On the performance of uncertainty estimation methods for deep-learning based image classification models

Luís Felipe P. Cattelan[1], **Danilo Silva**[1]

[1]Machine Learning and Applications Research Group (GAMA)
Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brazil

luisfelipe1998@gmail.com, danilo.silva@ufsc.br

Abstract. *Previous works have shown that modern neural networks tend to be overconfident; thus, for deep learning models to be trusted and adopted in critical applications, reliable uncertainty estimation (UE) is essential. However, many questions are still open regarding how to fairly compare UE methods. This work focuses on the task of selective classification and proposes a methodology where the predictions of the underlying model are kept fixed and only the UE method is allowed to vary. Experiments are performed for convolutional neural networks using Deep Ensembles and Monte Carlo Dropout. Surprisingly, our results show that the conventional softmax response can outperform most other UE methods for a large part of the risk-coverage curve.*

1. Introduction

In recent years, deep learning (DL) has consistently shown great success in predictive tasks in many different application areas. However, a major challenge in the adoption of DL models is not only that they are difficult to interpret ("black-box") but also that they tend to be overconfident even when they produce incorrect predictions, thus failing silently [Guo et al. 2017]. For DL models to be trusted and effectively used in critical applications, such as autonomous driving and medical diagnosis, it is essential that these models "know when they don't know", i.e., that they produce a reliable estimate of the uncertainty associated with their predictions [Ovadia et al. 2019, Abdar et al. 2021].

While many techniques have been proposed to estimate the uncertainty of a neural network [Ståhl et al. 2020, Gawlikowski et al. 2021, Manivannan 2020], there appears not be a consensus in the literature on how these techniques should be evaluated. With respect to classification models, the existing evaluation metrics for uncertainty estimation (UE) can be broadly arranged in three categories [Ding et al. 2020, Abdar et al. 2021]: probability calibration [Guo et al. 2017], which measures the degree to which a model's confidence in a prediction matches the empirical accuracy of that prediction; ability to detect out-of-distribution (OOD) samples; and selective classification (or classification with abstention) [Thulasidasan 2020, Geifman and El-Yaniv 2017], which refers to the performance of a classifier after a number of uncertain samples have been discarded. The suitability of any of these goals (which are often conflicting) is heavily dependent on the application; moreover, even within some category there are multiple proposed metrics each with their advantages and disadvantages [Gawlikowski et al. 2021, Galil et al. 2022, Ding et al. 2020]. An additional difficulty with the evaluation of OOD detection is that the definition of a representative OOD set is entirely subjective.

A good example of the current state of the art is the recent paper[Galil et al. 2022], which performs a comprehensive evaluation of deep uncertainty estimation under many different metrics (in all three categories) for a wide range of model architectures. In the case of selective classification, the paper argues that any scalar metric fails to completely summarize the UE performance of a model and instead advocates for the use of the risk-coverage (RC) curve, since the best performing model (in terms of selective accuracy) may depend on the operating point (the desired coverage).

However, the focus of [Galil et al. 2022] is on comparing different models (each using some arbitrary UE method), rather than on comparing different UE methods for the same model. This begs the question of whether the difference in, e.g., selective accuracy, is due to the intrinsic ability of a model to provide a reliable UE or simply due to the baseline accuracy of the model under full coverage—and, more generally, how to disentangle these two effects. Moreover, it also leaves open the question of which is the best UE method for a specific model.

A challenge in answering this last question is that the choice of an UE method often affects the underlying predictive model, since not all UE methods are directly compatible with all models in a plug-and-play fashion. This is clear for UE methods that require retraining the predictive model. Another, more subtle example is the use of ensemble methods for UE, which require the use of an ensemble model in the first place.

In this paper, we approach the aforementioned questions with the following methodology. We fix the predictive model and only allow the UE method to vary. For a fair comparison, we consider that a model is fixed when its predictions on a test set are fixed. In this initial study, we restrict attention to UE methods that can be computed deterministically from any of the model's ouputs or intermediate signals, without depending on the available data. This includes the conventional softmax response (also known as maximum class probability, MCP) and the entropy of the softmax output, for all models, as well as mutual information, predictive variance and mean variance, computed from the multiple realizations of the model's softmax output, in the case of ensemble models—but excludes any method that trains an auxiliary model to provide UE, such as [Geifman and El-Yaniv 2019, DeVries and Taylor 2018, Barnes and Barnes 2021, Corbière et al. 2021]. We focus on the task of selective classification [Geifman and El-Yaniv 2017] and, following [Galil et al. 2022], evaluate all methods using the RC curve.

We apply this methodology for CIFAR-10 and CIFAR-100 datasets and WideResnet 28-10 and ResNet-50 models. Suprisingly, our results show that the conventional softmax response has equal or better performance than most other UE methods for a large part of the RC curve. These results suggest that, for selective classification, ensemble methods do not always provide significantly better UE but simply a better predictive model.

The contributions of this paper are:

- We present a comparison between different UE methods for several types of models, including ensemble models (i.e., Deep Ensemble and Monte Carlo Dropout), using a selective classification metric and an evaluation methodology where we assume that the underlying model is kept fixed;
- Our results suggest that, for ensemble models, the softmax response applied to the

combined ensemble prediction is sufficient to achieve the same or better selective classification performance than more sophisticated UE methods that make use of the individual ensemble predictions.

2. Related Work

A first step to standardize the comparison between techniques in the literature is to compare them in the same context. Thus, [Nado et al. 2021] brings a collection of models and datasets baselines for the comparison between uncertainty estimation techniques. [Manivannan 2020] makes a more general comparison, comparing different ensemble methods and different uncertainty estimators from the fixed ensemble models. The difference between our work and theirs is that they focus on metrics that rely on the total number of misclassified samples detected and in the number of uncertain samples in a test set, while we focus on a selective classification metric. A disadvantage of their metric is that it does not take into account the initial performance of the model when no samples are discarded, leading to a potentially unfair comparison.

In [Galil et al. 2022] and [Ding et al. 2020], the applications and qualities of different metrics for uncertainty estimators from the literature are discussed. Also, different models are compared regarding these metrics, seeking conclusions as to which neural network architecture provides more information about uncertainties. Although the conclusions have great value, both works do not focus in comparing which is the best uncertainty estimator for a given model, bringing only the comparison between different models assuming it is the best quantification of confidence.

[Lakshminarayanan et al. 2017] proposes the deep ensemble method as an uncertainty estimator method and compare it to other methods. However, we claim that these comparisons are not fair, since the underlying model generates different predictions and performance. Indeed, the authors use as a metric the accuracy for different thresholds. The disadvantage of this metric is that it ignores the number of samples evaluated when the samples are filtered from the threshold and can be easily modified using post-hoc analysis, following a similar logic as presented by [Wang et al. 2021]. In this paper, the authors go beyond the investigation whether a model is calibrated, and instead propose to investigate if a model is *calibratable*. Applying the same reasoning in the context of selective classification, it is preferable to use a metric that does not rely on specific thresholds and is suitable for comparing different UE methods, which is the case of the risk-coverage curve adopted in the present paper.

3. Preliminaries

3.1. Selective classification

Let $S = \{(x_i, y_i)\}_{i=1}^N \subseteq (\mathcal{X} \times \mathcal{Y})^N$ be a set of labeled samples, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. A *selective model* [Geifman and El-Yaniv 2017] is a pair (h, g), where $h : \mathcal{X} \to \mathcal{Y}$ is a predictive model (such as a neural network) and $g : \mathcal{X} \to \mathbb{R}$ is a function that quantifies the model's confidence on its prediction for a given sample. For some specified threshold τ, the output of the selective model is given by

$$(h, g)(x) = \begin{cases} h(x) & \text{if} \quad g(x) \geq \tau \\ \text{abstain} & \text{if} \quad g(x) < \tau. \end{cases} \tag{1}$$

Thus, the model abstains from prediction on a sample x when its confidence $g(x)$ is low (or, equivalently, when its *uncertainty* $-g(x)$ is high). The fraction of samples for which a prediction is made, $c = (1/N)\sum_{i=1}^{N} \mathbb{1}[g(x_i) \geq \tau]$, where $\mathbb{1}[\cdot]$ is the indicator function, is called the *coverage* of the selective model.

For a selective model, one can define the *selective risk* [Geifman et al. 2018, Geifman and El-Yaniv 2019]

$$\hat{r}(h, g|S) = \frac{\sum_{i=1}^{N} \ell(h(x_i), y_i)\mathbb{1}[g(x_i) \geq \tau]}{\sum_{i=1}^{N} \mathbb{1}[g(x_i) \geq \tau]} \tag{2}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a given loss function. In what follows we assume the 0/1 loss, $\ell(\hat{y}, y) = \mathbb{1}[\hat{y} \neq y]$, which implies that $1 - \hat{r}$ is the selective accuracy of the model, i.e., the accuracy among the non-rejected samples.

Note that, by varying τ, it is generally possible to trade coverage for selective risk, i.e., a selective model can achieve a lower selective risk by increasing τ and thus rejecting more samples. Thus, for a fair comparison between selective models, it is necessary to consider both metrics $\hat{r}$ and c. A plot of $\hat{r}$ as a function of c is called the *risk-coverage* (RC) curve, which is commonly used to evaluate selective models.

Without loss of generality, we assume that $\mathcal{Y} = \{1, \ldots, C\}$, where C is the number of classes, and

$$h(x) = \arg \max_{j \in \{1,\ldots,C\}} (f(x))_j \tag{3}$$

where $f : \mathcal{X} \to [0, 1]^C$ returns a length-C vector. If $f(x)$ sums to 1 (such as when a softmax output is used), then $(f(x))_j$ can be interpreted as the model's estimated probability that sample x belongs to class j.

3.2. Ensemble methods

An ensemble is a combination of different models to create a more powerful one. Ensemble techniques are well known to increase machine learning performance [Dietterich 2000]. While many approaches to ensemble learning exist, in this paper we focus on the two most widely-used in the context of UE for deep learning: deep ensembles and Monte Carlo dropout. Both belong to the broad class of randomization-based approaches, where the ensemble components can be trained independently in parallel.

In the Deep Ensemble method [Lakshminarayanan et al. 2017], the same deep neural network is trained independently T times with different random initial values and a differently shuffling of the training data, resulting in models $f^1, \ldots, f^T$. The final ensemble model is obtained by simple averaging as

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} f^t(x). \tag{4}$$

Note that this method can be applied to any model architecture without changes; however, it requires training and storing T different models.

In the Monte Carlo Dropout (MCD) approach [Gal and Ghahramani 2016], a deep neural network with dropout layers is required. The dropout layers are made active not only during training but also during inference. In this way, a single network is trained but multiple inferences (with random dropped units) can be made. If T inferences (also called stochastic passes in this context) are made for a sample x, resulting in the predictions $f^1(x), \ldots, f^T(x)$, then the final output $f(x)$ is again obtained by simple averaging as (4). An advantage of MCD is that a single training is required and a single model needs to be stored; however, it requires architectural changes to add dropout layers if the network does not already have them.

3.3. Uncertainty Estimation Methods

3.3.1. Softmax response

Consider a neural network using a softmax activation function at the output layer. The *softmax response* (SR), as known as the *maximum class probability* (MCP), is given by

$$g_{\mathrm{SR}}(x) = \max_{j \in \{1,\ldots,C\}} (f(x))_j \tag{5}$$

The SR is the most natural UE method and is often used as a baseline [Hendrycks and Gimpel 2017].

3.3.2. Entropy

Another common technique is taking the entropy of the predictive distribution as the uncertainty parameter (thus the negative entropy as the confidence parameter):

$$g_E(x) = -H(f(x)) = \sum_{j=1}^{C} (f(x))_j \log(f(x))_j \tag{6}$$

3.3.3. Ensemble approaches

When using an ensemble model, the divergence between its multiple predictions can be used as an uncertainty parameter [Nair et al. 2020]. The most common ones are mutual information, softmax variance and predictive variance [Smith and Gal 2018]. Mutual information (MI) measures the difference between the average entropy of the individual predictive distributions and the entropy of the average predictive distribution:

$$g_I(x) = -I = \sum_{j=1}^{C} \left(\frac{1}{T} \sum_{t=1}^{T} \left((f^t(x))_j \log(f^t(x))_j \right) - (f(x))_j \log(f(x))_j \right) \tag{7}$$

where $f(x)$ is the predictive distribution of the ensemble after averaging, given by (4).

Softmax variance (SV) measures the average of the variance in the probability estimates for each class:

$$g_{SV}(x) = -\sigma_{SV}^2 = -\frac{1}{C} \sum_{j=1}^{C} \frac{1}{T} \sum_{t=1}^{T} \left((f^t(x))_j - (f(x))_j \right)^2 \tag{8}$$

Predictive variance (PV) measures the variance in the probability estimates of only the predicted class:

$$g_{PV}(x) = -\sigma^2_{PV} = -\frac{1}{T}\sum_{t=1}^{T}\left((f^t(x))_j - (f(x))_j\right)^2 |_{j=h(x)} \tag{9}$$

4. Methodology

The basic methodology of this paper is the following: for every inference technique in section 4.2, all applicable UE methods in section 3.3 are compared. This is done for each dataset and model architecture considered, as described below. In this manner, the predictions made by a given model on the test set will remain *fixed*, while only their confidence *ranking* (given by g) is allowed to vary by the choice of the UE method.

All codes, implementation and analysis can be found in Github/lfpc/Uncertainty_Estimation repository. All implementations were made using PyTorch [Paszke et al. 2019].

4.1. Datasets

The datasets used in the experiments are the CIFAR-10 and CIFAR-100 datasets, due to the simplicity and research relevance of these. During training, *random cropping* and *horizontal flip* data augmentations are applied.

4.2. Models

The architectures used are the WideResnet [Zagoruyko and Komodakis 2016] 28-10 and ResNet-50 [He et al. 2016], trained from scratch.

With these network architectures and for each dataset used, the following models are trained:

- A single model (no ensemble) as a baseline, which is referred to as a deterministic model;
- A Deep Ensemble with $T = 4$;
- A Monte Carlo Dropout ensemble with $T = 10$.

Note that the above choices relate only to how the main prediction of the model, $f(x)$, is computed. The ensemble approaches also produce auxiliary predictions $f^1(x), \ldots, f^T(x)$, which may be accessed by some UE methods.

Note also that there is no problem in considering ensemble approaches with different values of T, since these approaches are not directly compared. The values of T chosen are typical values used in practice, reflecting the fact that Deep Ensemble normally outperforms MCD for the same T at the expense of a much higher computational burden for training [Lakshminarayanan et al. 2017].

4.3. Metrics

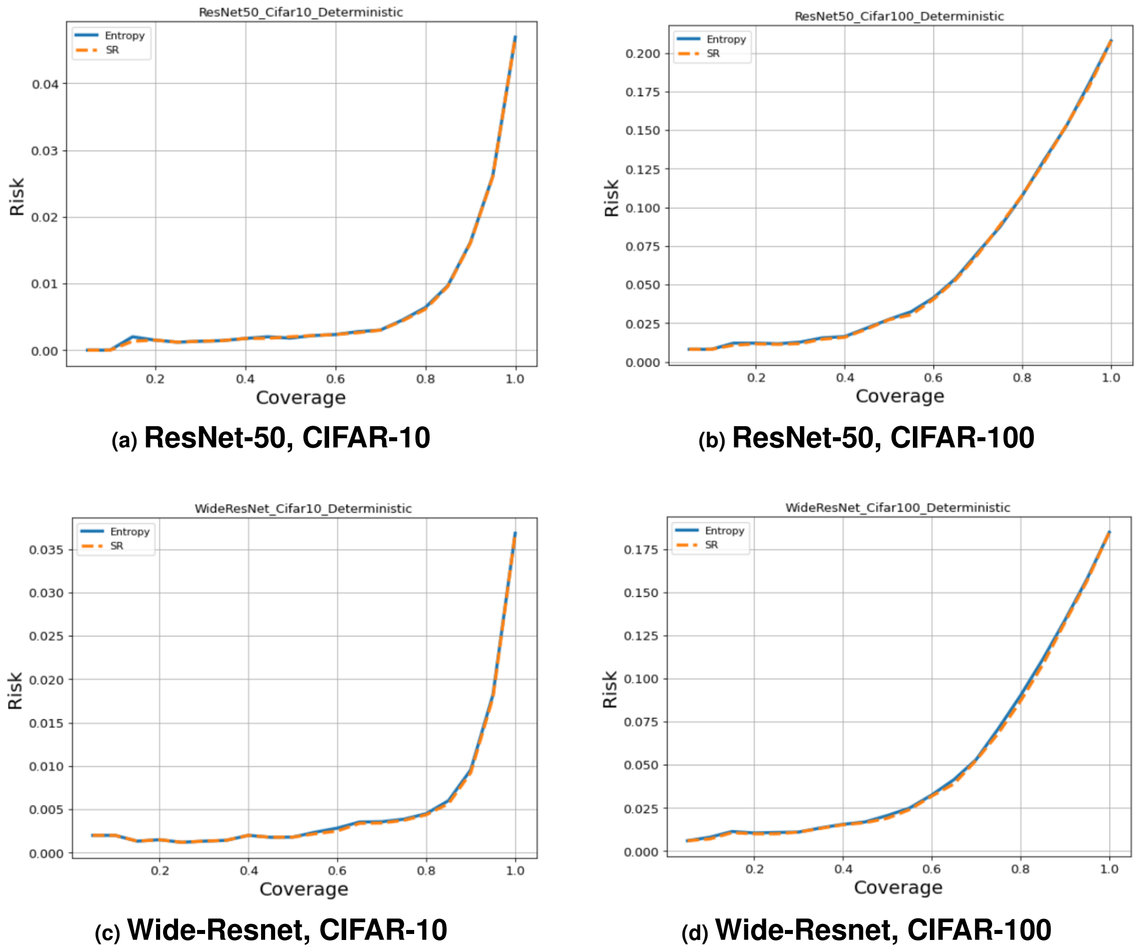
As already discussed and justified, in order to make the comparison between uncertainty estimators application-agnostic and to avoid total accuracy bias, we will use only the **RC curve (risk per coverage)** [Geifman et al. 2018, Geifman and El-Yaniv 2017], with the risk being the expected 0/1 loss (classification error). The curve is plotted defining the threshold which implies in the analysed coverage on the test set, and the risk is calculated with equation (2).

5. Results

5.1. Deterministic model

As discussed, the deterministic model has its SR as a baseline [Hendrycks and Gimpel 2017]. The figure 1 shows the RC curve in this case. It can be seen that the SR and the entropy of the deterministic model is effective in removing samples, since the accuracy keeps increasing with the coverage.

Moreover, entropy and SR shows similar performance, with a slight advantage for SR.



(a) **ResNet-50, CIFAR-10**

(b) **ResNet-50, CIFAR-100**

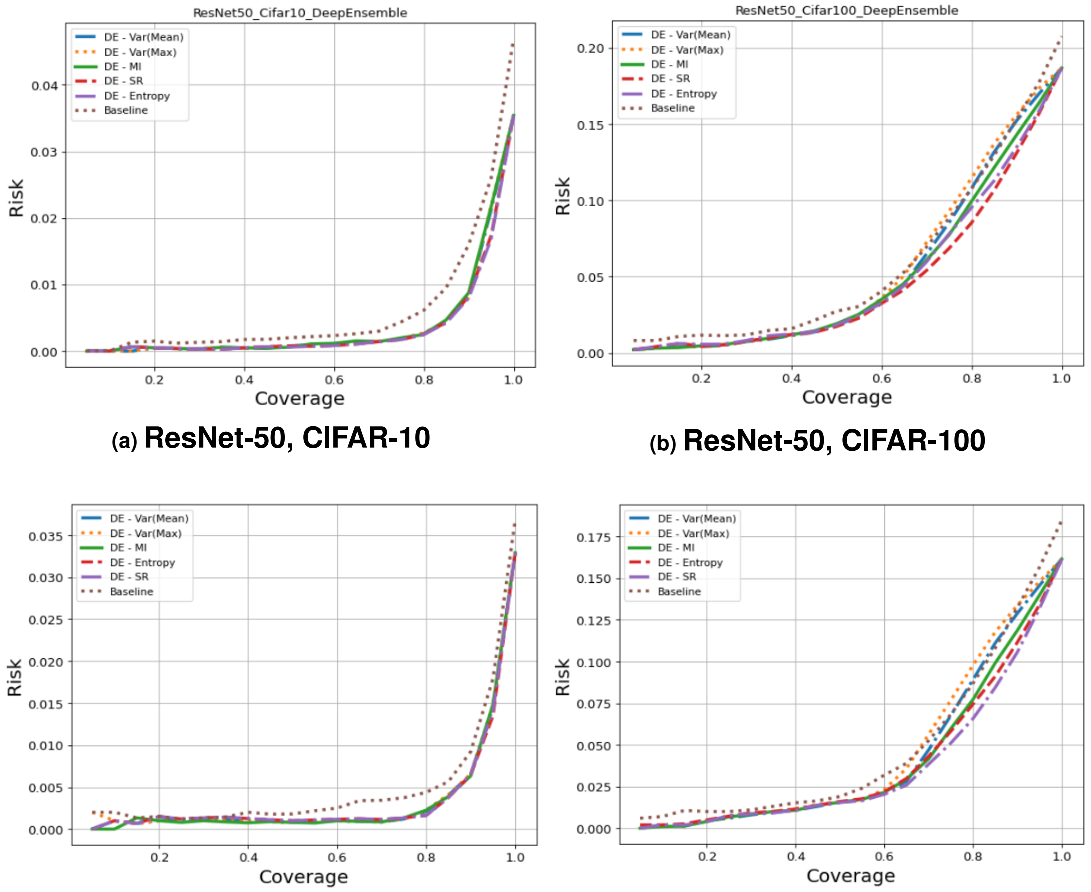
(c) **Wide-Resnet, CIFAR-10**

(d) **Wide-Resnet, CIFAR-100**

Figure 1. 0/1 loss x coverage for deterministic models - comparison between MCP and Entropy

5.2. Deep Ensemble

Fig. 2 presents comparisons for UE methods for the Deep Ensemble model. For CIFAR-10, for which the model has very high accuracy, it can be seen that all divergence methods show similar performance. For CIFAR-100, Softmax Response shows to be the most appropriate one. Baseline curve represents deterministic model with SR as uncertainty estimator.



(a) **ResNet-50, CIFAR-10**

(b) **ResNet-50, CIFAR-100**

(c) **Wide-Resnet, CIFAR-10**

(d) **Wide-Resnet, CIFAR-100**

Figure 2. Classification error per coverage using Deep Ensemble

5.3. Monte Carlo Dropout

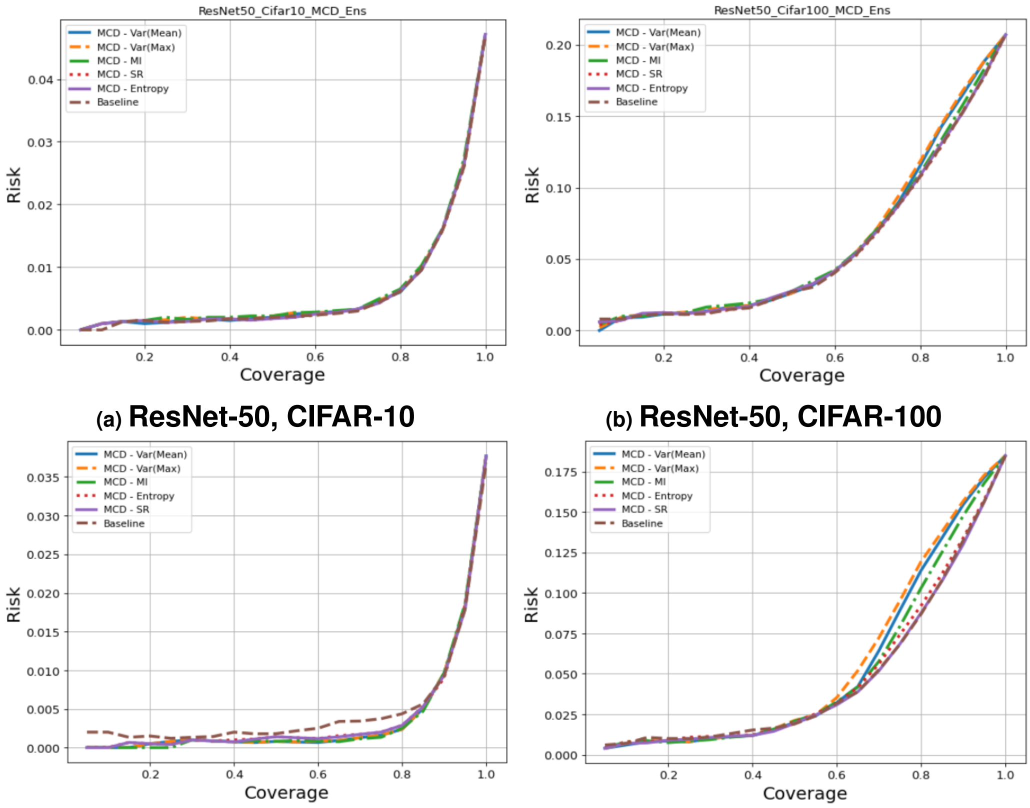
Figure 3 shows the RC curves for MCD. The patterns are similar to Deep Ensemble ones—SR is similar or better than divergence methods.

When compared to the baseline curve, it can be seen that the accuracy difference when using the MCD ensemble is not really significant. However, the new model has better results when removing wrong samples, especially for the Wide-ResNet.

6. Conclusions

In this paper, we compared different approaches for quantifying uncertainty in image classification models, using an evaluation metric where we assume that the underlying model is kept fixed. Ensemble methods are interpreted as the construction of a whole new model (possibly with different predictions), and not a pure uncertainty estimation technique. Also, the comparisons are performed in the risk-coverage domain, since it takes into account the initial model performance and is suitable for comparing across different uncertainty estimators, besides having a directly useful interpretation.

Although using the same architecture, ensemble methods change the base model—when applied, it possibly entails different accuracy and higher inference time. Thus, we have argued that it is not totally fair to compare UE performance for different



(a) **ResNet-50, CIFAR-10**

(b) **ResNet-50, CIFAR-100**

(c) **Wide-Resnet, CIFAR-10**

(d) **Wide-Resnet, CIFAR-100**

Figure 3. Classification error per coverage using Monte Carlo Dropout

models, as the predictions are different—for UE performance to be analyzed in a fair way, the predictions (or, at least, the accuracy at full coverage) should be the same. In particular, the high performance of Deep Ensembles at selective classification is partly explained by the fact that it is a better predictive model at full coverage.

In the context of selective classification, we compared several methods for evaluating uncertainty given ensemble predictions (especially, Deep Ensemble and Monte Carlo Dropout)—from the conventional softmax response to methods that exploit the divergence between multiple predictions.

When the predictions of the ensemble model are kept fixed, the SR response shows the best or close to best performance compared to other UE methods for a large part of the RC curve. This implies, in particular, that the Bayesian interpretation of Monte Carlo Dropout—from which one would naturally compute the UE as the predictive variance—is not fundamental for selective classification. Rather, MCD can be viewed simply as a type of ensemble that provides more reliable probability estimates and so is useful for selective classification.

For future work, we plan to expand our results by performing experiments with more models and datasets.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76(C):243–297.

Barnes, E. A. and Barnes, R. J. (2021). Controlled abstention neural networks for identifying skillful predictions for classification problems. *Journal of Advances in Modeling Earth Systems*, 13(12). e2021MS002573 2021MS002573.

Corbière, C., Thome, N., Saporta, A., Vu, T.-H., Cord, M., and Perez, P. (2021). Confidence Estimation via Auxiliary Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

DeVries, T. and Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Dietterich, T. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems: First International Workshop, MCS 2000, Lecture Notes in Computer Science*, pages 1–15.

Ding, Y., Liu, J., Xiong, J., and Shi, Y. (2020). Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 22–31.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org.

Galil, I., Dabbah, M., and El-Yaniv, R. (2022). Which models are innately best at uncertainty estimation?

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2021). A survey of uncertainty in deep neural networks.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Geifman, Y. and El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.

Geifman, Y., Uziel, G., and El-Yaniv, R. (2018). Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. *arXiv e-prints*, page arXiv:1805.08206.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th In-*

ternational Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Manivannan, I. (2020). A comparative study of uncertainty estimation methods in deep learning based classiAcation models. Technical report, Hochschule Bonn-Rhein-Sieg Ű University of Applied Sciences, Department of Computer Science.

Nado, Z., Band, N., Collier, M., Djolonga, J., Dusenberry, M. W., Farquhar, S., Filos, A., Havasi, M., Jenatton, R., Jerfel, G., Liu, J., Mariet, Z., Nixon, J., Padhy, S., Ren, J., Rudner, T. G. J., Wen, Y., Wenzel, F., Murphy, K., Sculley, D., Lakshminarayanan, B., Snoek, J., Gal, Y., and Tran, D. (2021). Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *CoRR*, abs/2106.04015.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. In Globerson, A. and Silva, R., editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569. AUAI Press.

Ståhl, N., Falkman, G., Karlsson, A., and Mathiason, G. (2020). Evaluation of uncertainty quantification in deep learning. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 556–568. Springer.

Thulasidasan, S. (2020). *Deep Learning with Abstention: Algorithms for Robust Training and Predictive Uncertainty*. PhD thesis, University of Washington.

Wang, D.-B., Feng, L., and Zhang, M.-L. (2021). Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In Richard C. Wilson, E. R. H. and Smith, W. A. P., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press.