UNCERTAINTY ESTIMATION WITH A VAE-CLASSIFIER HYBRID MODEL

Shuyu Lin[1]*, Ronald Clark*[2]*, Niki Trigoni*[1]*, Stephen Roberts*[1]

[1] University of Oxford, Oxford OX1 2JD, UK
[2] Imperial College London, South Kensington, London SW7 2AZ, UK

ABSTRACT

We propose a hybrid model that combines a generative unit and a discriminative classifier to quantify uncertainty in a classification task. The representation learning capability in the VAE module allows our method to learn more useful and generalizable features and outperform other purely discriminative classifiers when training labels are limited. With proper statistical treatment, the probabilistic encoder in our VAE module offers a convenient mechanism to express uncertainty for out-of-distribution (OOD) data. As a result, our method gives better calibrated uncertainty prediction. We demonstrate the effectiveness of our method on MNIST and a challenging medical image dataset for skin lesion diagnosis.

***Index Terms*—** Uncertainty estimation, Label efficient learning, Deep learning, Generative modelling

1. INTRODUCTION

Uncertainty estimation has been a key topic in machine learning research [1]. We have seen machine learning algorithms achieve state-of-the-art performance in a wide range of tasks, such as image recognition [2] and natural language processing [3]. However, to deploy these algorithms in real world applications, we need them to be able to quantity the uncertainty associated with their prediction. This is crucial for high-stakes applications, such as healthcare and autonomous driving, where an over-confident wrong decision can be detrimental. In many situations, the machine algorithms will need to work alongside humans. The ability of these algorithms to sensibly predict uncertainty can help to indicate when human intervention might be needed and also help to build trust between the human operators and the machine algorithms.

Evaluating a method's uncertainty performance is a challenging task. We focus on two aspects that we consider are important for real world applications. Firstly, we examine calibration, which measures uncertainty from a frequentist's perspective and highlights the discrepancy between a method's predictive confidence and the empirical accuracy [4, 5, 6]. Intuitively, a well-calibrated method will report a 80% confidence when the prediction is correct 80% of the time. Therefore, calibration can give a clear indication whether a method is over- or under-confident in its prediction. Secondly, we examine whether a method can report sensible uncertainty when out-of-distribution (OOD) data are presented [7]. As domain shifts are common in real world applications, being honest to what a method does not know and reporting low confidence accordingly is an important property.

In this paper, we propose a hybrid classifier that contains a generative unit - in the form of a variational autoencoder (VAE [8, 9]) - to map the input to a low-dimensional representation and a discriminative unit to make the class prediction. This formulation can be very useful in practice, as it allows the generative unit to be trained with additional unlabelled data or the same generative unit to be reused for different classification tasks. We show that our method gives sensible uncertainty estimation in challenging scenarios when the training labels are limited or when OOD data are encountered. We also evaluate our method's performance in a public medical dataset, HAM10k [10], for pigmented skin lesion diagnosis. We show that our method gives the most calibrated confidence prediction when other methods tend to be over-confident on their predictions.

2. RELATED WORK

Uncertainty estimation for neural networks (NNs) is a fast growing field. Bayesian neural networks introduce priors over the network parameters and then infer the posterior distributions of the parameters given the training data. This can be done using sampling-based approaches such as MCMC [11]. Alternatively, the posterior distribution can be inferred using approximations, such as Laplace approximation [12], variational inference [13, 14] and Monte-Carlo dropout [15]s. Non-Bayesian approaches, such as training ensembles of deterministic NN classifiers [6] or re-calibrating the predictive distribution with held-out validation data [16], have also shown effective performance in terms of classification accuracy and calibration quality.

3. OUR METHOD

In this section, we first describe how we introduce a generative module in the classification task. We show how to make class prediction and uncertainty estimation in our hybrid classifier setup. We then introduce how to train our model and the statistical treatment needed to consider OOD data.

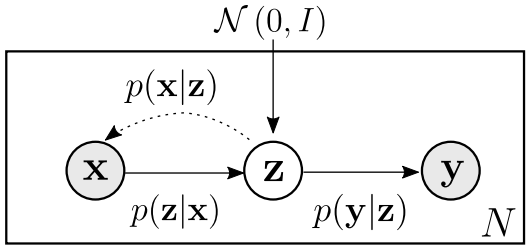

Fig. 1. **The graphical model for our hybrid classifier** that consists of a probabilistic encoder $p_\theta(\mathbf{z}|\mathbf{x})$ and a discriminative classifier $p_\theta(y|\mathbf{z})$.

3.1. Classification with a hybrid model

For our learning task, we assume a training dataset $\mathcal{D}_N$ that consists of N i.i.d. pairs of input-label samples, i.e. $\mathcal{D}_N = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where the input $\mathbf{x} \in \mathbb{R}^{L_x}$ represents a L_x-dimensional sample and the label y denotes one of K classes, i.e. $y \in \{1, \cdots, K\}$. During training, we aim to estimate the predictive distribution $p_\theta(y|\mathbf{x})$ over the class labels, where θ denotes the model parameters to be determined. To make a prediction for a new input observation $\boldsymbol{x}_t$, we compute the predictive distribution and the predicted label is sampled from this distribution, i.e. $y_t \sim p_\theta(y|\mathbf{x} = \boldsymbol{x}_t)$.

To estimate $p_\theta(y|\mathbf{x})$, we introduce a latent variable $\mathbf{z} \in \mathbb{R}^{L_z}$, as shown in Fig. 1. With $\mathbf{z}$, we can rewrite $p_\theta(y|\mathbf{x})$ as:

$$p_\theta(y|\mathbf{x}) = \int_{\mathbf{z}} p_\theta(y|\mathbf{z}) p_\theta(\mathbf{z}|\mathbf{x}) \mathrm{d}\mathbf{z} = \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})}\big[p_\theta(y|\mathbf{z})\big], \quad (1)$$

where $p_\theta(\mathbf{z}|\mathbf{x})$ denotes an encoder that maps an input sample to a distribution in the latent space and $p_\theta(y|\mathbf{z})$ denotes a classifier that is conditioned on the latent variable. The significance of this formulation is that we have introduced a latent representation symbolized by $\mathbf{z}$ in the class prediction task. This representation can be very useful in practice. For example, we can utilize additional unlabelled data (i.e. data that does not have labels for y) to learn a better encoder $p_\theta(\mathbf{z}|\mathbf{x})$ and/or we can reuse the same encoder for multiple tasks.

According to Eq. 1, we need to obtain both the encoder $p_\theta(\mathbf{z}|\mathbf{x})$ and the classification module $p_\theta(y|\mathbf{z})$ to evaluate the predictive distribution $p_\theta(y|\mathbf{x})$. In this paper, we derive the latent representation with a variational auto-encoder (VAE, [8, 9]) and implement the classifier via a multi-layer perceptron (MLP [17]) with a softmax activation at the last layer. Putting the two modules together, we compute the overall predictive distribution $p_\theta(y|\mathbf{x})$ via sampling, as an analytical form for the integration in Eq. 1 is not available. Specifically, we draw M samples of the latent features from the encoder, i.e. $z_m \sim p_\theta(\mathbf{z}|\boldsymbol{x}_i)$, and approximate the expectation by taking the average of these samples, i.e. $\mathbb{E}_{p_\theta(\mathbf{z}|\boldsymbol{x}_i)}\big[p_\theta(y|\mathbf{z})\big] = \frac{1}{M}\sum_{m=1}^M p_\theta(y|\boldsymbol{z}_m)$. The class prediction $\hat{y}$ is made by selecting the most probable label according to $p_\theta(y|\mathbf{x})$ and the confidence τ is reported by the corresponding probability, i.e. $\hat{y} = \arg\max_k p_\theta(y|\mathbf{x}),\ \tau = \max_k p(y = k|\boldsymbol{x}_i)$.

3.2. Optimizing model parameters

A VAE [8, 9] contains a probabilistic encoder $q_\theta(\mathbf{z}|\boldsymbol{x}_i)$ that maps an input sample $\boldsymbol{x}_i$ to the latent space, and a decoder $p_\phi(\mathbf{x}|\boldsymbol{z}_i)$ that generates an input sample from a latent feature $\boldsymbol{z}_i$. We make the further assumption that the encoder is parameterized by a diagonal Gaussian, i.e. $q_\theta(\mathbf{z}|\boldsymbol{x}_i) = \mathcal{N}(\mu_\theta(\boldsymbol{x}_i), \sigma_\theta^2(\boldsymbol{x}_i))$ and that latent representations for the samples are confined to a prior distribution $p(\mathbf{z})$, which is chosen to be a unit Gaussian distribution. To estimate the parameters for the encoder $p_\theta(\mathbf{z}|\mathbf{x})$, we maximize VAE's evidence lower bound (ELBO) objective $\mathcal{L}_{\text{ELBO}} = \frac{1}{N}\sum_i \mathcal{L}(\boldsymbol{x}_i)$, where $\mathcal{L}(\boldsymbol{x}_i)$ is defined as follows:

$$\mathcal{L}(\boldsymbol{x}_i) = \mathbb{E}_{q_\theta(\mathbf{z}|\boldsymbol{x}_i)}[\log p_\phi(\boldsymbol{x}_i|\mathbf{z})] - D_{\text{KL}}[q_\theta(\mathbf{z}|\boldsymbol{x}_i)\|p(\mathbf{z})]. \quad (2)$$

We train a VAE on the available training input samples, i.e. $\mathcal{D}_x = \{\mathbf{x}_i\}_{i=1}^N$. We then take the optimized VAE encoder as the encoder in Eq. 1, i.e. $p_\theta(\mathbf{z}|\mathbf{x}) = q_\theta(\mathbf{z}|\mathbf{x})$.

To optimize the parameters in the MLP, we maximize the log likelihood of the class prediction with respect to $\mathcal{D}_N$, i.e. $\mathcal{L}_{\text{LL}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{e}_{y_i} \log p_\theta(y|\boldsymbol{x}_i)$, where $\boldsymbol{e}_i$ denotes the standard base vector with 1 at i-index and 0 elsewhere.

The best classification performance would be achieved when both the VAE module and the MLP module are trained together, as the encoder is also optimized to give good class prediction. In practice, it is more useful to train the modules in separate stages, as the input and the label might not be available at the same time or the same input might be applied for a different classification task. To reflect the performance in reality, we first optimize the encoder's parameters by maximizing the ELBO objective and then optimize the MLP's parameters by maximizing the log likelihood objective.

3.3. Considering out-of-distribution data

One issue with the standard VAE formulation described in the previous section is that it does not handle out-of-distribution (OOD) data sensibly. Very often, the encoder projects OOD samples to sharp Gaussians at locations that correspond to specific in-distribution (ID) samples. This causes the latent classifier to make over-confident prediction on these samples. We address this issue by incorporating a hypothesis parameter H in the encoder module, as follows:

$$p(\mathbf{z}|\boldsymbol{x}_i) = \int_H p(\mathbf{z}|\boldsymbol{x}_i, H) p(H|\boldsymbol{x}_i) \mathrm{d}H \quad (3)$$

Given a data sample $\boldsymbol{x}_i$, H has two possible outcomes $\{H_\mathcal{D}, H_{\neg\mathcal{D}}\}$, where $H_\mathcal{D}$ denotes $\boldsymbol{x}_i$ is ID data, $H_{\neg\mathcal{D}}$ denotes the opposite and $p(H_\mathcal{D}|\boldsymbol{x}_i) + p(H_{\neg\mathcal{D}}|\boldsymbol{x}_i) = 1$.

During training, we assume that all data samples are ID, so $p(H_\mathcal{D}|\boldsymbol{x}_i) = 1$ and the encoder simplifies to the original VAE's encoder, i.e. $p(\mathbf{z}|\boldsymbol{x}_i, H_\mathcal{D}) = \mathcal{N}(\mu_\theta(\boldsymbol{x}_i), \sigma_\theta^2(\boldsymbol{x}_i))$. During inference, the input sample $\boldsymbol{x}_i$ can be either be OOD

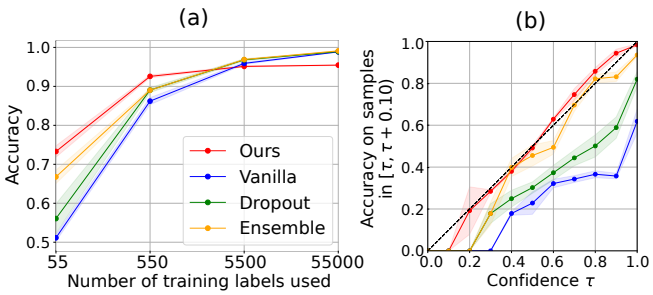

Fig. 2. **Label efficiency study**. Classifiers are trained with 55, 550, 5500 or 55K training labels available. (a) Test accuracy. (b) calibration plot for classifiers trained at 55 labels. Our method outperforms other approaches in accuracy and uncertainty calibration at limited label settings.

or ID. Therefore, we have to evaluate the encoder by considering both scenarios, according to the following equation:

$$p(\mathbf{z}|\boldsymbol{x}_i) = p(\mathbf{z}|\boldsymbol{x}_i, H_{\mathcal{D}})p(H_{\mathcal{D}}|\boldsymbol{x}_i) + p(\mathbf{z}|\boldsymbol{x}_i, H_{\neg\mathcal{D}})p(H_{\neg\mathcal{D}}|\boldsymbol{x}_i), \tag{4}$$

where $p(\mathbf{z}|\boldsymbol{x}_i, H_{\mathcal{D}})$ denotes the original VAE's encoding distribution. We define the other three probabilities as follows:

$$p(H_{\mathcal{D}}|\boldsymbol{x}_i) = \frac{1}{1 + \exp(a \cdot \text{ELBO}(\boldsymbol{x}_i) + b)}, \tag{5}$$

$$p(H_{\neg\mathcal{D}}|\boldsymbol{x}_i) = 1 - p(H_{\mathcal{D}}|\boldsymbol{x}_i), \tag{6}$$

$$p(\mathbf{z}|\boldsymbol{x}_i, H_{\neg\mathcal{D}}) = \mathcal{N}(0, \boldsymbol{\mu}^2_{\text{max}}), \tag{7}$$

where the hyperparameters a and b are chosen using the validation set so that $p(H_{\mathcal{D}}|\boldsymbol{x}_i)$ is 0.5 at -3 standard deviation away from the mean ELBO and 0.05 at the minimum ELBO; $\boldsymbol{\mu}_{\text{max}}$ denotes the maximum absolute value of the encoding mean across all latent dimensions in the validation set; ELBO($\boldsymbol{x}_i$) evaluates the ELBO objective for an input sample $\boldsymbol{x}_i$ according to Eq. 2. The above equations ensure that an input sample with a high ELBO value will be assigned a high probability for $p(H_{\mathcal{D}}|\boldsymbol{x}_i)$ and its encoding will be primarily governed by the VAE's encoder - sharp and localized in the latent space. On the other hand, an input sample with low ELBO value will be assigned a high probability for $p(H_{\neg\mathcal{D}}|\boldsymbol{x}_i)$ and its encoding will be dominated by an uninformative wide Gaussian. Such an encoder allows the latent classifier to behave more sensibly, as the predictive distribution (averaged over multiple samples from the latent distribution) will tend to spread out over different class labels for the OOD data.

4. EXPERIMENTS AND RESULTS

We evaluate our method on two datasets: 1) MNIST [18], which contains 60k/10k train/test images and corresponding labels of handwritten digits; 2) HAM10k [10], a medical dataset for pigmented skin lesion diagnosis, which contains

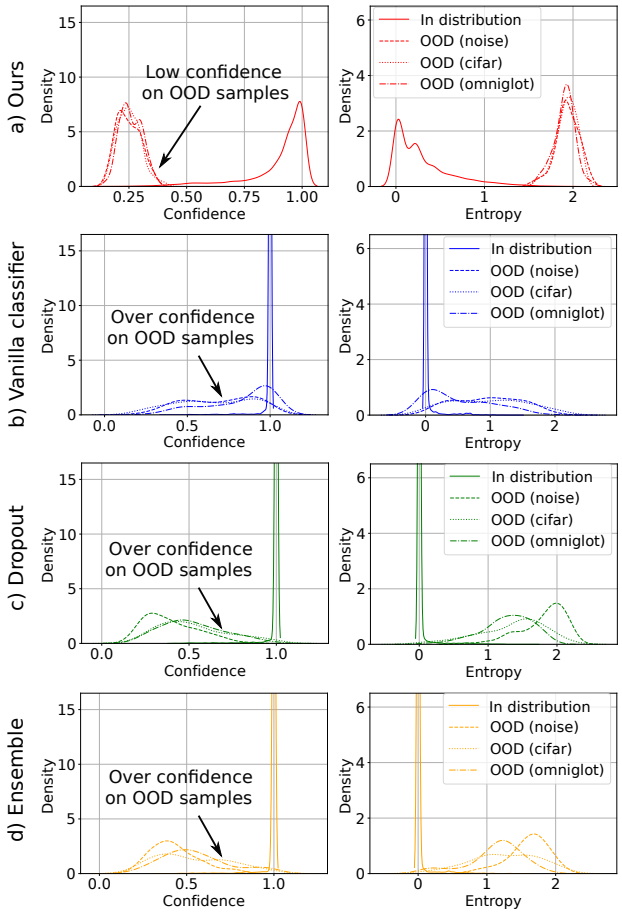

Fig. 3. **Uncertainty estimation for out-of-distribution (OOD) data**. Our method keeps the predicted confidence on OOD data low, whereas other approaches tend to make high confidence prediction on OOD data.

10015 dermatoscopic images and corresponding lesion labels. We compare our method against three baseline approaches: a standard neural network classifier (Vanilla [7]), a classifier trained with Monte-Carlo dropout at a rate 0.1 (Dropout [15]) and ensembles of $M = 3$ vanilla classifiers (Ensemble [6]). We use the same model architecture for all the approaches - i.e. LeNet [19] for MNIST and ResNet18 [20] with a 3-layer MLP for HAM10k. We evaluate two metrics - classification accuracy and calibration. The calibration is evaluated via a calibration plot, where the test samples are grouped into 10 buckets according to their confidence τ (defined in Sect. 3.1). The accuracy of each group is evaluated and plotted against the centre confidence of each bucket. A well calibrated classifier will generate line along the diagonal.

We first show that our method works better than other approaches when training labels are limited. We train different approaches with different amount of training labels available on the MNIST dataset. As shown in Fig. 2a, our method gives the best accuracy among all approaches when extremely lim-

Table 1. Classification accuracy on HAM10k test set.

	Ours	Vanilla	Dropout	Ensemble
Accuracy	74.9 ± 0.3	82.0 ± 1.1	82.3 ± 0.5	**84.0** ± 0.3

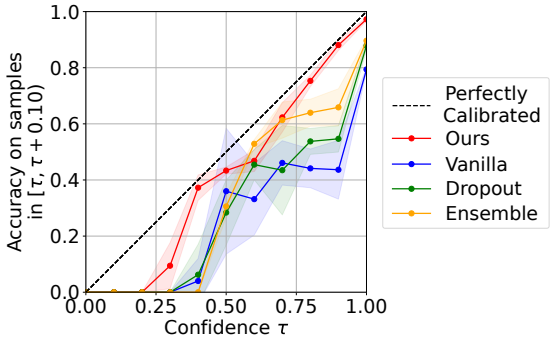

Fig. 4. **Calibration on skin lesion detection**. Our method gives the best calibrated uncertainty prediction.

ited training labels are available (e.g. 55 or 550). This indicates that the generative unit helps our method to derive better representation that compensates the limitation in the amount of the training labels available. Furthermore, our method offers well-calibrated uncertainty estimation that follows the diagonal almost perfectly in Fig. 2b, whereas the other approaches show various levels of over confidence as they are likely over-fitted to the limited training examples.

We now investigate how different approaches react to OOD data. We train all approaches on the MNIST training data and we plot the confidence and the entropy distribution of all the classifiers on the test set of MNIST (in-distribution) and the test set of three other datasets - random noise, CIFAR10 and OMNIGLOT (OOD). The results are shown in Fig. 3. All three baseline approaches make high confident prediction (confidence > 0.5 or entropy < 1.0) on a significant distribution of the OOD data. In contrast, our method keeps all the OOD data in the low confidence and high entropy range. This indicates our method do not make over-confident prediction on unfamiliar samples, which is a very desirable property with high-stakes applications such as healthcare.

We now evaluate our method on a medical dataset HAM10k, where the goal is to predict skin lesion type from a dermatoscopic image. We train all classifiers until the validation loss has converged. Prior training our latent classifier, we first train the VAE module for 200 epochs when the ELBO objective on the validation set has converged. The test accuracy on all classifiers is reported in Table 1, where our performance seems to lag behind other approaches with the ensemble classifier giving the best accuracy.

However, we see a different picture, when we investigate the calibration plot in Fig. 4. Although reporting a lower accuracy, our method is the best calibrated approach among all. This indicates that the majority of the wrong predictions made by our approach occur at the low confidence regime. As the

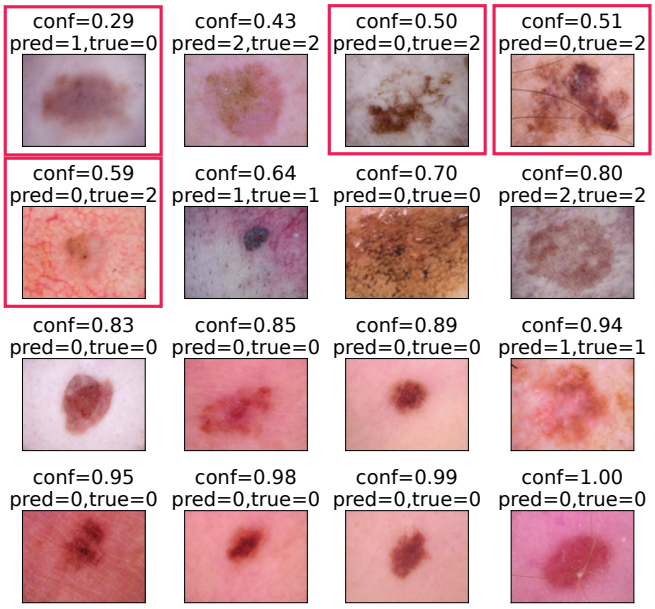

Fig. 5. **Skin lesion classification examples from our approach**. The prediction accuracy is well calibrated with the confidence, i.e. the wrong prediction (highlighted in red square) tends to take place at low confidence.

predicted confidence increases, our method becomes more accurate. The other approaches render a varying degree of overconfidence. Their mis-calibrated accuracy at high confidence prediction is worrying. For example, the accuracy for the ensemble classifier at the confidence bracket $[0.8, 0.9)$ is 65.5%. This indicates 34.5% of these high-confidence prediction is wrong. In applications such as cancer diagnosis, such overly confident wrong prediction can be detrimental.

We demonstrate a random batch of the lesion prediction made by our method in Fig. 5. The wrong prediction are highlighted with the red squares, which occurs more often at the low confidence cases. This example shows that our method can be potentially used as an assistive diagnosis tool for clinicians. The predictions with low confidence will be sent to a clinician for further investigation, whereas the predictions at high confidence only need to be randomly verified.

5. CONCLUSION

We propose a hybrid model that utilizes a VAE's representation learning capability to derive compact, low-dimensional features and a light-weight MLP classifier to make the class prediction. With proper statistical treatment, the probabilistic encoder in our method appropriately expresses its familiarity for an input sample in the spread of the latent distribution. This further allows our classifier to give better calibrated uncertainty estimation. We show the effectiveness of our approach in a challenging skin lesion diagnosis task.

6. REFERENCES

[1] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. 2012, vol. 25, Curran Associates, Inc.

[3] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2013.

[4] A. Philip Dawid, "The well-calibrated bayesian," *Journal of the American Statistical Association*, vol. 77, pp. 605–610, 1982.

[5] Morris H. Degroot and Stephen E. Fienberg, "The comparison and evaluation of forecasters.," *The Statistician*, vol. 32, pp. 12–22, 1983.

[6] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NIPS*, 2017.

[7] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations*, 2017.

[8] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[9] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 1278–1286.

[10] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 2018.

[11] Max Welling and Yee W Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.

[12] David J. C. Mackay, "Bayesian methods for adaptive models," 1992.

[13] Alex Graves, "Practical variational inference for neural networks," in *NIPS*, 2011.

[14] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra, "Weight uncertainty in neural networks," *ArXiv*, vol. abs/1505.05424, 2015.

[15] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *ArXiv*, vol. abs/1506.02142, 2016.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," *ArXiv*, vol. abs/1706.04599, 2017.

[17] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning internal representations by error propagation," 1986.

[18] Y. LeCun, "The mnist database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.

[19] Yann André LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," 1998.

[20] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.