

---

# Multi-Loss Sub-Ensembles for Accurate Classification with Uncertainty Estimation

---

**Omer Achrack**

Intel (EGI)

omer.achrack@intel.com

**Raizy Kellerman**

Intel (Mobileye)

raizy.kellerman@intel.com

**Ouriel Barzilay**

DermaDetect Ltd. / Intel (Mobileye)

ouriel@dermadetect.com

## Abstract

Deep neural networks (DNNs) have made a revolution in numerous fields during the last decade. However, in tasks with high safety requirements, such as medical or autonomous driving applications, providing an assessment of the model's reliability can be vital. Uncertainty estimation for DNNs has been addressed using Bayesian methods, providing mathematically founded models for reliability assessment. These models are computationally expensive and generally impractical for many real-time use cases. Recently, non-Bayesian methods were proposed to tackle uncertainty estimation more efficiently. We propose an efficient method for uncertainty estimation in DNNs achieving high accuracy. We simulate the notion of multi-task learning on single-task problems by producing parallel predictions from similar models differing by their loss. This multi-loss approach allows one-phase training for single-task learning with uncertainty estimation. We keep our inference time relatively low by leveraging the advantage proposed by the Deep Sub-Ensembles method. The novelty of this work resides in the proposed variational inference with a simple and convenient training procedure, while remaining competitive in terms of computational time. We conduct experiments on SVHN, CIFAR10, CIFAR100 as well as ImageNet using different architectures. Our results show improved accuracy on the classification task and competitive results on several uncertainty measures.

## 1 Introduction

Deep neural networks have achieved impressive results in various A.I. areas such as computer vision, natural language processing and audio analysis. The common approach for classification in supervised learning is to employ softmax activation on the logits vector. While being efficient and intuitive, it lacks the ability to provide any kind of uncertainty measure that would yield an estimation of what the model does not know (e.g. a lack of expertise on a given input) [22, 36, 29]. These models generally tend to be over-confident as mentioned in [36, 34]. When applied to life-critical tasks, such as autonomous driving and medical diagnosis, the results might be devastating. In such use cases, the model should have the essential ability to determine what it knows and what it does not. Ideally, this ability should be reflected by a proper confidence measure associated to the model prediction, likely to prevent rare, but costly events of over-confident mispredictions. Bayesian techniques [31, 32, 18] offer a theoretically grounded approach to infer uncertainty from a model, but their computational resources consumption often make them impractical. Several non-Bayesian methods have been proposed for providing uncertainty measures [13, 28]. While being more computationally attractive

than Bayesian methods, these methods still require multiple forward passes at inference time to obtain a reliable uncertainty measure. Recently, the Deep Sub-Ensembles (DSE) [46] method proposes a simplification of the Deep Ensembles (DE) [28] method by sharing a common network trunk and composing different heads, called *sub-ensembles*.

We propose an enhancement to the DSE method by associating a different loss function to each sub-ensemble. As a result, training can be performed in an end-to-end manner, instead of separately for each network branch, saving computational resources. Furthermore, our method improves the model’s accuracy, and offers competitive uncertainty estimation results, considering the attractive run time and convenient training procedure, as shown in Section 4.

We organize the rest of the paper as follows: Section 2 briefly overviews the recent methods addressing model confidence. Section 3 describes the details of the proposed method. In Section 4, we show our experimental setup and results, which are discussed in Section 5. We conclude and propose directions for future work in Section 6.

## 2 Related work

**Types of uncertainty** In Bayesian learning, there are two main types of uncertainty one can model [11]. Aleatoric uncertainty (from Greek *alea*, “rolling a dice”), is a **data property** capturing the noise inherent to observations and thus, can’t be reduced even if more data is collected. On the other hand, Epistemic uncertainty (from Greek *episteme*, “knowledge”) is a **model property** and can be reduced if more data is collected. This uncertainty measure stems from the training process: formally, it estimates a posterior distribution over the weights  $w$  given the data set  $\mathcal{D}$ , i.e.  $\mathcal{P}(w|\mathcal{D})$ , in contrast to  $\mathcal{P}(y|x, w)$  where  $x$  and  $y$  are respectively the network’s input and output. The epistemic uncertainty measure could also be interpreted as the result of an insufficiency in the training data at certain regions of the input space, [22], affecting in turn prediction reliability. In this work, we address epistemic uncertainty estimation.

**Non-Bayesian methods for uncertainty estimation** Many recent studies seek to provide uncertainty estimation using non-Bayesian methods. MC-dropout [13] approximates at inference time the posterior distribution of the weights by incorporating a dropout [19, 14, 6] mechanism. By performing the forward pass of the network several times, the method generates different predictions induced by randomized dropouts on the activations, to model the epistemic uncertainty. In Deep Ensembles [28], several differently-initialized models are trained with random shuffling of the training data to obtain model variations given a training set. In this way, each model is treated as a sample from  $\mathcal{P}(w|\mathcal{D})$ . At inference time, the method averages probability prediction vectors over the different models, and quantifies the uncertainty by means of the averaged probability’s entropy. While outperforming MC-dropout [46], DE is expensive in terms of memory and compute resources for both training and inference. Tagasovska et al. [44] propose a single model estimating both types of uncertainty, using quantile regression for aleatoric uncertainty and out-of-distribution detection, by using a set of different linear classifiers for epistemic uncertainty. In [1], the authors propose to model the deep feature space at different intermediate layers of the network using multivariate Gaussian or Gaussian Mixture Models (GMM) for enabling the model to detect out-of-distribution samples, via Mahalanobis distance [5]. This work, by performing a separate modeling for each class, extends [30] where the feature distributions are estimated for the entire data set as a whole. The post-training modeling required by these methods is relatively expensive since it involves computing multivariate Gaussians or GMMs. Jiang et al. [21] train KNN classifiers using the intermediate feature space for each class at selected layers. The KNN scores are then used to determine the prediction’s reliability. Kendall et al. [22] combine the modeling of both uncertainty measures: the aleatoric uncertainty estimator is trained by a dedicated branch of the network, while the epistemic uncertainty is modeled using MC-dropout. Another recent approach [42] models the posterior probability of the predicted class by directly learning the alpha coefficients of the Dirichlet distribution [40]. The uncertainty was quantified as the ratio of the number of classes to the alpha coefficients sum. Corbiere et al. [9] train a network aiming at predicting the confidence, based, in an innovative manner, on the true class probability (TCP), rather than on the maximum class probability like softmax. Their network includes a separate branch trained to regress the TCP with  $L2$  norm as a dedicated loss function for that branch. In the widely-used Temperature Scaling (TS) [34] method, a TS parameter is used to normalize the final logits vector for control on the “peakiness” of the model’s feature vector before the final softmax activation layer. The parameter can either be manually tuned using a dedicated

validation set, or being part of the predicted values as in [36]. According to [29] combining TS [34] with MC-dropout, the TS method significantly reduces the expected calibration error (ECE) on several data sets. Post-Hoc Calibration [26] aims at learning a re-calibration matrix for reducing the ECE on the predictions produced by the softmax activation [27]. This is done under the assumption that the re-calibration matrix coefficients are sampled from a Dirichlet distribution modeled by a fully connected layer.

**Bayesian methods for uncertainty estimation** The method proposed in [3] models the epistemic uncertainty by directly learning  $\mathcal{P}(w|\mathcal{D})$ . At inference time, the model parameters are sampled from the learned posterior distribution to produce multiple predictions and perform variational inference. An extension of this method by [45] provides an open-source implementation with additional Bayesian RNN and LSTM layers. However, these approaches use the relaxed assumption of diagonal-dominant covariance matrices for the posterior distribution of the weights in each layer. Another innovative approach [12] proposes to perturb Adam [23] gradients during training and show improvements on the uncertainty estimation performed with MC-dropout. The Stochastic Weight Averaging (SWA) method [20] proposes to train a model from scratch until convergence, and in a next step, to train it further for a smaller number of epochs. At each of those additional epochs, a snapshot of the model weights is collected. These sets of weights are then averaged to produce a model approximating Deep Ensembles. While significantly improving run-time requirements, this method achieves impressive results on several data sets, as mentioned by [33], which recently extended SWA by approximating, from the learned models, the posterior distribution over the weights. The authors also relax the diagonal-dominant assumption on the covariance matrix and use instead low rank covariance matrices for estimating the posterior distributions over the weights of the whole network.

### 3 Method

**Deep Ensembles and Deep Sub-Ensembles** In Deep Ensembles [28],  $M$  models are separately trained on a data set  $D = \{x_i, y_i\}$  via Maximum Likelihood Estimation (MLE) optimization, implemented as Stochastic Gradient Descent (SGD) or any of its variants [39]. Once the training of the different models has been performed, each classifier  $i$  provides a prediction vector  $\hat{y}_i$  representing the probability distribution over the defined classes  $\{1 \dots C\}$ . These  $M$  distributions are eventually averaged to compose the final prediction  $\hat{y}$  (see Equation 1 and e.g. [29, 46]).

$$\hat{y} = \frac{1}{M} \sum_{i=1}^M \hat{y}_i \quad (1)$$

Alternatively, in DSE [46], the network shares a trunk and sequentially train each of the sub-ensembles. Since our approach is similar to the DSE method, we depict the difference in the network architectures of DE and DSE in Figure 1.

**Multiple views on the same task** We propose to train the whole network from scratch in a single phase (in contrast to DE or SWA) and in an end-to-end manner by combining the advantages of the paradigms of variational inference (VI) and multi-task learning. In multi-task learning [38, 17], a single network, with shared layers and a dedicated head for each task, is trained in an end-to-end manner to perform simultaneously different tasks on the same input data. This usually improves the results on *each task* separately. Inspired by this notion, we propose to train, on a *single task*, a network with DSE-like architecture where each of the cloned heads is optimized using a different loss function. The rationale behind that lays on the conjecture that each head could benefit from the shared knowledge obtained from other "experts" in the same field. A comparison between the training procedures of our approach and DSE [46], is demonstrated in Figure 2. We next review the different loss functions that were utilized for assessing the proposed approach. Sensoy et al. [42] propose a loss allowing quantification of the classification uncertainty using Dirichlet distribution. Relaxed softmax [36] predicts an additional parameter used for normalizing the other features, similarly to [34], for dealing with miscalibrated softmax probability vectors. Xu et al. [47] address the classification objective from a different perspective: their classifier is trained using an information-theoretic-based loss aiming at overcoming data with noisy labels. With the same goal of robustness against noisy labels, Amid et al. [2] propose to couple a novel activation function, called Tempered softmax, with

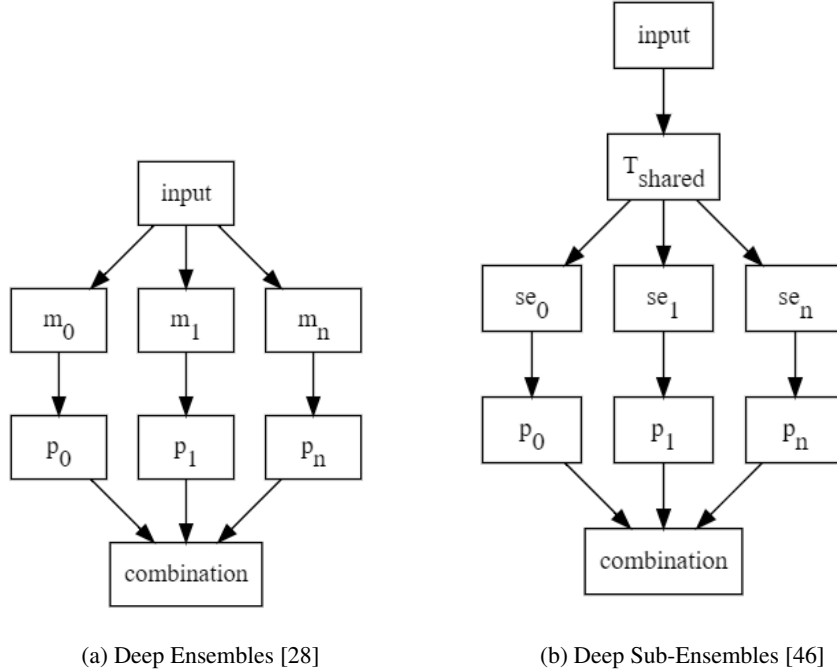


Figure 1: Schematic comparison of Deep ensembles [28] vs. Deep Sub-Ensembles [46] (as depicted in [46]). Deep ensembles do not share any layer, while deep sub-ensembles train the trunk only once and share that trunk across sub-ensembles.

an adequate loss<sup>1</sup>.  $L1$  and  $L2$  norms have commonly been used as loss functions for classification and regression problems. These loss functions are distance-based and very similar one to another or to the cross-entropy loss already used in the relaxed softmax head. However and as these norms provided notable improvements, we included  $L1$  and  $L2$  to the set of selected loss functions. After convergence of the training, we perform variational inference (VI) as described in Equation 1 and similarly to DSE [46].

**Uncertainty estimation measures** Estimation of the model uncertainty might be done in several manners. One option would be to use the normalized entropy of the averaged probability vector as in [29].

$$u = \frac{1}{\log(C)} \sum_{i=1}^C \bar{y}_i \log(\bar{y}_i) \quad (2)$$

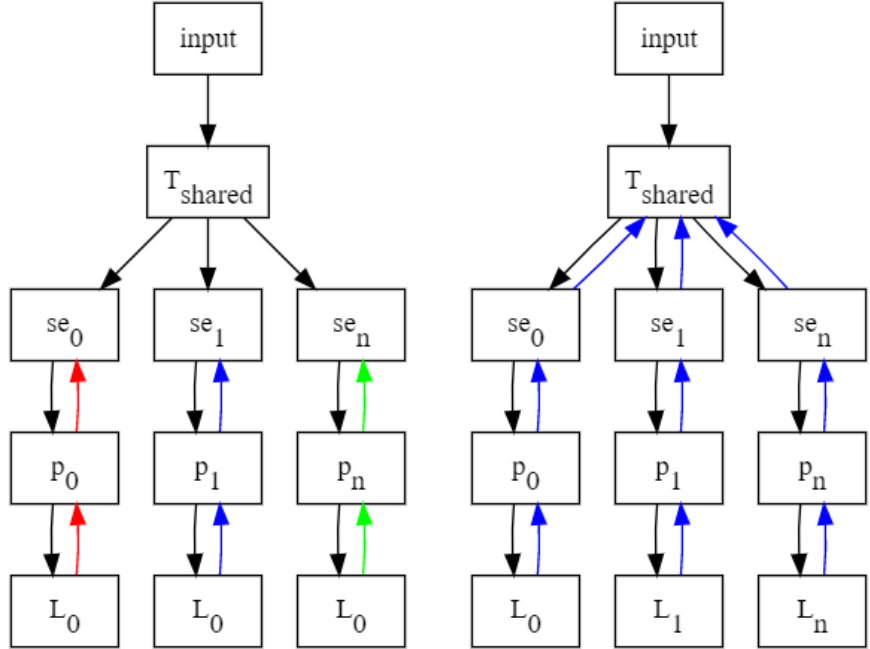
where  $\bar{y}$  indicates the **averaged** probability vector and  $C$  is the number of classes. Alternatively, the normalized entropy could be averaged for each prediction vector  $y_i$  produced by the sub-ensemble  $se_i$ .

$$u = \frac{1}{M} \frac{1}{\log(C)} \sum_{j=1}^M \sum_{i=1}^C \hat{y}_i \log(\hat{y}_i) \quad (3)$$

where  $\hat{y}$  indicates a **single** probability vector. We also introduce a metric for uncertainty measure on variational inference setups, estimated as the variance of the predicted class  $j$  across the different

---

<sup>1</sup>As accuracy was not improved using Tempered softmax loss, the results on these experiments were omitted from section 4



(a) Training procedure of the DSE approach [46]: each branch  $se_i$  is trained separately using the same loss function.

(b) The proposed approach: the training is end-to-end. Each branch is optimized with its own loss function.

Figure 2: Schematic view of the training procedures used in our approach and in [46]. Black arrows represent the forward pass and colored arrows depict the back-propagation paths in the network. Different colors denote different training phases.

predictions  $\hat{y}_{j_i}$ , formally:

$$\begin{aligned}
 u &= \sum_{i=1}^M (\hat{y}_{j_i} - \mu)^2 \\
 j &= \arg \max_{\hat{y}} \\
 \mu &= \frac{1}{M} \sum_{i=1}^M \hat{y}_{j_i}
 \end{aligned} \tag{4}$$

## 4 Experiments

**Experiment design intent** In order to make the approach available for practical use cases, we employ a modest number of heads, even for scalable methods. We employ the loss functions mentioned in section 3 and compare the obtained results with MC-dropout [13], Deep Ensembles [28], Deep Sub-Ensembles[46] as well as Stochastic Weight Averaging (SWA) [20].

**Experimental setup** We use state-of-the-art neural network architectures in order to evaluate our approach under various conditions. From the family of residual architectures, we choose MobileNet-v2 [41], ResNet50[16] and Xception[7]. Additionally, we include VGG-16 [43]. Each of these architectures was trained and tested using Keras[8] on CIFAR10 [24], CIFAR100[25] and ImageNet [10] data sets respectively, for residual architectures. VGG-16 was trained and tested on SVHN [35]. In all experiments<sup>2</sup>, the learning rate was set as  $1e^{-3}$ . We calculate the accuracy and Brier score [4]

<sup>2</sup>The goal is to examine our method and not necessarily find out the best training procedure, hence we chose as simple as possible a training procedure for a uniform and fair comparison

defined as  $BS = \frac{1}{C} \sum_{i=1}^C (t_i^* - p(y=i|x))^2$  where  $t_i^* = 1$  if  $i = y^*$ , and 0 otherwise. Additionally, we compute the expected calibration error (ECE) [37, 15]. The models with best accuracy on the test set among 300 epochs of training were selected for evaluation. In the DSE [46] method and ours, the split point between the trunk and the sub-ensembles was defined with regards to the internal structure of the selected architecture. Specifically, the split point was set for MobileNet-v2, at the last two inverted residual blocks (with 160 and 320 channels, respectively); for ResNet50, at the last residual block (with 2048 channels); for VGG16, at the last 3 convolution blocks (with 128, 256 and 512 channels respectively). For Deep Ensembles, we trained the vanilla model, with no dropout nor any other regularization. We chose relaxed softmax [36] as a base loss for experiments with MobileNet-v2, VGG16 and Xception, while for ResNet50, we chose the standard softmax for the sake of experimental diversity. After the training converged, we additionally trained the models for 30 epochs, and selected the best  $N$  models to compose the ensemble. Those models were also used to calculate the Stochastic Weight Averaging [20] model, with no additional fine-tuning. For VGG16 and MobileNet-v2,  $N = 4$ ; for ResNet50,  $N = 3$ ; and for Xception  $N = 2$ . In MC-dropout, the dropout probability was set as 0.5 for all architectures. The dropout layer was placed after each block depending on the internal structure of the architecture. For the sake of fair evaluation, we set the various key parameters in the most similar fashion for the different methods. Specifically,  $N$  was shared across methods given a pair of architecture/data-set, i.e.  $N$  ensembles,  $N$  heads for DSE, and  $N$  stochastic forward passes for MC-dropout. Implementation details and results are summarized in Table 2 and in Table 1, respectively.

Table 1: Basic accuracy and calibration comparison

Arch/Data-set	Method	Accuracy	ECE	Brier Score
VGG16/SVHN	MC-dropout	0.965	0.05	0.005
	DE	0.963	0.037	0.006
	DSE	0.962	0.652	0.051
	SWA	0.962	0.033	0.006
	Ours	0.966	0.262	0.013
MobileNet-v2/CIFAR10	MC-dropout	0.823	0.055	0.025
	DE	0.873	0.029	0.019
	DSE	0.861	0.034	0.02
	SWA	0.871	0.048	0.022
	Ours	0.875	0.1	0.023
ResNet50/CIFAR100	MC-dropout	0.669	0.029	0.022
	DE	0.709	0.074	0.019
	DSE	0.609	0.101	0.028
	SWA	0.713	0.021	0.019
	Ours	0.679	0.04	0.022
Xception/ImageNet	MC-dropout	0.581	0.011	0.023
	DE	0.801	0.028	0.022
	DSE	0.802	0.030	0.042
	SWA	0.816	0.027	0.022
	Ours	0.832	0.029	0.030

**Uncertainty measures comparison** The uncertainty measure is designed to assess the quality of the model predictions. Failure prediction could be addressed by finding a good separator over a selected uncertainty measure, such as eqs. (2) to (4). We average each of the uncertainty measures on the correct predictions and on the incorrect ones separately. A pronounced difference between the mean uncertainty measures of the true and false predictions might indicate the existence of a good separator. Based on the application needs, a threshold on the uncertainty measure could be set, in such a case. The results are shown in Table 3.

**Accuracy vs. Number of heads** To evaluate the effect of increasing the number of heads in our method, we train Mobile-NetV2 [41] on CIFAR10 and VGG16 [43] on SVHN with 1 to 4 heads and

Table 2: Implementation details

Arch/Data-set	Method	Epochs	Loss	Split
VGG16/SVHN	MC-dropout	300	[36]	-
	DE	300+30	[36]	-
	DSE	300+80x4	[36]	Blocks 3-5
	SWA	300+30	[36]	-
	Ours	300	[36], [42], [47],MSE	Blocks 3-5
MobileNet-v2/CIFAR10	MC-dropout	300	[36]	-
	DE	300+30	[36]	-
	DSE	300+80x4	[36]	Last 2 blocks
	SWA	300+30	[36]	-
	Ours	300	[36], [42], MAE,MSE	Last 2 blocks
ResNet50/CIFAR100	MC-dropout	300	softmax	-
	DE	300+30	softmax	-
	DSE	300+80x3	softmax	Last block
	SWA	300+30	softmax	-
	Ours	300	softmax,[42],MSE	Last block
Xception/ImageNet	MC-dropout	150	[36]	-
	DE	150+30	[36]	-
	DSE	150+50x2	[36]	From 7th middle-flow block onward
	SWA	150+50	[36]	-
	Ours	150	[36], [42]	From 7th middle-flow block onward

Table 3: Uncertainty comparison

Arch/Data set	Method	eqn2		eqn3		eqn4	
		True	False	True	False	True	False
VGG16/SVHN	MC-dropout	0.043	0.349	0.033	0.236	0.006	0.072
	DE	0.009	0.186	0.005	0.107	0.003	0.056
	DSE	0.893	0.904	0.679	0.709	0.148	0.129
	SWA	0.004	0.116	NA	NA	NA	NA
	Ours	0.343	0.544	0.113	0.247	0.163	0.156
MobileNet-v2/CIFAR10	MC-dropout	0.193	0.486	0.163	0.406	0.017	0.044
	DE	0.058	0.28	0.033	0.156	0.016	0.081
	DSE	0.092	0.337	0.079	0.337	0.006	0.03
	SWA	0.021	0.16	NA	NA	NA	NA
	Ours	0.11	0.217	0.0004	0.002	0.061	0.061
ResNet50/CIFAR100	MC-dropout	0.229	0.514	0.21	0.495	0.004	0.009
	DE	0.276	0.549	0.237	0.472	0.024	0.043
	DSE	0.567	0.673	0.403	0.517	0.151	0.103
	SWA	0.207	0.482	NA	NA	NA	NA
	Ours	0.305	0.545	0.273	0.514	0.023	0.015
Xception/ImageNet	MC-dropout	0.275	0.563	0.357	0.448	0.013	0.036
	DE	0.357	0.534	0.364	0.443	0.026	0.046
	DSE	0.757	0.823	0.433	0.663	0.132	0.129
	SWA	0.203	0.452	NA	NA	NA	NA
	Ours	0.317	0.499	0.225	0.431	0.164	0.157

compare their obtained accuracy. The heads were added in the following order: for CIFAR10, [36], [42], Mean Squared Error (MSE), Mean Absolute Error (MAE); for SVHN, [36], [42], [47], MSE. The results are summarized in Figure 3.

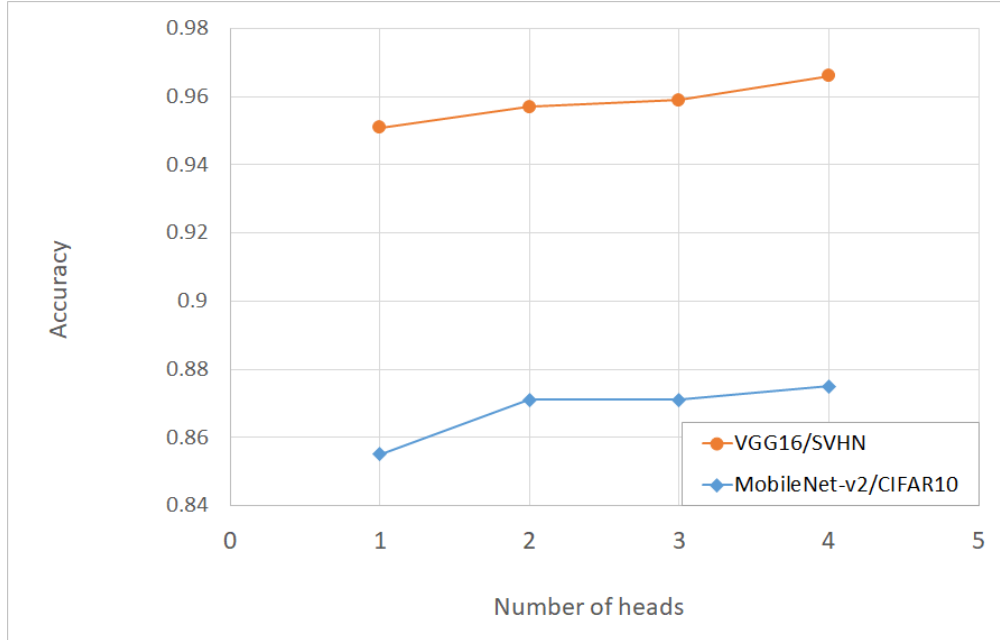


Figure 3: Accuracy vs. number of heads on different data sets

## 5 Discussion

**Accuracy and calibration analysis** As demonstrated in Table 1, the proposed method improves accuracy on SVHN and CIFAR10 as well as ImageNet. On these data sets, the expected calibration error does not surpass each of the compared methods but remains competitive ( $ECE \leq 5\%$ ). Intuitively, the higher ECE could stem from the fact that the models, having trained their own features, may come to provide different *opinions* on the same query image, in contrast to DSE using more similar models, or SWA using a single model. This conjecture yet remains to be rigorously assessed. It is worth noting that the ECE is lower than the one reported in [29], presumably due to the use of novel loss functions in our method [42, 36]. In regards to Brier score, our results are almost identical to the best methods we compare against (absolute diff.  $\leq 0.008$ ), as can be seen in Table 1. On CIFAR100, accuracy was improved over MC-dropout and DSE, though did not outperform DE and SWA. Deep Ensembles is a lot heavier framework than ours and SWA is known as one of the best known methods in the field.

**Uncertainty measures** Uncertainty is often measured by performing out-of-distribution analysis, by training a model on a subset of the classes considered in-distribution and testing on the remaining classes. We alternatively propose to use uncertainty measures for pointing out mispredictions even on in-distribution data. Intuitively, we should expect from a good uncertainty measure, provided a well-performing network, to exhibit a great difference between the sets of correct and incorrect predictions. The largest the difference is, the easier will it be to determine a threshold on the uncertainty measure, separating correct from incorrect predictions. The threshold is use-case dependent and should be set with regards to the allowed rate of false positives or false negatives.

## 6 Conclusion

In this work, we propose to adopt a multi-task approach for single-task variational inference and uncertainty estimation. We show that the approach improves accuracy over the most relevant variational inference methods on several data sets. Additionally, its relatively low run-time requirements make it feasible for practical use cases. Furthermore, we introduce an intuitive way to evaluate uncertainty estimation and to detect mispredictions, even on in-distribution data. For future work, more sophisticated forms of dropout such as [14, 6] could be used to perform variational inference. A regularizer to the different sub-ensembles could be introduced for enhanced control on the consistency of the different predictions, while preserving the gain in accuracy obtained by the diversity in the



heads. It is also worth noting that [9, 21, 1] could be combined with our method, i.e. training a Confid-Net branch for each network head, or using deep feature modeling based on the shared latent space. Aiming at improving separations between true and false predictions might also be an interesting field for active research. The uncertainty measures employed in this work could be used to evaluate the best uncertainty separator for detecting mispredictions, and by such, to bring the system to reach some level of awareness of its own limitations.

## References

- [1] Nilesh A Ahuja et al. “Probabilistic modeling of deep features for out-of-distribution and adversarial detection”. In: *arXiv preprint arXiv:1909.11786* (2019).
- [2] Ehsan Amid et al. “Robust Bi-Tempered Logistic Loss Based on Bregman Divergences”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14987–14996.
- [3] Charles Blundell et al. “Weight uncertainty in neural networks”. In: *arXiv preprint arXiv:1505.05424* (2015).
- [4] Glenn W Brier. “Verification of forecasts expressed in terms of probability”. In: *Monthly weather review* 78.1 (1950), pp. 1–3.
- [5] Mahalanobis Prasanta Chandra et al. “On the generalised distance in statistics”. In: *Proceedings of the National Institute of Sciences of India*. Vol. 2. 1. 1936, pp. 49–55.
- [6] Liyan Chen, Philip Gautier, and Sergul Aydore. “DropCluster: A structured dropout for convolutional networks”. In: *arXiv preprint arXiv:2002.02997* (2020).
- [7] François Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.
- [8] François Chollet et al. *Keras*. <https://keras.io>. 2015.
- [9] Charles Corbière et al. “Addressing Failure Prediction by Learning Model Confidence”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 2898–2909.
- [10] J. Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR09*. 2009.
- [11] Armen Der Kiureghian and Ove Ditlevsen. “Aleatory or epistemic? Does it matter?” In: *Structural safety* 31.2 (2009), pp. 105–112.
- [12] Mohammad Emtiyaz Khan et al. “Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam”. In: *arXiv preprint arXiv:1806.04854* (2018).
- [13] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. 2016, pp. 1050–1059.
- [14] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. “Dropblock: A regularization method for convolutional networks”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 10727–10737.
- [15] Chuan Guo et al. “On calibration of modern neural networks”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1321–1330.
- [16] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [17] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.
- [18] 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*. 1993, pp. 5–13.
- [19] Geoffrey E Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *arXiv preprint arXiv:1207.0580* (2012).
- [20] Pavel Izmailov et al. “Averaging weights leads to wider optima and better generalization”. In: *arXiv preprint arXiv:1803.05407* (2018).
- [21] Heinrich Jiang et al. “To trust or not to trust a classifier”. In: *Advances in neural information processing systems*. 2018, pp. 5541–5552.
- [22] Alex Kendall and Yarin Gal. “What uncertainties do we need in bayesian deep learning for computer vision?” In: *Advances in neural information processing systems*. 2017, pp. 5574–5584.
- [23] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [24] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. “CIFAR-10 (Canadian Institute for Advanced Research)”. In: (). URL: <http://www.cs.toronto.edu/~kriz/cifar.html>.
- [25] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. “CIFAR-100 (Canadian Institute for Advanced Research)”. In: (). URL: <http://www.cs.toronto.edu/~kriz/cifar.html>.

- [26] Meelis Kull, Telmo M Silva Filho, Peter Flach, et al. “Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration”. In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5052–5080.
- [27] Meelis Kull et al. “Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 12295–12305.
- [28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in neural information processing systems*. 2017, pp. 6402–6413.
- [29] Max-Heinrich Laves et al. “Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference”. In: *arXiv preprint arXiv:1909.13550* (2019).
- [30] Kimin Lee et al. “A simple unified framework for detecting out-of-distribution samples and adversarial attacks”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 7167–7177.
- [31] David JC MacKay. “A practical Bayesian framework for backpropagation networks”. In: *Neural computation* 4.3 (1992), pp. 448–472.
- [32] David JC MacKay. “Bayesian methods for adaptive models”. PhD thesis. California Institute of Technology, 1992.
- [33] Wesley J Maddox et al. “A simple baseline for bayesian uncertainty in deep learning”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 13132–13143.
- [34] Azadeh Sadat Mozafari et al. “Attended Temperature Scaling: A Practical Approach for Calibrating Deep Neural Networks”. In: *arXiv preprint arXiv:1810.11586* (2018).
- [35] Yuval Netzer et al. “Reading digits in natural images with unsupervised feature learning”. In: (2011).
- [36] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. “Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection”. In: (2018).
- [37] Alexandru Niculescu-Mizil and Rich Caruana. “Predicting good probabilities with supervised learning”. In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 625–632.
- [38] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015, pp. 91–99.
- [39] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).
- [40] Kotz Samuel, Balakrishnan Narayanswamy, and L Johnson Norman. *Continuous multivariate distributions, models and applications*. 2004.
- [41] Mark Sandler et al. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.
- [42] Murat Sensoy, Lance Kaplan, and Melih Kandemir. “Evidential deep learning to quantify classification uncertainty”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 3179–3189.
- [43] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *CoRR* abs/1409.1556 (2014). URL: <http://arxiv.org/abs/1409.1556>.
- [44] Natasa Tagasovska and David Lopez-Paz. “Single-Model Uncertainties for Deep Learning”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 6414–6425.
- [45] Dustin Tran et al. “Bayesian layers: A module for neural network uncertainty”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14633–14645.
- [46] Matias Valdenegro-Toro. “Deep Sub-Ensembles for Fast Uncertainty Estimation in Image Classification”. In: *arXiv preprint arXiv:1910.08168* (2019).
- [47] Yilun Xu et al. “L\_DMI: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 6222–6233.