Appendix for the work: Calibration of Deep Probabilistic Models with Decoupled Bayesian Neural Networks

1. Calibration Results: Explicit Techniques

Table 1. ECE 15(%) and ACC(%) comparing model uncalibrated, calibrated with TS, with MFVI, with MFVILR and with decoupled NE.

CIFAR10

	uncalibrated		Temp Scal		MFVI		MFVILR*		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
WideResNet 28x10	96.13	1.84	96.13	0.52	96.08	0.24	95.94	0.44	95.70	1.40
DenseNet 121	95.49	2.64	95.49	1.01	95.26	0.60	95.29	0.43	95.08	2.21
DenseNet 169	95.49	2.66	95.49	0.83	95.29	0.51	95.37	0.38	95.09	2.27
Dual Path Network 92	95.18	3.00	95.18	1.07	95.03	0.73	94.96	0.62	94.52	2.62
ResNet 101	93.46	4.27	93.46	1.20	93.38	0.78	93.11	0.63	93.37	3.38
VGG 19	93.68	4.41	93.68	1.71	93.67	0.84	93.52	0.65	93.44	3.19
Preactivation ResNet 18	94.93	3.16	94.93	0.57	94.73	0.45	94.8	0.44	94.74	2.43
Preactivation ResNet 164	93.91	4.10	93.91	0.44	93.82	0.33	93.89	0.30	93.92	3.04
ResNext 29_8x16	94.79	2.83	94.79	0.74	94.61	0.73	94.58	0.71	94.69	2.10
Wide ResNet 40x10	95.01	3.00	95.01	0.92	95.08	0.59	94.97	0.41	95.03	2.87

SVHN

	uncalibrated		Temp Scal		MFVI		MFVILR*		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
WideResNet 40x10	96.95	1.26	96.95	1.17	96.90	1.15	96.82	1.11	96.70	0.84
Densenet-121	96.76	2.02	96.76	1.09	96.70	0.72	96.74	1.06	96.38	1.04
Densenet-169	96.70	0.36	96.70	1.02	96.59	0.45	96.62	0.60	96.68	0.87
ResNet 50	96.47	0.89	96.47	1.03	96.33	0.86	96.35	0.87	96.39	1.42
Preactivation ResNet 164	96.20	2.54	96.20	1.08	96.08	0.92	96.09	0.85	96.02	1.53
Wide ResNet 16x8	96.88	0.71	96.88	1.32	96.82	0.74	96.92	0.70	97.00	0.83
Preactivation ResNet 18	96.15	1.57	96.15	0.65	96.05	1.10	96.05	0.83	95.88	0.59
WideResNet 28x10	96.62	1.48	96.62	0.93	96.54	1.03	96.78	0.74	96.31	1.03

CIFAR100

	uncalibrated		Temp Scal		MFVI		MFVILR		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
WideResNet 28x10	80.39	4.85	80.39	4.28	77.59	2.46	78.54*	2.59	78.879	7.31
DenseNet 121	78.80	8.72	78.80	3.48	75.90	2.53	76.53*	2.47	78.09	8.91
DenseNet 169 Network	79.05	8.88	79.05	3.76	75.58	2.39	77.22*	2.45	78.38	8.93
ResNet 101	72.00	11.41	72.00	1.53	68.59	1.61	70.31*	1.75	71.40	12.77
VGG 19	72.70	17.63	72.70	4.80	71.94	6.00	71.61*	6.07	70.60	16.49
Preactivation ResNet 18	76.60	10.78	76.90	3.15	74.30	1.76	74.51*	1.59	75.70	9.23
Preactivation ResNet 164	73.28	15.75	73.28	2.05	70.77*	1.46	71.16	2.20	73.04	11.29
ResNext 29_8x16	77.88	9.68	77.88	2.81	73.97*	2.58	71.13	3.77	77.35	6.41
Wide ResNet 40x10	76.74	14.77	76.74	3.77	76.17	1.88	76.51*	1.79	77.67	10.21

ADIENCE

	uncalibrated		Temp Scal		MFVI		MFVILR*		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
VGG-19	94.54	4.20	94.54	0.77	94.53	0.44	94.51	0.46	94.02	2.08
DenseNet 121	93.96	4.90	93.96	0.96	94.03	0.61	94.03	0.55	94.60	3.20

VGGFACE2

	uncalibrated		Temp Scal		MFVI		MFVILR*		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
MobileNet	96.76	0.93	96.76	0.37	-	-	96.76	0.29	96.74	0.72
SeNet	96.96	2.50	96.96	0.68	-	-	96.97	0.41	97.02	1.05
VGG	94.84	0.57	94.84	0.60	-	-	94.87	0.41	94.89	0.61

CARS

	uncalibrated		Temp Scal		MFVI		MFVILR*		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
DenseNet-169	88.98	5.78	88.98	1.94	-	-	85.27	1.98	89.34	6.02
DenseNet-121	88.87	5.83	88.87	1.67	-	-	85.43	1.31	89.26	5.83
ResNet-18	86.56	7.00	86.56	1.51	-	-	83.33	1.58	86.12	5.94
ResNet-50	89.84	5.07	89.84	2.06	-	-	87.04	1.73	89.55	4.93
ResNet-101	89.71	5.36	89.71	1.83	-	-	85.63	1.34	89.78	4.81

BIRDS

	uncalibrated		Temp Scal		MFVI		MFVILR*		NE	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
DenseNet-169	77.49	12.63	77.49	1.92	-	-	75.19	1.67	78.43	5.82
DenseNet-121	76.77	12.38	76.77	2.29	-	-	74.03	1.62	77.84	6.43
ResNet-18	72.49	15.62	72.49	2.84	-	-	71.40	2.44	73.66	5.43
ResNet-50	76.43	12.95	76.43	2.71	-	-	75.56	1.76	77.54	4.51
ResNet-101	78.15	12.53	78.15	2.27	-	-	75.40	1.93	78.15	4.06

The above tables show the ECE and accuracy results for decoupled technique used to compute the average results from the main work. With an * we mark which model was the best on validation. In general we see that by only applying LR we achieve better calibration and increase the accuracy in the models where the BNN slightly degraded the accuracy.

As we state in the main article, for CARS and BIRDS we suffered from accuracy degradation during training, no matter how big the topology was. Thus, we only report results using BNN-LR. The high dimensionality of these tasks might be the reason for this degradation. The minimization is correctly performed because the NNL is correctly minimized, but the high variance of the estimator does not allows to reach a good optimum. For VGGFACE2 we only trained BNN-LR because of the high training time required. Remember BNN-LR converge faster and requires less time per batch than BNN. The details of the architecture and training algorithm parameters are provided in the GitHub.

Moreover, as we mentioned in the main work, we found some instabilities when using shallow architectures on 2-dimensional problems in the BNN-LR setting. More precisely, we found that on the first backward operation, the gradient w.r.t the variance σ_i of the distribution of each weight w_i saturate. We analyze the gradient w.r.t this parameter and realized that in the case of BNN-LR the gradient scales quadratically with the logit value, with a normalization factor that depends linearly on the number of hidden units. This means that with sallower topologies this normalization factor is not enough to compensate the potentially high numerator. On the other hand, in standard BNNs, the gradient of the variance scales linearly with the logit value, and these instabilities do not appear, allowing for shallower architectures. We try to solve this problem by re-scaling the logit values, constraining the variance parameter, or controlling the parameter initialization. We solved the problem for just a few epochs before the model again saturated. In practice using a bigger topology was a better solution to the problem.

2. Calibration Results: Implicit Techniques

In this section we show the results used to compute the average for the implicit techniques. Due to the high computation cost of some of these techniques we use only three databases and a subset of the DNNs considered. The usage of dropout or not depends on whether we use it in the decoupled techniques. The first table shows de results on different predictive samples for the Monte Carlo Dropout approach.

Table 2. Table showing ACC and ECE for the Monte Carlo Dropout technique and the three databases considered in this work. We show results on different Monte Carlo predictive samples. With and * we mark the ones used for the average result displayed in the main work. The results from the work are computed using the 100 poredictive samples.

MONTE CARLO DROPOUT

		CIFAR10		CIFAR100		SVHN	
	MC samples	ACC	ECE	ACC	ECE	ACC	ECE
DenseNet-121	1	94.70	3.77	77.75	11.24	96.39	2.92
	25	93.84	1.31	75.53	2.34	96.49	1.34
	50	93.84	1.03	75.70	2.24	96.50	1.34
	75	93.79	1.17	75.64	1.99	96.51	1.31
	100	93.78	1.02	75.66	2.01	96.50	1.30
WideResNet28x10	1	95.25	3.06	77.98	11.82	96.69	0.56
	25	94.47	1.38	79.35	3.48	96.88	1.04
	50	94.43	1.34	79.39	3.30	96.87	1.03
	75	94.43	1.34	79.37	3.37	96.88	0.99
	100	94.44	1.35	79.36	3.33	96.88	1.02
WideResNet40x10	1	95.35	3.20	78.15	12.65	-	-
	25	95.14	1.80	78.44	5.25	-	-
	50	95.17	1.76	78.45	5.18	-	-
	75	95.14	1.75	78.48	5.16	-	-
	100	95.15	1.71	78.56	5.13	-	-
WideResNet16x8	1	-	-	-	-	96.86	0.44
	25	-	-	-	-	96.95	0.59
	50	-	-	-	-	96.93	0.50
	75	-	-	-	-	96.92	0.46
	100	-	-	-	-	96.90	0.46

The next table shows the results using Network Ensembles. We both provide the baseline result (1 ensemble) alongside with the 5 ensemble. We experiment both with the default adversarial value as noted in the original work and also with and adversarial factor such that the perturbation norm is below the quantification error. The result of WideResnet-40x10 with adversarial and dropout 0.3 is not used in the final average due to its bad performance. This somehow illustrate the problematic of hyperparameter search for this technique. Combining the default adversarial and a dropout WideResnet-40x10 (which in the Wide Resnet original work is reported as one of the best performing models on CIFAR100) presents very bad performance.

Table 3. Table showing the results for the original Network Ensembles. * shows the results used for the paper.

NETWORK ENSEMBLES

			CIFAR10				CIFAR100				SVHN			
			1 ensemble		5 ensemble		1 ensemble		5 ensemble		1 ensemble		5 ensemble	
	Dropout Rate	Adversarial Factor	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
WideResNet 28x10	0.0	0.05	96.12	2.22	96.91	*0.59	-	-	-	-	-	-	-	-
	0.0	0.00976	95.75	2.379	96.48	*0.65	-	-	-	-	-	-	-	-
	0.0	0.0195	-	-	-	-	-	-	-	-	96.55	1.85	97.27	0.58*
	0.0	0.00781	-	-	-	-	80.21	5.60	82.86	2.42*	-	-	-	-
WideResNet 40x10	0.0	0.00976	96.3	2.24	96.96	*0.62	-	-	-	-	-	-	-	-
	0.3	0.0	-	-	-	-	77.24	12.91	79.76	5.04*	-	-	-	-
	0.0	0.00781	-	-	-	-	80.39	9.36	82.89	2.36*	-	-	-	-
	0.3	0.00781	-	-	-	-	64.27	20.96	66.34	10.86	-	-	-	-
WideResNet 16x8	0.0	0.01	-	-	-	-	-	-	-	-	96.27	0.94	97.37	0.87*
	0.0	0.0195	-	-	-	-	-	-	-	-	97.00	0.5	97.22	0.69*

Finally the next table shows the results for the MMCE technique. We use the default hyperparameter as provided in the original work.

Table 4. Table showing ACC and ECE for the Monte Carlo Dropout technique and the three databases considered in this work. We show results on different Monte Carlo predictive samples. With and * we mark the ones used for the average result displayed in the main work. The results from the work are computed using the 100 poredictive samples.

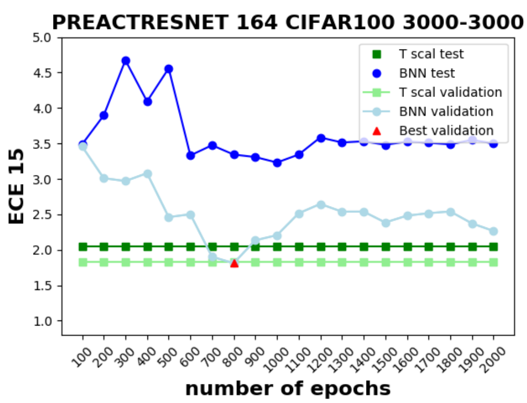
MMCE

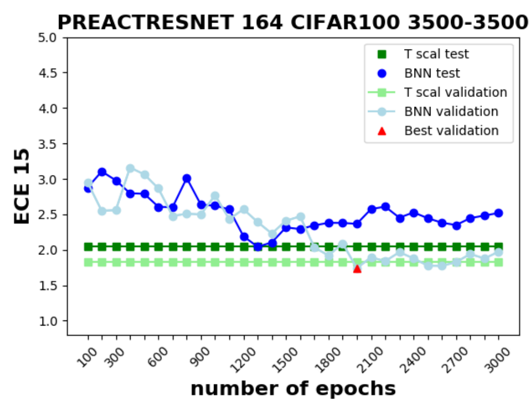
		CIFAR10		CIFAR100		SVHN	
	Dropout	ACC	ECE	ACC	ECE	ACC	ECE
DenseNet-121	0.0	93.72	2.38	73.02	6.41	96.65	1.76
WideResNet-28x10	0.0	95.58	1.21	74.98	7.04	-	-
WideResNet-16x8	0.0	-	-	-	-	96.65	0.49

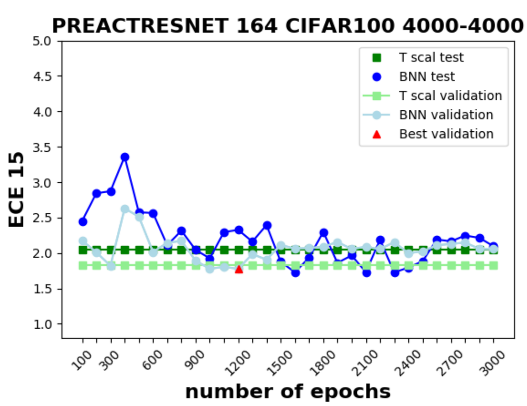
3. Robustness of Bayesian Neural Networks

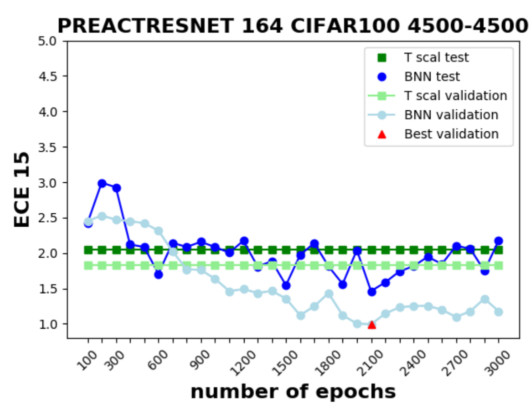
3.1. Increasing topology increases calibration performance

This subsection shows how the calibration performance is improved by increasing the expressiveness of the likelihood model in the MFVI approach. The number on the title indicate the model topology. For instance, 3500-3500 means two hidden layers of 3500 neurons each.



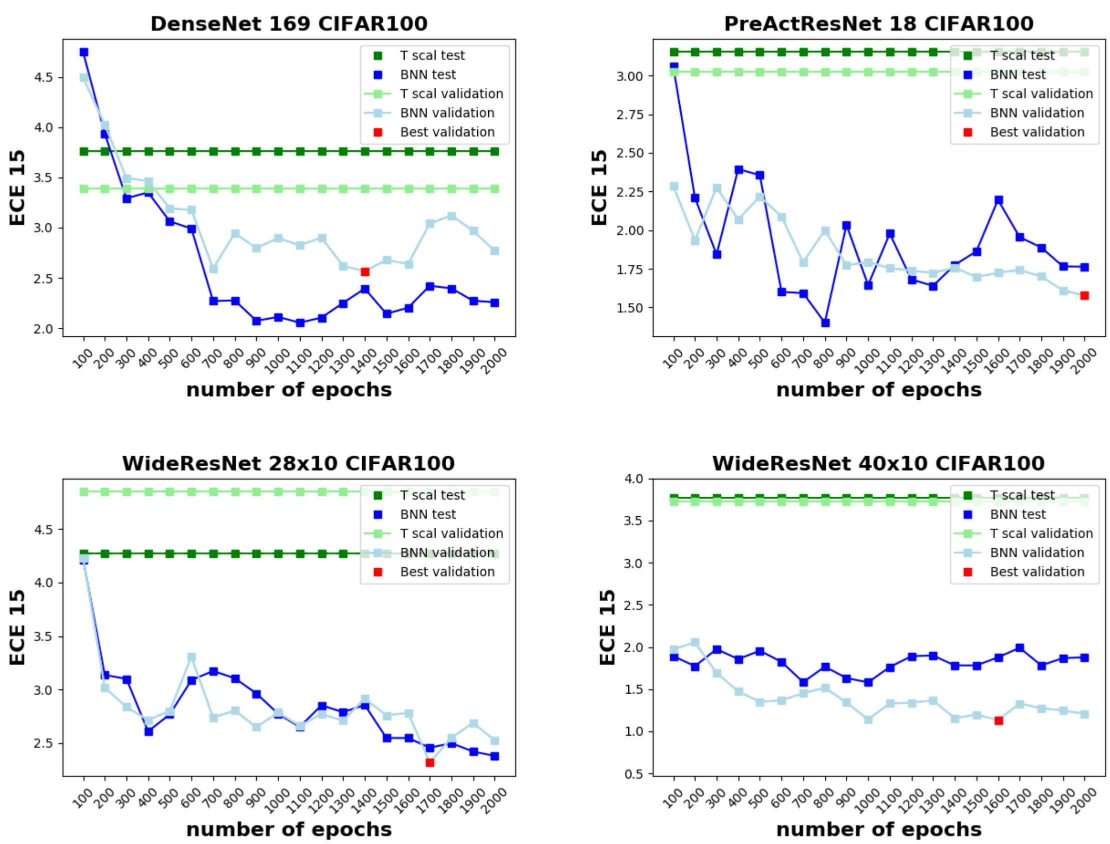


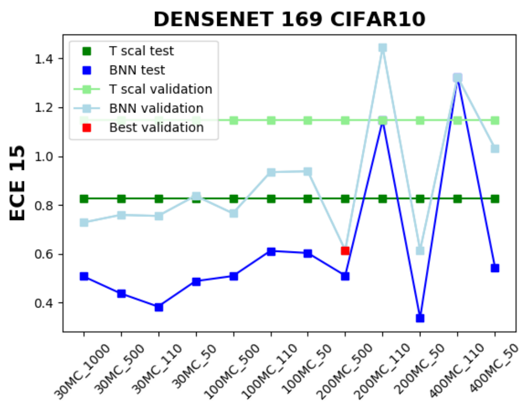


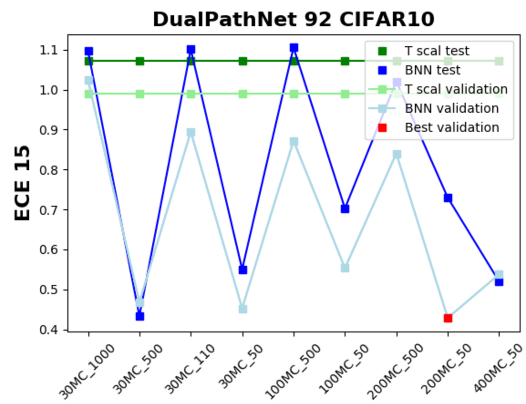


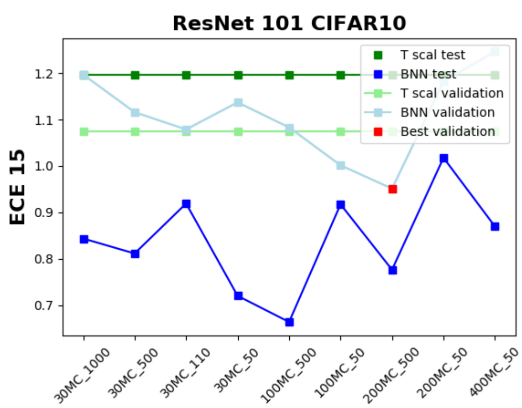
3.2. Comparison of different models and datasets

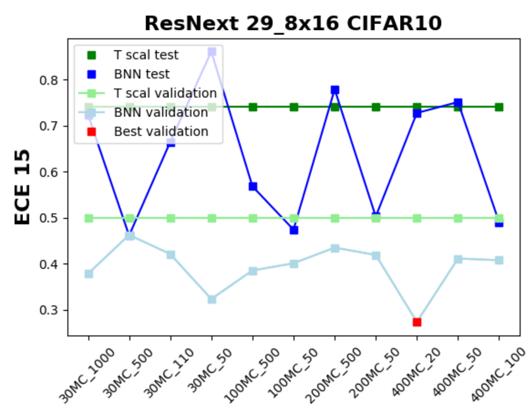
This section illustrate the robustness of the BNN when used for improving calibration performance over TS. We show figures comparing BNN with TS on different networks and datasets. We see how clearly different configurations of the BNNs outperform TS.

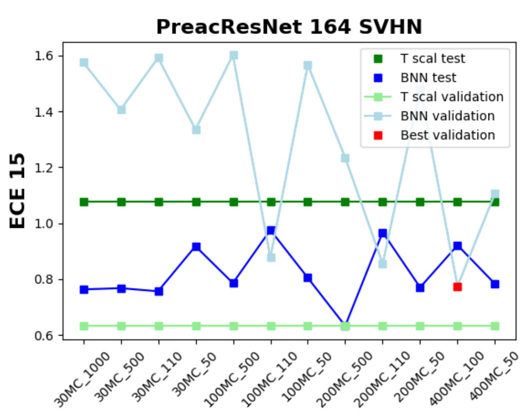



DENSENET 169 CIFAR10
ECE 15
T scal test
BNN test
T scal validation
BNN validation
Best validation
30MC_1000
30MC_500
30MC_110
30MC_50
100MC_500
100MC_110
100MC_50
200MC_500
200MC_110
200MC_50
400MC_110
400MC_50


DualPathNet 92 CIFAR10
ECE 15
T scal test
BNN test
T scal validation
BNN validation
Best validation
30MC_1000
30MC_500
30MC_110
30MC_50
100MC_500
100MC_50
200MC_500
200MC_50
400MC_50


ResNet 101 CIFAR10
ECE 15
T scal test
BNN test
T scal validation
BNN validation
Best validation
30MC_1000
30MC_500
30MC_110
30MC_50
100MC_500
100MC_50
200MC_500
200MC_50
400MC_50


ResNext 29_8x16 CIFAR10
ECE 15
T scal test
BNN test
T scal validation
BNN validation
Best validation
30MC_1000
30MC_500
30MC_110
30MC_50
100MC_500
100MC_50
200MC_500
200MC_50
400MC_20
400MC_50
400MC_100


PreacResNet 164 SVHN
ECE 15
T scal test
BNN test
T scal validation
BNN validation
Best validation
30MC_1000
30MC_500
30MC_110
30MC_50
100MC_500
100MC_110
100MC_50
200MC_500
200MC_110
200MC_50
400MC_100
400MC_50