Ghahramani, 2015; Kingma et al., 2015), Bayes by back propagation (Blundell et al., 2015) or Multiplicative Normalizing Flows (Louizos & Welling, 2017), our method is able to cope with any probabilistic model no matter its computation limitations, as applying this Bayesian techniques directly to a deep convolutional model is highly expensive, because it requires several convolutional forward-backwards during optimization and prediction stages. This means that is suitable both for tasks in which we use a pre-trained model; or when we directly train a deep neural network to output calibrated distributions. We will show that we can use fast fully connected neural networks, and thus inference does not implies a decrease in computational performance mainly because forward through a DNN is much more expensive compared to fully connected and because Monte Carlo integration can be fully parallelized. Thus, using fast BNNs does not suppose a significant decrease in performance w.r.t temp-scal, as we will later discuss.

Finally, although recent advances in generative modeling (Kingma & Welling, 2014; Rezende et al., 2014) allow efficient learning of BNNs, our work reflects that basic variational approximations have several computation limitations directly related to the expressiveness needed by the task. We expect that by applying more sophisticated BNN approximations we can improve calibration and reduce the complexity of the topologies. This mean we expect to improve calibration and efficiency performance by reducing the gap between the variational and the true posterior distributions. Our results open new perspectives to improve the variatonal approximation by using local reparameretization (Kingma et al., 2015), normalizing flows (Rezende & Mohamed, 2015; Kingma et al., 2016), auxiliary variables (Agakov & Barber, 2004; Maaløe et al., 2016; Ranganath et al., 2016) and combinations of them (Louizos & Welling, 2017). Moreover, we believe that our results might foster further research in e.g. regularization (Blundell et al., 2015; Gal & Ghahramani, 2015), uncertainty treatment (Gal & Ghahramani, 2016; Louizos & Welling, 2017), dropout strategies(Gal & Ghahramani, 2015; Kingma et al., 2015), robustness to adversarial examples (Louizos & Welling, 2017) and improvements of the approximate posterior (Kingma et al., 2016; Ranganath et al., 2016; Louizos & Welling, 2017; Shu et al., 2018; Kim et al., 2018).

The paper is organized as follows. We first provide a brief description on how we can overcome the limitations presented in the Bayesian framework when using Neural Networks to model the likelihood distribution with a description of our off-line approach and model set-up. We support our work with experiments that clearly outperform temp-scal, evidencing the robustness of our proposal in a wide set of situations. Finally we discuss our results and approach, showing pros and cons and how can we manage to solve them.

2 RELATED WORK

To our knowledge, temp-scal (Guo et al., 2017) is the best technique to improve calibration over a list of classical ways of improving calibration, such as histogram binning (Zadrozny & Elkan, 2001), isotonic regression (Zadrozny & Elkan, 2002b), Platt scaling (Platt, 1999) , Bayesian binning into quantiles (Naeini et al., 2015). For a recent description and performance comparison with modern neural networks see Guo et al. (2017).

On the other hand, there are several works that study overconfident predictions and model uncertainty in different contexts but not present a notion of calibration performance. Gal & Ghahramani (2015) connect Bernoulli dropout with BNNs and Gal & Ghahramani (2016) Gaussian processes with classical dropout regularized networks showing how can we get uncertainty estimates from this networks. Kingma et al. (2015) formalize Gaussian dropout as a Bayesian approach. In Louizos & Welling (2017), novel BNNs are proposed, mixing inverse autoregressive flows (Kingma et al., 2016) and auxiliary variables (Ranganath et al., 2016; Maaløe et al., 2016). In Lakshminarayanan et al. (2017), network ensembles are used with adversarial training for a confident score and a notion of calibration is presented, however a calibration measurementents is not compared against any recent work as we do. Both Louizos & Welling (2017) and Lakshminarayanan et al. (2017) evaluate uncertainty by training on one dataset and use it on another, expecting a maximum entropy output distribution.

In Pereyra et al. (2017), an entropy term is added to the log-likelihood to relax overconfidence. Chen et al. (2018) propose a model that uses probes of the individual layers of the neural network classifier to create a confidence score for the neural network output. DeVries & Taylor (2018) trains a second output obtained from the penultimate layer of the classifier to be confident by interpolation of the softmax output and the true value, scaled by this score. Lee et al. (2018) proposed a generative approach for detecting out-of-distribution samples but evaluates calibration performance comparing their method with normal cross entropy minimization, using temp-scal as the calibration technique.

Finally, (Zhang et al., 2018) propose an inference algorithm scalable and asymptotically accurate as MCMC algorithms.

3 NEURAL NETWORKS AND BAYESIAN INFERENCE

Bayesian statistics make use of the predictive distribution to infer a random variable by computing the expected value of all the possible likelihood distributions. This is done under the posterior distribution of the likelihood parameters. When using neural networks to model the likelihood, several intractabilities arise and must be solved with approximations.

In this case we use monte carlo integration to solve the integral and approximated amortized inference to approximate the posterior distribution, as Markov Chain Monte Carlo algorithms are very expensive to get samples from the posterior in big dataset tasks. Formally, given a labelled dataset $\mathcal{X} = \{(x_i, t_i)\}_{i=1}^{N}$ we train our variational distribution by maximization of the Evidence Lower Bound:

$$ELBO = \mathop{\mathbb{E}}_{\theta \sim q_\phi(\theta)} \{\log p(t, \theta|x)\} + H[q_\phi(\theta)] \tag{1}$$

where $H[q_\phi(\theta)]$ denotes the entropy of the variational distribution $q_\phi(\theta)$. Maximizing the ELBO w.r.t the variational parameters ϕ ensures that the variational distribution approximates the true posterior. Variational Inference and has been studied for the last years (Jordan et al., 1999; Graves, 2011; Challis & Barber, 2012; Mnih & Gregor, 2014; Rezende et al., 2014; Kingma & Welling, 2014). An alternative view of this scenario is given by a MDL principle (Minimum Description Length) (Hinton & van Camp, 1993) yielding the same training criteria.

We use 2-hidden layer fully connected networks to parametrize the conditional likelihood, with ReLU activations. Our variational distribution takes the form of a factorized Gaussian distribution and we use isotropic Gaussian as the prior over the parameter. We train an unbiased estimator of equation 1 because expectations under the variational distribution cannot be solved. However, with these parametric choices, we can optimize a modified version of equation 1 as we can analytically solve the entropy expectation (check appendix B from Kingma & Welling (2014)), and thus reduce the variance of the ELBO estimator. We approximate the likelihood expectation with Monte Carlo integration and data expectation with batch optimization yielding our final training criteria:

$$ELBO = \frac{1}{N} \sum_{(x,t) \sim p_d(x,t)} \{ \frac{1}{K} \sum_{\theta \sim q_\phi(\theta)} [\log p(t|x, \theta)] - \beta \cdot D_{KL}\{q_\phi(\theta)//p(\theta)\}\} \tag{2}$$

where N is the batch size and K the number of samples of the Monte Carlo integration. We introduce a hyperparameter β to control the importance given to the D_{KL} term, following Blundell et al. (2015). We train using a so-called reparameretization trick (Rezende et al., 2014; Kingma & Welling, 2014) to compute lower-variance unbiased estimators of the gradient of the $ELBO$ under the variational parameters in one backward step per value of K. After maximization of the $ELBO$, we make predictions by solving:

$$p(t|x, \mathcal{X}) \approx \frac{1}{M} \sum_{i=1}^{M} p(t|x, \theta_i)\,;\ \theta_i \sim q_\phi(\theta) \tag{3}$$

this means that we get M samples of the network parameters, perform M forward steps through the network, and averaging the softmax activations. We choose the value M using a validation set.

As we commented, our approach acts offline w.r.t the deep convolutional model. This means that the input to our Bayesian stage is made up from the the projections of a sample (in this case an image) to the logit space of a deep convolutional model. For predictions over a test sample we first forward

through the deep model, and then produce calibrated probability distributions using equation 3. For its simplicity we will refer to this approach as our mean-field approximation, but we *remark* that it is not the classical mean-field variational inference (Jordan et al., 1999).

We compare against temp-scal, as to our knowledge is the state of the art in calibration. Formally, given $\mathcal{X}$, temp-scal maximize the log-likelihood of the conditional distribution $p(t|x/T)$ w.r.t T. In this case x represents also the logits of the deep convolutional model.

4 EXPERIMENTS

For our experiments we have used CIFAR10 and CIFAR100 databases (Krizhevsky et al., a;b); SVHN (Netzer et al., 2011); and a GENDER recognition task (Eidinger et al., 2014). We use a validation set randomly taken from the training set with 5000 samples for CIFAR10 and CIFAR100, 10000 samples for SVHN and 4005 for GENDER. This validation set is used to train the temp-scal parameter and to choose the number of Monte Carlo samples, M in equation 3. We report results for the best model on validation for all the tested configurations. We optimize the ELBO using Adam (Kingma & Ba, 2014), as it performed better than stochastic gradient descent on a quick check[1]. We used $\beta = 0.1$ from the set $\{1, 0.1, 0.01\}$ as it behaves better on validation.

In order to compare our experiments with uncalibrated and temp-scal calibrated probabilities, we used an unbiased estimator of the Expected Calibration Error (ECE) computed as in Guo et al. (2017), with 15 bins. The ECE measures the expected value of the difference between accuracy and confidence:

$$ECE = \sum_{i=1}^{15} \frac{|B_i|}{N} |acc(B_i) - conf(B_i)| \tag{4}$$

where N is the number of total samples; B_i represents the set of samples whose predictions t confidence lie in bin i; $conf(B_i)$ is the average confidence and $acc(B_i)$ is the accuracy of that bin. We also report the accuracy of our models. This is because a classifier can be perfectly calibrated, but useless.

We evaluate our proposed method on several state of art configurations of computer vision neural networks over the mentioned datasets: Wide Residual Networks (Zagoruyko & Komodakis, 2016), Residual Networks (He et al., 2016b), Densely Connected Neural Networks (Huang et al., 2017), Pre-Activation Residual Networks (He et al., 2016a), Dual Path Networks (Chen et al., 2017), VGG (Simonyan & Zisserman, 2014) and ResNext (Xie et al., 2017). The results reported in this work are obtained from some pretrained neural networks. Details on our github.

4.1 RESULTS

We present the different results for the models and databases in table 1. The most important point is that calibration is improved by a wide margin in every model except for two models in CIFAR100 and one model in SVHN. Table 1 shows that in average our proposed calibration method outperforms temp-scal with an insignificant accuracy loss. This means that high expressive models can cope with the calibration task as long as uncertainty is correctly modelled, and thus, we propose an alternative hypothesis to the one given in Guo et al. (2017) where the authors argued that calibration space is simple. We argue that if using high complex models outperform simple ones is because the distribution of the calibration space is also complex.

We realized that more expressive models are needed by more complex tasks, like CIFAR100. As example, ResNet-18 GENDER uses BNNs of two layers with two neurons per layer while WideResNet 40x10 on CIFAR100 uses two layers of 2000. This reflects that, when dimensionality increases, more expressiveness is needed in the BNNs. Another important point observed in SVHN (see Densenet-169, ResNet-50 and WideResNet 16x8) is that temp-scal has degraded calibration by a factor of three in the worst case. In general BNNs do not degrade the calibration.

One drawback of our mean-field approach is that in some cases we obtain slight accuracy degradation. Accuracy degradation is more relevant only for CIFAR100, however our BNN method reduces ECE15

[1]Experiments where performed using PyTorch and for details on the models and code to reproduce experiments check our github https://github.com/2019submission/bnn.2019

Table 1: ECE 15(%) and Accuracy (%) comparing model uncalibrated, calibrated with temp-scal and with BNN

	CIFAR10					
	uncalibrated		Temp Scal		BNN	
	Acc	ECE	Acc	ECE	Acc	ECE
WideResNet 28x10	96.13	1.835	96.13	0.518	96.08	0.243
DenseNet 121	95.49	2.643	95.49	1.011	95.26	0.600
DenseNet 169	95.49	2.664	95.49	0.826	95.29	0.511
Dual Path Network 92	95.18	2.995	95.18	1.072	95.03	0.730
ResNet 101	93.46	4.268	93.46	1.196	93.38	0.776
VGG 19	93.68	4.412	93.68	1.708	93.67	0.843
Preactivation ResNet 18	94.93	3.155	94.93	0.570	94.73	0.455
Preactivation ResNet 164	93.91	4.102	93.91	0.437	93.82	0.331
ResNext 29_8x16	94.79	2.833	94.79	0.741	94.61	0.728
Wide ResNet 40x10	95.01	3.001	95.01	0.921	95.08	0.594
average	94.81	3.191	94.81	0.9	94.70	0.581

	SVHN					
	uncalibrated		Temp Scal		BNN	
	Acc	ECE	Acc	ECE	Acc	ECE
WideResNet 40x10	96.95	1.26	96.95	1.17	96.90	1.15
Densenet-121	96.76	2.021	96.76	1.092	96.69	0.716
Densenet-169	96.70	0.363	96.70	1.016	96.59	0.453
ResNet 50	96.47	0.886	96.47	1.030	96.33	0.857
Preactivation ResNet 164	96.20	2.539	96.20	1.079	96.08	0.921
Wide ResNet 16x8	96.88	0.710	96.88	1.318	96.82	0.739
Preactivation ResNet 18	96.15	1.574	96.15	0.645	96.05	1.096
average	96,587	1,336	96,587	1.05	96,494	0.847

	CIFAR100					
	uncalibrated		Temp Scal		BNN	
	Acc	ECE	Acc	ECE	Acc	ECE
WideResNet 28x10	80.39	4.853	80.39	4.276	77.59	2.456
DenseNet 121	78.8	8.724	78.8	3.476	75.9	2.534
ResNet 101	72	11.413	72	1.533	68.7	1.612
VGG 19	72.7	17.631	72.7	6.00	71.94	6.918
Preactivation ResNet 18	76.6	10.780	76.9	3.152	74.3	1.763
Preactivation ResNet 164	73.28	15.754	73.28	2.046	70.77	1.461
ResNext 29_8x16	77.88	9.678	77.88	2.811	73.97	2.581
DenseNet 169	79.05	8.883	79.05	3.758	75.58	2.393
Wide ResNet 40x10	76.74	14.767	76.74	3.765	76.17	1.876
average	76,36	11,387	76,36	3,424	73,88	2,622

	GENDER					
	uncalibrated		Temp Scal		BNN	
	Acc	ECE	Acc	ECE	Acc	ECE
VGG-19	90.60	8.08	90.60	3.96	90.50	2.70
DenseNet 121	90.035	8.803	90.035	3.077	89.961	1.547
ResNet 18	90.42	8.45	90.42	3.8	90.44	3.082
average	90.352	8.444	90.352	3.612	90.30	2.443

by a factor of two in some cases in this task. Moreover, in some cases we are able to improve both, accuracy and calibration, see WideResNet 40x10 for CIFAR10 or ResNet-18 for GENDER dataset. We cannot conclude that BNNs are calibrating at the cost of loosing accuracy. This motivate us towards further research on this accuracy degradation as we expect to solve it with more sophisticated approximations, and we let this for further research. Possible hypothesis for this degradation are that either the gap between the variational and the true posterior is very large, or the expressiveness of the likelihood model is not enough to deal with the particular logits space distribution. We found that increasing or reducing the expressiveness of the likelihood in some models yield better accuracy and better calibration, and we think there is still a margin of improvement.

Finally, we realized that the BNNs are suitable and robust in the experiments carried out. In many experiments we found that all the tested configurations clearly outperform temp-scal, as an example see figure 2.

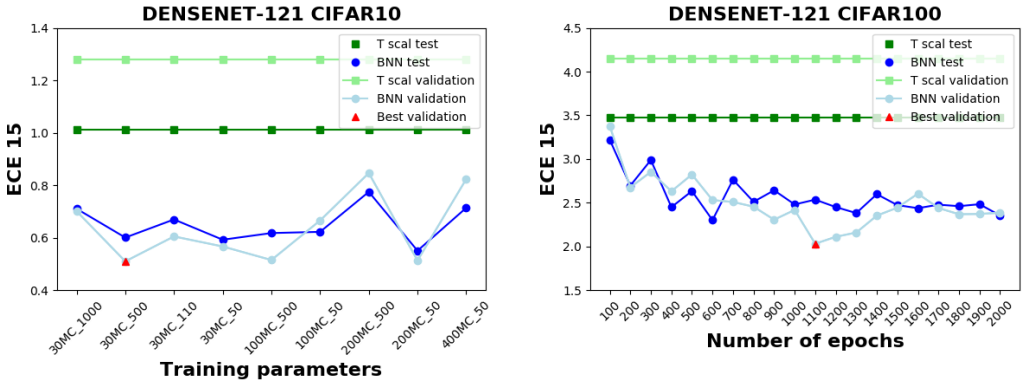

Figure 2: This figures compares the ECE performance for temp-scal and BNN in test and validation. On the left (CIFAR10) we show the performance of different training parameters. For example 30MC_500 means that the ELBO was optimized using 30 MC steps to estimate the expectation and 500 epochs of Adam optimization. On the right (CIFAR100) we show the performance of a BNN trained with different number of epochs up to 2000, showing the robustness against the course of learning.

5 DISCUSSION AND FUTURE WORK

This work has focused on showing that as long as uncertainty is correctly addressed we can improve calibration of state-of-the-art neural networks over temp-scal, which is the state-of-the-art to the best of our knowledge. Our offline approach can be applied to any deep model making it a desirable choice. We can calibrate outputs of any model without the need to apply Bayesian techniques directly to them, compromising our performance only to the number of classes and, being able to apply any state-of-the-art Bayesian Neural Networks techniques to the calibration model.

For our variatonal approximation we have chosen the most simple amortized variational inference distribution, whose training is based on the reparameretization trick (Rezende et al., 2014; Kingma & Welling, 2014). Our proposed approach provides several benefits. First, the variational parameters are predicted in a single forward step, from where we sample the neural network parameters. Second, it easily optimizes this variational parameters, as the gradients have low variance and can be computed by back propagation in a single backward step. Finally, it uses general parameters to approximate the intractable posterior distribution, avoiding the definition of a per sample variational parameter. The results obtained in our experiments are significant in terms of ECE reduction, thus calibration. However, we expect to improve even more the calibration performance and efficiency, i.e lower topologies, and solve the accuracy degradation, by using more sophisticated approximations such as local reparameretization (Kingma et al., 2015), auxiliary variables (Agakov & Barber, 2004; Ranganath et al., 2016), normalizing flows (Kingma et al., 2016; Louizos & Welling, 2017; Huang et al., 2018; van den Berg et al., 2018) and Monte Carlo Dropout (Gal & Ghahramani, 2015), among others. Uncertainty provided by dropout regularized networks (Gal & Ghahramani, 2016) cannot

be used for this purpose as the theoretical framework ensures uncalibration (see section 5.1.1) of the appendix, so although we can estimate uncertainty from dropout point estimate networks, the predictive distribution is not calibrated. On the other hand, as recently introduced (Cremer et al., 2018), amortized inference lead to an additional gap in the bound, in addition to the D_{KL} gap between the true and variational posterior and recent proposals mitigate this effect (Shu et al., 2018; Kim et al., 2018).

Regarding computation performance, BNNs are more inefficient in comparison to temp-scal. However as our Bayesian approach uses fast fully connected and because both approaches need to forward through a deep convolutional model (so in a real application for both methods we have to run through the deep convolutional network), we can mitigate this problem by using parallel computing platforms and libraries such as CuBLAS to finally evaluate the predictive distribution in only one step. Also, our approach is faster than directly applying Bayesian methods over the convolutional model because expectations of equation 2 require several deep convolutional forward and backward steps and also avoids the need to retrain them.

6 CONCLUSIONS

In this work we have shown that a Bayesian reasoning with high expressive models can improve calibration over much simpler methods that perform well, concluding that the calibration space is not simple. We have also shown that the BNN are robust models as they perform very well in many configurations, making them suitable for other applications. Moreover we show that can be applied in an efficient way. We let for future work investigating more sophisticated approximations with a clear objective: improve calibration and improve efficiency.

REFERENCES

Felix V. Agakov and David Barber. An auxiliary variational method. In *Neural Information Processing, 11th International Conference, ICONIP 2004, Calcutta, India, November 22-25, 2004, Proceedings*, pp. 561–566, 2004. doi: 10.1007/978-3-540-30499-9_86. URL https://doi.org/10.1007/978-3-540-30499-9_86.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1613–1622. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045290.

N. Brümmer and D. van Leeuwen. On calibration of language recognition scores. In *Proc. of Odyssey*, San Juan, Puerto Rico, 2006.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1721–1730, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2788613. URL http://doi.acm.org/10.1145/2783258.2788613.

Edward Challis and David Barber. Affine independent variational inference. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pp. 2195–2203, 2012. URL http://papers.nips.cc/paper/4512-affine-independent-variational-inference.

Tongfei Chen, Jirí Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam. Confidence scoring using whitebox meta-models with linear classifier probes. *CoRR*, abs/1805.05396, 2018. URL http://arxiv.org/abs/1805.05396.

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4467–4475. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7033-dual-path-networks.pdf.

I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. In *Knowledge Discovery in Databases: PKDD 2004*, volume 3202 of *Lecture Notes in Computer Science*, Heidelberg - Berlin, 2004. Springer. doi: https://doi.org/10.1007/978-3-540-30116-5_14.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1086–1094, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/cremer18a.html.

Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *CoRR*, abs/1802.04865, 2018.

Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *Trans. Info. For. Sec.*, 9(12):2170–2179, December 2014. ISSN 1556-6013. doi: 10.1109/TIFS.2014.2359646. URL https://doi.org/10.1109/TIFS.2014.2359646.

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *CoRR*, abs/1506.02158, 2015. URL http://arxiv.org/abs/1506.02158.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1050–1059. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045502.

Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2348–2356. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. On integrating a language model into neural machine translation. *Comput. Speech Lang.*, 45(C):137–148, September 2017. ISSN 0885-2308. doi: 10.1016/j.csl.2017.01.014. URL https://doi.org/10.1016/j.csl.2017.01.014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/guo17a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016b. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, , and Brian Kingsbury. Deep neural networks for acoustic modelling in speech recognition. the shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. doi: 10.1109/MSP.2012.2205597.

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, pp. 5–13, New York, NY, USA, 1993. ACM. ISBN 0-89791-611-5. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron C. Courville. Neural autoregressive flows. In *ICML*, 2018.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. Semi-amortized variational autoencoders. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2683–2692, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kim18e.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2575–2583. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5666-variational-dropout-and-the-local-reparameterization-trick.pdf.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4743–4751. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6581-improved-variational-inference-with-inverse-autoregressive-flow.pdf.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, Cambridge, MA, USA, 2009. ISBN 0262013193, 9780262013192.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). a. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). b. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6402–6413. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference On Learning Representations*, abs/1711.09325, 2018.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2218–2227, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/louizos17a.html.

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1445–1454. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045543.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a. URL http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119, USA, 2013b. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1791–II–1799. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3045092.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2901–2907. AAAI Press, 2015. ISBN 0-262-51129-0. URL http://dl.acm.org/citation.cfm?id=2888116.2888120.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011. URL https://www-cs.stanford.edu/~twangcat/papers/nips2011_housenumbers.pdf.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, Bonn, Germany, 2005. doi: 10.1145/1102351.1102430.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548, 2017.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pp. 61–74. MIT Press, 1999.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 324–333, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/ranganath16.html.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32Nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1530–1538. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045281.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/rezende14.html.

Rui Shu, Hung H. Bui, Shengjia Zhao, Mykel J. Kochenderfer, and Stefano Ermon. Amortized inference regularization. *CoRR*, abs/1805.08913, 2018. URL http://arxiv.org/abs/1805.08913.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pp. 4278–4284, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806.

Zoltán Tüske, Ralf Schlüter, and Hermann Ney. Investigation on LSTM recurrent n-gram language models for speech recognition. In *Interspeech*, pp. 3358–3362. ISCA, 2018.

Rianne van den Berg, Leonard Hasenclever, Jakub M. Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. *CoRR*, abs/1803.05649, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, 2017.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proceeding of the Eight International Conference on Knowledge Discovery and Data Mining (KDD'02)*, 2002a. doi: 10.1145/775047.775151.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655658.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 694–699, New York, NY, USA, 2002b. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775151. URL http://doi.acm.org/10.1145/775047.775151.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

Yichuan Zhang, José Miguel Hernández-Lobato, and Zoubin Ghahramani. Variational measure preserving flows. *CoRR*, abs/1805.10377, 2018. URL http://arxiv.org/abs/1805.10377.