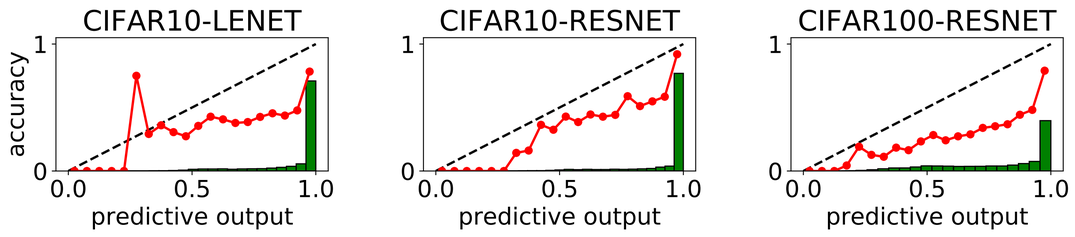
Calibrating Deep Convolutional Gaussian Processes

G. L. Tran[1], E. V. Bonilla[2], J. P. Cunningham[3], P. Michiardi[1], M. Filippone[1]

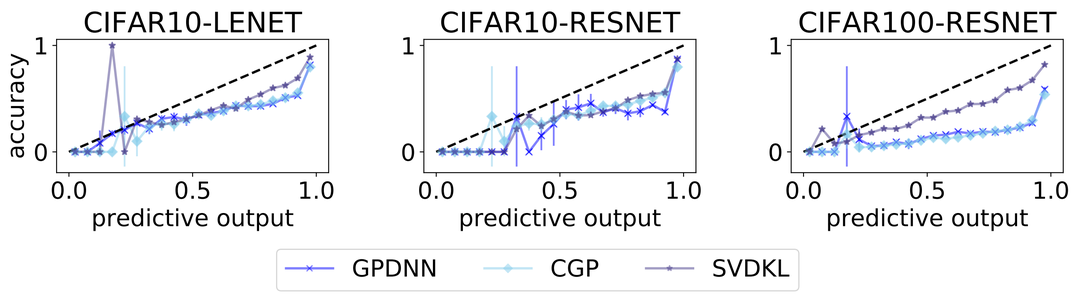
1 - EURECOM, France 2 - CSIRO's Data61 and UNSW, Australia 3 - Columbia University, USA

Miscalibration of Convolutional Neural Networks

- The classification networks in a decision making system must not only be accurate, but also should indicate when they are likely to be incorrect, i.e., model's calibration.
- Convolutional Neural Networks (CNNs) were found to be miscalibrated.



- The current combinations of CNNs and Gaussian Processes (GPs) are miscalibrated.



Gaussian Processes with Random Features

- Convolutional features $\mathbf{C}(\mathbf{X}|\mathbf{\Psi})$ are fed into a GPs.
- Low-rank approximation for kernel matrix:

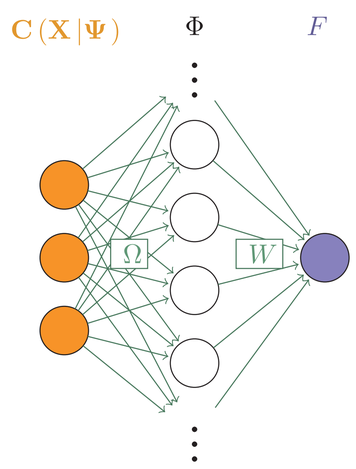
$$\mathbf{K} \approx \mathbf{\Phi}\mathbf{\Phi}^T$$

- Taking a weight-space view of a GP:

$$\mathbf{F} = \mathbf{\Phi}\mathbf{W}$$

- The priors over the weights are:

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$



- The order-one ARC-COSINE kernel:

$$k_{\text{arc}}^{(1)}(\mathbf{x}_i, \mathbf{x}_j|\mathbf{\Psi}, \boldsymbol{\theta}) = \frac{\sigma^2}{\pi} \left\|\Lambda^{-\frac{1}{2}}\mathbf{c}(\mathbf{x}_i|\mathbf{\Psi})\right\| \left\|\Lambda^{-\frac{1}{2}}\mathbf{c}(\mathbf{x}_j|\mathbf{\Psi})\right\| [\sin(\alpha) + (\pi - \alpha)\cos(\alpha)], \text{ where}$$

$$\boldsymbol{\theta} = (\sigma, \mathbf{\Lambda} = \text{diag}(\ell_1^2, \ldots, \ell_d^2)) \text{ and } \alpha = \cos^{-1}\frac{\left(\Lambda^{-\frac{1}{2}}\mathbf{c}(\mathbf{x}_i|\mathbf{\Psi})\right)^T \left(\Lambda^{-\frac{1}{2}}\mathbf{c}(\mathbf{x}_j|\mathbf{\Psi})\right)}{\left\|\Lambda^{-\frac{1}{2}}\mathbf{c}(\mathbf{x}_i|\mathbf{\Psi})\right\| \left\|\Lambda^{-\frac{1}{2}}\mathbf{c}(\mathbf{x}_j|\mathbf{\Psi})\right\|}$$

can be approximated using Rectified Linear Units (ReLU):

$$\mathbf{\Phi}_{\text{arc}} = \sqrt{\frac{2\sigma^2}{N_{\text{RF}}}} \max(\mathbf{0}, \mathbf{C}(\mathbf{X}|\mathbf{\Psi})\,\mathbf{\Omega}), \text{ where } \mathbf{\Psi} \text{ are convolutional parameters or filters}$$

- $\mathbf{\Omega}$ can be structured through a series of Hadamard transformations of diagonal matrices $\mathbf{D}_i$ with elements randomly sampled from $\{-1, +1\}$:

$$\mathbf{\Omega} = \frac{\sqrt{d}}{l}\mathbf{H}\mathbf{D}_1\mathbf{H}\mathbf{D}_2\mathbf{H}\mathbf{D}_3, \text{ where } d \text{ is the dimensionality of the convolutional features}$$

Stochastic Variational Learning

- The lower bound on the log-marginal likelihood $\mathcal{L} = \log[p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})]$:

$$\mathcal{L} \geq \text{E}_{q(\mathbf{W},\mathbf{\Omega},\mathbf{\Psi})}(\log[p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi}, \boldsymbol{\theta})]) - \text{KL}[q(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi}) \,\|\, p(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi})]$$

where $q(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi})$ approximates $p(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi}|\mathbf{X}, \mathbf{Y})$.

- Factorization of approximate posterior:

$$q(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi}) = q(\mathbf{W})\,q(\mathbf{\Omega})\,q(\mathbf{\Psi}) = \prod_i q(\mathbf{W}_i) \prod_i q(\mathbf{\Omega}_i) \prod_i q(\mathbf{\Psi}_i), \text{ where}$$

$$q(\mathbf{W}_i) = \pi_w \mathcal{N}([\mathbf{M}_w]_r, \sigma^2\mathbf{I}) + (1 - \pi_w)\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$
$$q(\mathbf{\Omega}_i) = \pi_\Omega \mathcal{N}([\mathbf{M}_\Omega]_r, \sigma^2\mathbf{I}) + (1 - \pi_\Omega)\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$
$$q(\mathbf{\Psi}_i) = \pi_\Psi \mathcal{N}([\mathbf{M}_\Psi]_r, \sigma^2\mathbf{I}) + (1 - \pi_\Psi)\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

$\mathbf{M}_w, \mathbf{M}_\Omega$ and $\mathbf{M}_\Psi$ are called as variational parameters; $\pi_w, \pi_\Omega; \pi_\Psi$ are probabilities and $\sigma^2 \approx 0$.

- Assuming a normal prior, the KL term can be approximated:

$$\text{KL}[q(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi}) \,\|\, p(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi})] \approx \frac{\pi_w}{2}\|\mathbf{M}_w\|^2 + \frac{\pi_\Omega}{2}\|\mathbf{M}_\Omega\|^2 + \frac{\pi_\psi}{2}\|\mathbf{M}_\psi\|^2$$

- The expectation can be unbiasedly estimated using Monte Carlo and mini-batch of size m:

$$\text{E}_{q(\mathbf{W},\mathbf{\Omega},\mathbf{\Psi})}(\log[p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathbf{\Omega}, \mathbf{\Psi}, \boldsymbol{\theta})]) \approx \frac{n}{m}\frac{1}{N_{\text{MC}}}\sum_{i=1}^{N_{\text{MC}}}\sum_{k\in\mathcal{I}_m} \log\left[p\left(\mathbf{y}_k|\mathbf{x}_k, \tilde{\mathbf{W}}^{(i)}, \tilde{\mathbf{\Omega}}^{(i)}, \tilde{\mathbf{\Psi}}^{(i)}, \boldsymbol{\theta}\right)\right]$$

- Due to the assumption of $\sigma^2 \approx 0$, we can sample from the above posteriors as follows:

$$\tilde{\mathbf{W}}^{(i)} = \mathbf{M}_w\text{diag}[\tilde{\mathbf{z}}_w^i] \quad \text{and} \quad \tilde{\mathbf{\Omega}}^{(i)} = \mathbf{M}_\Omega\text{diag}[\tilde{\mathbf{z}}_\Omega^i] \quad \text{and} \quad \tilde{\mathbf{\Psi}}^{(i)} = \mathbf{M}_\Psi\text{diag}[\tilde{\mathbf{z}}_\Psi^i]$$

where $\tilde{\mathbf{z}}_w^i, \tilde{\mathbf{z}}_\Omega^i$ and $\tilde{\mathbf{z}}_\Psi^i$ are sampled from $\text{Bernoulli}(\pi_w), \text{Bernoulli}(\pi_\Omega)$ and $\text{Bernoulli}(\pi_\Psi)$.

Experiments

- Competing methods:

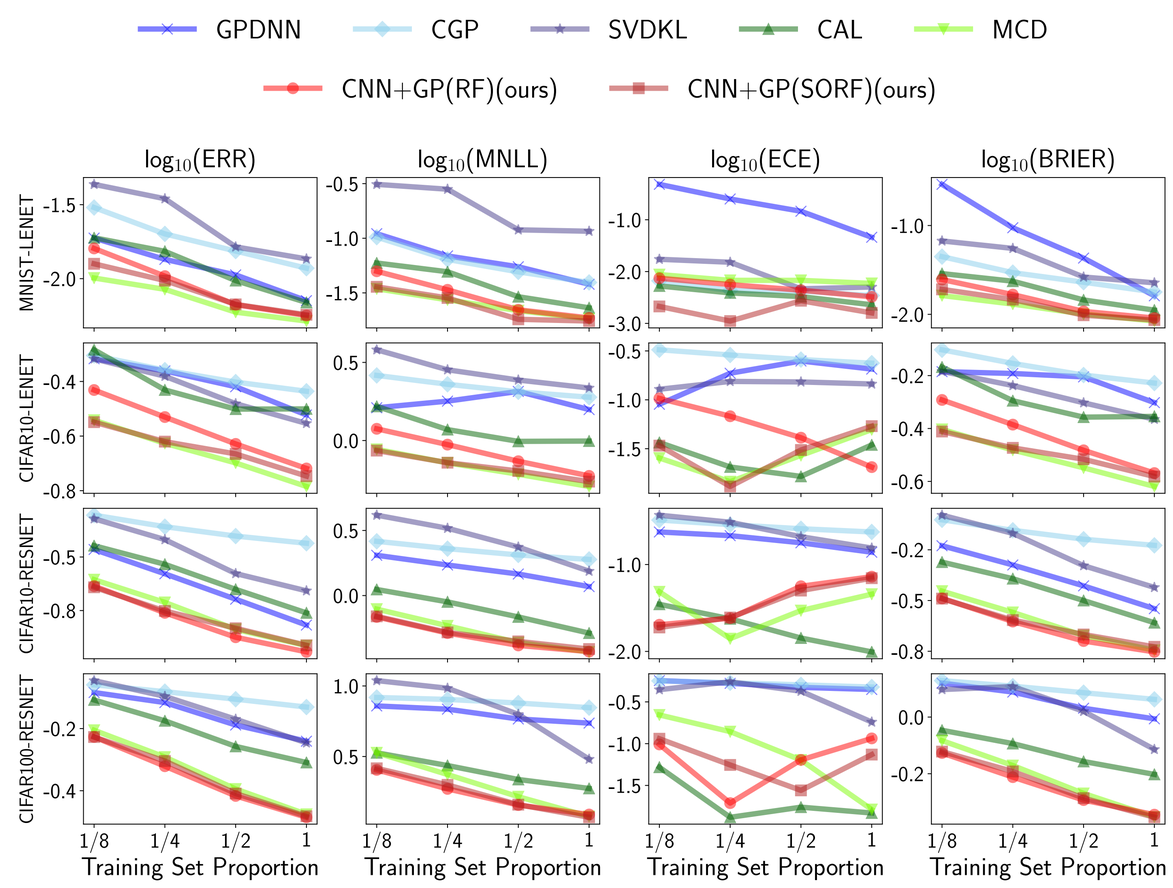


Figure 1: Comparison of our CNN+GP(RF) and CNN+GP(SORF) with competitors.

- Reliability diagram:

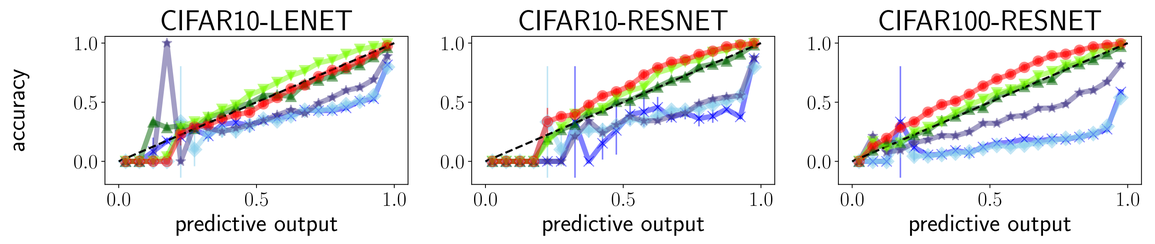


Figure 2: Reliability diagrams of our CNN+GP(RF) in comparison with competitors.

- Combination of CNNs and deep GPs:

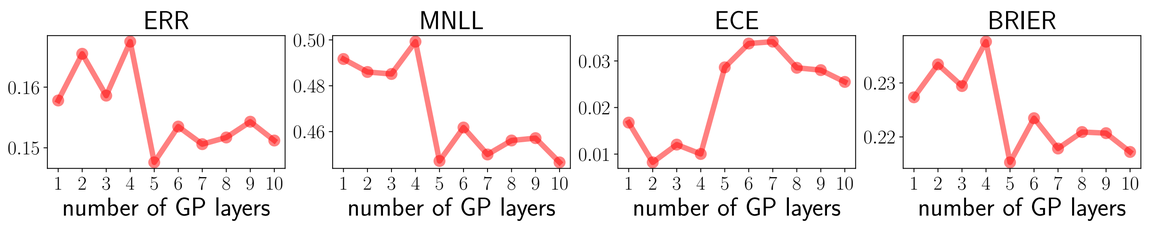


Figure 3: Performance of the proposed model when varying the number of GP layers on top of a LENET convolutional structure on CIFAR10 dataset.

- Ability to detect out-of-distribution samples:

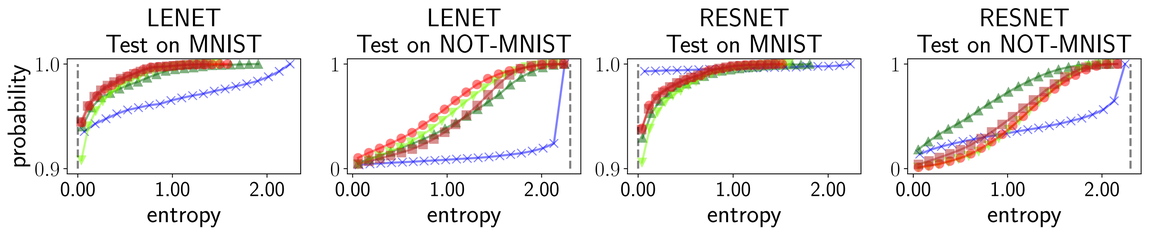


Figure 4: Cumulative distribution function plot of predictive entropies when the models trained on MNIST are tested on MNIST and NOT-MNIST.

- Substitution of fully connected layers with Structured Orthorgonal Random Features (SORF):

Table 1: CIFAR10

METHOD	ERR	MNLL	ECE	BRIER
SORF	0.172	0.522	0.063	0.250
MCD	0.181	0.591	0.110	0.276

Table 2: CIFAR100

METHOD	ERR	MNLL	ECE	BRIER
SORF	0.459	1.806	0.127	0.612
MCD	0.594	2.434	0.058	0.732

Contributions

- Show that current combinations of CNNs and GPs are miscalibrated.
- Propose a novel combination of CNNs and GPs that is well-calibrated.
- Extend the model by replacing the last layer of CNNs with deep GPs.
- Obtain a compact approximation of GP by using SORF.

References

[1] Guo et al. *On Calibration of Modern Neural Networks*, ICML 2017.
[2] Bradshaw et al. *Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks*, https://arxiv.org/abs/1707.02476.
[3] Wilk et al. *Convolutional Gaussian Processes*, NeurIPS 2017.
[4] Wilson et al. *Stochastic Variational Deep Kernel Learning*, NeurIPS 2016.
[5] Cutajar et al. *Random Feature Expansions for Deep Gaussian Processes*, ICML 2017.
[6] Yu et al. *Orthogonal Random Features*, NeurIPS 2016.
[7] Gal and Gharamani. *Dropout as a Bayesian Approximation*, ICML 2016.