Beyond First-Order Uncertainty Estimation with Evidential Models for Open-World Recognition

Charles Corbière, Marc Lafon, Nicolas Thome, Matthieu Cord, Patrick Pérez

Beyond First-Order Uncertainty Estimation with Evidential Models for Open-World Recognition

Charles Corbière [1,2] **Marc Lafon** [1] **Nicolas Thome** [1] **Matthieu Cord** [2,3] **Patrick Pérez** [2]

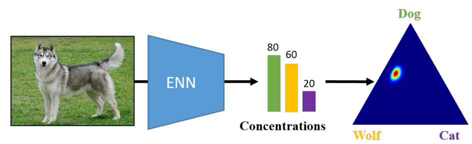
Abstract

In this paper, we tackle the challenge of jointly quantifying in-distribution and out-of-distribution (OOD) uncertainties. We introduce *KLoS*, a KL-divergence measure defined on the class-probability simplex. By leveraging the second-order uncertainty representation provided by evidential models, KLoS captures more than existing first-order uncertainty measures such as predictive entropy. We design an auxiliary neural network, *KLoSNet*, to learn a refined measure directly aligned with the evidential training objective. Experiments show that KLoSNet acts as a class-wise density estimator and outperforms current uncertainty measures in the realistic context where no OOD data is available during training. We also report comparisons in the presence of OOD training samples, which shed a new light on the impact of the vicinity of this data with OOD test data.

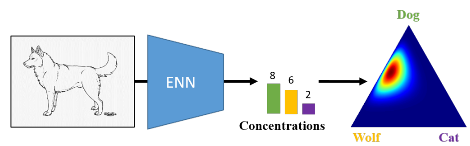
1. Introduction

Obtaining reliable predictive uncertainty estimates is crucial to safely deploy models in open-world conditions (Bendale & Boult, 2015). Notable progress has been made with the renewal of Bayesian neural networks (MacKay, 1992) and ensembling (Lakshminarayanan et al., 2017). These techniques describe an implicit probability density over the predictive categorical distribution obtained from sampling. A recent class of models, coined *evidential* (Sensoy et al., 2018; Malinin & Gales, 2019; Joo et al., 2020), proposes instead to explicitly learn the concentration parameters of a Dirichlet distribution over output probabilities. They have been shown to improve generalisation (Joo et al., 2020) and OOD detection (Nandy et al., 2020)

Based on the subjective logic framework (Josang, 2016), evidential models capture different sources of uncertainty. First-order uncertainty relates to the expectation of the Dirichlet distribution and is caused by conflicting evidence, e.g., class confusion. Second-order uncertainty expresses the lack of evidence in a prediction (Shi et al., 2020), which is characterized by the spread of the Dirichlet distribution. For instance, huskies share lots of features with wolves although being a breed of dog, which leads to a large 1st-order uncertainty due to class confusion in Fig. 1a. In presence of a drawing of a husky, Fig. 1b, a similar class confusion is expected, but a lower amount of evidence due to the distribution shift.



(a) Natural image of a husky
MCP $= 0.50$, entropy $= 0.97$, KLoS $= 97.85$



(b) Drawing of a husky
MCP $= 0.50$, entropy $= 0.97$, KLoS $= 104.71$

Figure 1. **Limitations of 1st-order uncertainty measures**. (a) In-distribution image with conflicting evidence. (b) OOD image with same class confusion is supposed to have larger total uncertainty, which is correctly reflected by its KLoS score.

Surprisingly, previous works do not leverage the distribution over probabilities on the simplex to derive such a joint measure of the two sources of uncertainty. Some methods focus on OOD detection by characterizing only the distribution spread, e.g., using mutual information (Malinin & Gales, 2019). Approaches targeting total uncertainty actually reduce probability distributions on the simplex to their expected value and compute first-order uncertainty measures, e.g., predictive entropy (Sensoy et al., 2018). However, these measures are invariant to the spread of the distribution (Fig. 1), whereas uncertainty caused by class confusion and lack of evidence should be cumulative, a property naturally fulfilled by the predictive variance in Bayesian regression (Murphy, 2012). In addition, some methods for evidential models use auxiliary data during training to enforce higher

[1]CEDRIC, Conservatoire National des Arts et Métiers, Paris, France [2]valeo.ai, Paris, France [3]LIP6, Sorbonne University, Paris, France. Correspondence to: Charles Corbière <charles.corbiere@valeo.com>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

distribution spread on OOD inputs. But when deprived access to OOD training data, the low-dispersion behavior is not guaranteed for all OOD examples (Charpentier et al., 2020; Sensoy et al., 2020) and 2nd-order uncertainty measures struggle to discriminate them from in-distribution examples.

Contributions. We introduce *KLoS*, a KL-divergence measure on the simplex based on the Dirichlet distribution. KLoS provides richer estimates than standard first-order measures by considering both 1st-order and 2nd-order uncertainties. Noting that KLoS naturally reflects the training objective used in evidential models, we propose to learn an auxiliary model, *KLoSNet*, to regress the values of this objective for training samples and to improve uncertainty estimation. Experiments on simultaneous detection of misclassifications and OOD samples show the benefits of KLoSNet thanks to its class-wise density estimator behaviour, a crucial property in the absence of OOD training data. We also shed a new light on the impact of the type of OOD training samples for existing measures.

2. Capturing 1^{st}- and 2^{nd}-Order Uncertainties

2.1. Background: Evidential Neural Networks

Let us consider a training dataset $\mathcal{D}$ composed of N *i.i.d.* samples $(\boldsymbol{x}, y)$ drawn from an unknown joint distribution $p(\mathbf{x}, \mathbf{y})$. We denote $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_C)$ the random variable over categorical probabilities, where $\sum_{c=1}^{C} \pi_c = 1$, and which lives on the (C-1)-dimensional simplex $\triangle^{C-1}$. Bayesian models and ensemble approaches approximate the posterior predictive distribution $p(y|\boldsymbol{\pi}, \boldsymbol{x})$ by marginalizing over a network's parameters $\boldsymbol{\theta}$ via Monte-Carlo sampling or explicit ensembling. But this comes at the cost of multiple forward passes. Evidential Neural Networks (ENN) propose instead to model explicitly the posterior distribution over categorical probabilities by a Dirichlet distribution,

$$q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x}) = \mathrm{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}\big) = \frac{\Gamma(\alpha_0)}{\prod_{c=1}^{C}\Gamma(\alpha_c)} \prod_{c=1}^{C} \pi_c^{\alpha_c - 1}, \quad (1)$$

whose concentration parameters $\boldsymbol{\alpha} = \exp f(\boldsymbol{x}, \boldsymbol{\theta})$ are output by a network f with parameters $\boldsymbol{\theta}$; Γ is the Gamma function and $\alpha_0 = \sum_{c=1}^{C} \alpha_c$. Precision α_0 controls the sharpness of the density with more mass concentrating around the mean as α_0 grows. By conjugate property, the predictive distribution for a new point $\boldsymbol{x}^*$ is $P(\mathbf{y} = c|\boldsymbol{x}^*, \mathcal{D}) \approx \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x}^*)}[\pi_c] = \frac{\exp f_c(\boldsymbol{x}^*, \boldsymbol{\theta})}{\sum_{k=1}^{C} \exp f_k(\boldsymbol{x}^*, \boldsymbol{\theta})}$, which is the output of a network with softmax activation.

ENN training is formulated as a variational approximation to minimize the KL divergence between the distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\pi}|\boldsymbol{x})$ and the true posterior $p(\boldsymbol{\pi}|\boldsymbol{x}, y)$. Following Joo et al. (2020), we use the non-informative uniform prior $p(\boldsymbol{\pi}|\boldsymbol{x}) = \mathrm{Dir}\big(\boldsymbol{\pi}|\mathbf{1}\big)$. The evidential training objective thus reads:

$$\mathcal{L}_{\text{var}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{N} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \Big(\psi(\alpha_y) - \psi(\alpha_0) + \lambda \mathrm{KL}\big(\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \parallel \mathrm{Dir}(\boldsymbol{\pi}|\mathbf{1})\big)\Big), \quad (2)$$

where ψ is the digamma function and with hyperparameter $\lambda > 0$. In particular, minimizing this loss enforces training sample's precision α_0 to remain close to $C + 1/\lambda$.

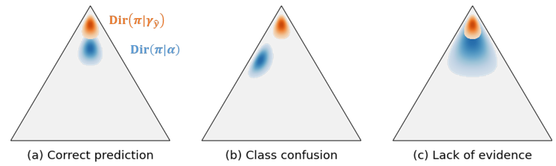


Figure 2. Illustration of KLoS behavior in absence of uncertainty (a), with class confusion (b) and with lack of evidence (c).

2.2. A KL-Divergence Measure on the Simplex

By explicitly learning a distribution of the categorical probabilities $\boldsymbol{\pi}$, evidential models can distinguish first-order from second-order uncertainty on the simplex. Inputs with large first-order uncertainty due to class confusion will have a distribution closer to the simplex center. Conversely, inputs with large second-order uncertainty are expected to present flat distributions, reflecting the lack of evidence of the model on these points. To encompass both types of uncertainty, an efficient measure needs to encapsulate both the sharpness of the distribution and its location on the simplex.

We introduce a novel measure, named *KLoS* for "KL on Simplex", that computes the KL divergence between the model's output and a sharp Dirichlet distribution with concentrations $\boldsymbol{\gamma}_{\hat{y}}$ focused on the *predicted* class $\hat{y}$:

$$\mathrm{KLoS}(\boldsymbol{x}) \triangleq \mathrm{KL}\Big(\mathrm{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}\big) \parallel \mathrm{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\gamma}_{\hat{y}}\big)\Big), \quad (3)$$

where $\boldsymbol{\alpha} = \exp f(\boldsymbol{x}, \boldsymbol{\theta})$ are model's output and $\boldsymbol{\gamma}_{\hat{y}} = (1, \ldots, 1, \tau, 1, \ldots, 1)$ are the uniform concentration parameters except for the predicted class with $\tau = 1/\lambda + 1$.

The lower KLoS is, the more certain the prediction is. Correct predictions will have Dirichlet distributions close to the posterior distribution and will thus be associated with a low uncertainty measure (Fig. 2a). Samples with high class confusion will present concentration parameters closer to simplex's center than the target Dirichlet objective, resulting in a higher KLoS measure (Fig. 2b). KLoS also penalizes samples having a different precision α_0 than the precision $\alpha_0^* = \tau + C - 1$ of the target $\boldsymbol{\gamma}_{\hat{y}}$. For instance, samples with lacking evidence, i.e., having smaller precision than α_0^* (Fig. 2c), receive a larger KLoS score. Since in-distribution samples are enforced to have precision close to α_0^* during training, KLoS will be effective to detect various types of OOD samples whose precision is far from α_0^*, and acts as a class-wise density estimator (see also Section 3.1). In contrast, second-order uncertainty measures, such as the mutual information, assume that OOD samples have smaller α_0, a property difficult to fulfill for models trained only with in-distribution samples (see Section 3.3).

2.3. Improving First-Order Uncertainty Representation with Confidence Learning

When the model misclassifies an example, i.e., the predicted class $\hat{y}$ differs from the ground truth y, KLoS measures the distance between the ENN's output and the wrongly estimated posterior $p(\boldsymbol{\pi}|\boldsymbol{x}, \hat{y})$. Measuring instead the distance to the true posterior distribution $p(\pi|\boldsymbol{x}, y)$ (green region in Fig. 3) would more likely yield a greater value, reflecting the fact that the classifier made an error. Thus, a better measure for misclassification detection would be:

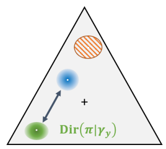


Figure 3. KLoS*

$$\mathrm{KLoS}^*(\boldsymbol{x}, y) \triangleq \mathrm{KL}\Big(\mathrm{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}\big) \parallel \mathrm{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\gamma}_y\big)\Big), \quad (4)$$

where $\boldsymbol{\gamma}_y$ corresponds to the uniform concentrations except for the *true* class y with $\tau = 1/\lambda + 1$.

Connecting KLoS* with evidential training objective. Choosing such value for τ results in KLoS* matching the objective function in Eq. (2). This means that KLoS* is explicitly minimized during training for in-distribution samples and reflects the fact that the model is confident about its prediction if its score is close to zero.

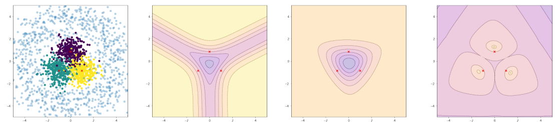
Obviously, the true class of an output is not available when estimating confidence on test samples. We propose to learn KLoS* by introducing an auxiliary confidence neural network, termed KLoSNet, with parameters $\boldsymbol{\omega}$, which outputs a confidence prediction $C(\boldsymbol{x}, \boldsymbol{\omega})$. KLoSNet consists of a small decoder, composed of several dense layers attached to the penultimate layer of the original classification network. During training, we seek $\boldsymbol{\omega}$ such that $C(\boldsymbol{x}, \boldsymbol{\omega})$ is close to KLoS$^*(\boldsymbol{x}, y)$, by minimizing

$$\mathcal{L}_{\text{KLoSNet}}(\boldsymbol{\omega}; \mathcal{D}) = \frac{1}{N} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \big\| C(\boldsymbol{x}, \boldsymbol{\omega}) - \mathrm{KLoS}^*(\boldsymbol{x}, y) \big\|^2. \quad (5)$$

At test time, we now directly use KLoSNet's scalar output $C(\boldsymbol{x}, \boldsymbol{\omega})$ as our uncertainty estimate.

3. Experiments

We evaluate our approach against various baselines: 1$^{\text{st}}$-order uncertainty metrics (Maximum Class probability (*MCP*) and predictive entropy (*Entropy*)), 2$^{\text{nd}}$-order metrics (mutual information (*Mut. Inf.*), differential entropy (*Diff. Ent.*) and expected pairwise KL-divergence (*EPKL*)), post-training methods for OOD detection (*ODIN* (Liang et al., 2018) and *Mahalanobis* (Lee et al., 2018)) and for misclassification detection (*ConfidNet* (Corbière et al., 2019)). Uncertainty measures are derived from the same evidential model trained with $\lambda = 10^{-2}$. We rely on the learned classifier to train our auxiliary confidence model KLoSNet, using the same training set. Except in Section 3.3, we consider setups where no OOD data is available for training. All training details are available in Appendix A.



(a) Toy data (b) Entropy (c) Mut. Inf. (d) KLoS

Figure 4. Visualisation of different uncertainty measures on a toy dataset. Yellow (resp. purple) indicates high (resp. low) certainty.

3.1. Synthetic Experiment

We analyse the behavior of the KLoS measure and the limitations of existing first- and second-order uncertainty metrics on a 2D synthetic dataset composed of three Gaussian-distributed classes with equidistant means and identical isotropic variance (Fig. 4). OOD samples are drawn from a ring around the in-distribution dataset and are only used for evaluation. Fig. 4b shows that Entropy correctly assigns large uncertainty along decision boundaries, which is convenient to detect misclassifications, but yields low uncertainty for points far from the distribution. Surprisingly, Mut. Inf. (Fig. 4c) decreases when moving away from the training data. This behavior is due to the linear nature of the toy dataset where models assign higher concentration parameters far from decision boundaries, hence smaller spread on the simplex, as also noted by Charpentier et al. (2020). Additionally, Mut. Inf. does not reflect the uncertainty caused by class confusion along decision boundaries. In contrast, KLoS allows discriminating both misclassifications and OOD samples from correct predictions as uncertainty increases far from in-distribution samples for each class (Fig. 4d). By measuring a distance between the model's output and a class-wise target distribution, we can observe that KLoS acts as a density estimator for each class.

3.2. Comparative Experiments

When jointly detecting in-distribution misclassifications and OOD samples, correct predictions are considered as positive samples while misclassified inputs and OOD examples constitute negative samples. Following standard practices (Hendrycks & Gimpel, 2017), we use the area under the ROC curve (AUROC) to evaluate threshold-independent

Table 1. Comparative experiments on CIFAR-10. Misclassification (Mis.), OD and simultaneous (Mis+OOD) detection results (mean % AUROC and std. over 5 runs). Bold type indicates significantly best performance ($p < 0.05$) according to paired t-test.

			LSUN		**TinyImageNet**		**STL-10**	
	Method	Mis.	OOD	Mis+OOD	OOD	Mis+OOD	OOD	Mis+OOD
CIFAR-10 VGG-16	MCP (Hendrycks & Gimpel, 2017)	87.6 ±1.6	79.7 ±1.1	84.9 ±1.1	80.3 ±1.5	85.2 ±1.5	60.3 ±1.2	75.2 ±1.4
	Entropy (Sensoy et al., 2018)	83.5 ±2.4	83.8 ±0.3	87.9 ±0.2	82.3 ±0.4	87.2 ±0.4	60.1 ±1.2	75.0 ±1.4
	ConfidNet (Corbière et al., 2019)	90.2 ±0.8	82.1 ±1.5	87.6 ±1.1	83.5 ±0.6	88.3 ±0.7	61.5 ±1.6	77.2 ±1.1
	Mut. Inf. (Malinin & Gales, 2019)	84.1 ±1.5	84.6 ±0.6	85.1 ±1.0	80.6 ±0.8	83.4 ±1.1	61.3 ±0.8	65.0 ±2.5
	Diff. Ent. (Malinin & Gales, 2018)	86.8 ±1.0	85.6 ±0.5	87.2 ±0.7	82.7 ±0.7	85.8 ±0.8	62.0 ±1.0	75.4 ±1.3
	EPKL (Malinin, 2019)	83.9 ±1.5	84.5 ±0.7	85.1 ±1.0	80.4 ±0.8	83.2 ±1.2	61.3 ±0.8	73.8 ±1.1
	ODIN (Liang et al., 2018)	86.0 ±2.0	79.5 ±1.2	83.8 ±1.5	79.6 ±1.9	84.0 ±2.0	54.7 ±1.5	65.0 ±2.6
	Mahalanobis (Lee et al., 2018)	91.2 ±0.3	**88.9** ±0.2	**91.3** ±0.1	**86.4** ±0.2	90.2 ±0.1	63.4 ±0.2	78.8 ±0.3
	KLoSNet (Ours)	**92.5** ±0.6	87.6 ±0.9	**91.7** ±0.9	**86.6** ±0.9	**91.2** ±0.8	**67.7** ±1.4	**81.8** ±0.9

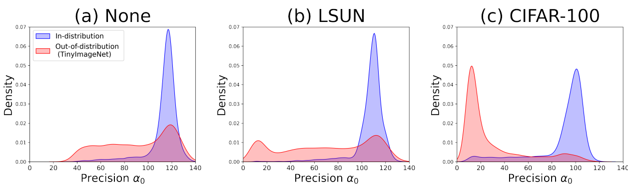
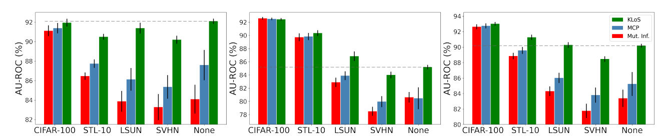


Figure 5. Effect of OOD training data on precision α_0.



(a) Mis. (b) OOD (c) Mis.+OOD

Figure 6. Comparative results with varying OOD training datasets.

performance. Experiments are conducted with VGG-16 (Simonyan & Zisserman, 2015) architectures on CIFAR-10 (Krizhevsky, 2009). We also report experiments on CIFAR-100 and with ResNet-18 (He et al., 2016) architecture in Appendix B.1. Along with simultaneous detection results, we also provide separate results for misclassifications detection and OOD detection respectively in Table 1.

On OOD detection, density estimation-based methods such as Mahalanobis and KLoSNet outperform other methods, including 2$^{\text{nd}}$-order measures. Indeed, for settings where OOD training data is not available, there is no guarantee that every OOD sample will result in lower predicted concentration parameters as shown by the density plot of precision α_0 in Fig. 5a. This stresses the importance of class-wise density estimation. While Mahalanobis may sometimes be slightly better than KLoSNet for OOD detection, it performs significantly less well in misclassification detection. As a result, KLoSNet appears to be the best measure in every simultaneous detection benchmark. For instance, for CIFAR-10/STL-10 benchmark, KLoSNet achieves 81.8% AUROC while the second best, Mahalanobis, scores 78.8%. We also observe that KLoSNet improves significantly misclassification detection, even compared to dedicated methods such as ConfidNet. In Appendix B.2, we provide a detailed ablation study to evaluate the impact of confidence learning.

3.3. Effect of Training with OOD Samples

The literature on evidential models only deals with an OOD training set somewhat related to the in-distribution dataset, e.g., CIFAR-100 for models trained on CIFAR-10. In Fig. 6, we vary the OOD training set used to train an evidential model with the reverse KL-divergence loss (Malinin & Gales, 2019) and evaluate performances using TinyImageNet as OOD test set. As expected, using CIFAR-100 as OOD training data improves performance for every measure (MCP, Mut. Inf. and KLoS). However, the boost provided by training with OOD samples depends highly on the chosen dataset: The performance of Mut. Inf. decreases from 92.6% AUC with CIFAR-100 to 82.9% when switching to LSUN, and even becomes worse with SVHN (78.5%) compared to using no OOD data (80.6%). We also note that KLoS outperforms or is on par with MCP and Mut. Inf. in every setting. Most importantly, using KLoS on models without OOD training data yields better detection performance than other measures taken from models trained with inappropriate OOD samples, here being every OOD dataset other than CIFAR-100.

4. Discussion

We propose KLoSNet, an auxiliary model to estimate the uncertainty of a classifier for both in-domain and out-of-domain inputs. Experiments demonstrate its effectiveness on simultaneous detection of misclassifications and of OOD samples, and reveal its class-wise density estimation behavior. Far from being the panacea, using training OOD samples depends critically on the choice of these samples for existing uncertainty measures. Conversely, KLoS is more robust to this choice and can alleviate their use altogether.

References

Bendale, A. and Boult, T. Towards open world recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems*, 2020.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017.

Joo, T., Chung, U., and Seo, M.-G. Being bayesian about categorical probability. In *Proceedings of the International Conference on Machine Learning*, 2020.

Josang, A. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016.

Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of International Conference on Learning Representations*, 2018.

MacKay, D. J. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

Malinin, A. *Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment*. PhD thesis, University of Cambridge, 2019.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018.

Malinin, A. and Gales, M. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.

Murphy, K. P. *Machine learning : A Probabilistic Perspective*. MIT Press, 2012.

Nandy, J., Hsu, W., and Lee, M. Towards maximizing the representation gap between in-domain and out-of-distribution examples. In *Advances in Neural Information Processing Systems*, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018.

Sensoy, M., Kaplan, L., Cerutti, F., and Saleki, M. Uncertainty-aware deep classifiers using generative models. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Shi, W., Zhao, X., Chen, F., and Yu, Q. Multifaceted uncertainty estimation for label-efficient deep learning. In *Advances in Neural Information Processing Systems*, 2020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

A. Experimental Setup

In this section, we provide comprehensive details about the datasets, the implementation and the hyperparameters of the experiments shown in Section 3.

Synthetic Data. The training dataset (Fig. 4a) consists of 1,000 samples $(\boldsymbol{x}, y)$ from a distribution $p(\mathbf{x}, \mathbf{y})$ over $\mathbb{R}^2 \times \{1, 2, 3\}$ defined as:

$$p(\mathbf{x} = \boldsymbol{x}, \mathbf{y} = y) = \frac{1}{3}\mathcal{N}(\mathbf{x} = \boldsymbol{x} | \boldsymbol{\mu}_y, \sigma^2 \mathrm{I}_{2\times 2}), \quad (6)$$

where $\boldsymbol{\mu}_1 = (0, \sqrt{3}/2)$, $\boldsymbol{\mu}_2 = (-1, -\sqrt{3}/2)$, $\boldsymbol{\mu}_3 = (1, -\sqrt{3}/2)$ and $\sigma = 4$. The marginal distribution of $\mathbf{x}$ is a Gaussian mixture with three equally weighted components having equidistant centers and equal spherical covariance matrices.

The test dataset consists of 1,000 other samples from this distribution. Finally, we construct an out-of-distribution (OOD) dataset following Malinin & Gales (2019), by sampling 100 points in $\mathbb{R}^2$ such that they form a 'ring' with large noise around the training points. Some OOD samples will be close to the in-distribution while others will be very far (see Fig. 3 of the paper). The number of OOD samples has been carefully chosen so that it amounts approximately to the number of test points misclassified by the classifier. Classification is performed by a simple logistic regression.

A set of five models is trained for 200 epochs using the evidential training objective (Eq. 7 of the paper) with regularization parameter $\lambda = 5\text{e-}2$ and Adam optimizer with learning rate 0.02. Uncertainty metrics – MCP, Entropy, Mut. Inf., Malahanobis and KLoS – are computed from these models.

Image Classification Datasets. Experiments are conducted using CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009). They consist in 32×32 natural images featuring 10 object classes for CIFAR-10 and 100 classes for CIFAR-100. Both datasets are composed with 50,000 training samples and 10,000 test samples. We further randomly split the training set to create a validation set of 10,000 images.

OOD datasets are TinyImageNet[1] – a subset of ImageNet (10,000 test images with 200 classes) –, LSUN (Yu et al., 2015) – a scene classification dataset (10,000 test images of 10 scenes) – and STL-10 – a dataset similar to CIFAR-10 but with different classes. We downsample each image of TinyImageNet, LSUN and STL-10 to size 32×32.

Training Details. We implemented in PyTorch (Paszke et al., 2019) a VGG-16 architecture (Simonyan & Zisserman, 2015) in line with the previous works of (Malinin & Gales, 2019; Charpentier et al., 2020; Nandy et al., 2020), with fully-connected layers reduced to 512 units. In both experiments, the models are trained for 200 epochs with a batch size of 128 images, using a stochastic gradient descent with Nesterov momentum of 0.9 and weight decay 5e-4. The learning rate is initialized at 0.1 and reduced by a factor of 10 at 50% and 75% of the training progress. Images are randomly horizontally flipped and shifted by ± 4 pixels as a form of data augmentation.

[1]https://tiny-imagenet.herokuapp.com/

Balancing Misclassification and OOD Detection The models used in the experiments present high predictive performances as shown in Table 2. Most often, there are much fewer misclassifications in the test set. Hence, joint detection performances might be dominated by the evaluation of the quality of OOD detection. To mitigate this unbalance, we propose to consider the following scheme based on oversampling. Let $\mathcal{A}_{\mathrm{M}}$ be the subset of in-distribution test examples that are misclassified by the observed model and $\mathcal{A}_{\mathrm{O}}$ the set of OOD test samples. We randomly sample $\kappa|\mathcal{A}_{\mathrm{O}}|$ points in $\mathcal{A}_{\mathrm{M}}$, with $\kappa = 1$. Supposing $|\mathcal{A}_{\mathrm{O}}| \geq |\mathcal{A}_{\mathrm{M}}|$, this corresponds to oversampling the set of misclassifications. This over-sampled set is then added to the OOD set to form the negative examples for detection training. The set of correct predictions remains the same. We observed that the variance in AUROC due to this sampling is negligible and we report only the mean.

	CIFAR-10	CIFAR-100
Train	99.0 ±0.1	91.2 ±0.2
Val	93.6 ±0.1	70.6 ±0.3
Test	93.0 ±0.3	70.1 ±0.4

Table 2. Mean accuracies (%) and std. over five runs.

KLoSNet. We start from the pre-trained evidential model described above. As detailed in Section 2.3, KLoSNet consists of a small decoder attached to the penultimate layer of the main network. In CIFAR experiments, this corresponds to VGG16's fc1 layer of size 512. This auxiliary neural network is composed of five fully-connected layers of size 400, except for the last layer obviously. KLoSNet decoder's weights ω are trained for 100 epochs with ℓ_2 loss (Eq. 11 in the main paper) and with Adam optimizer with learning rate 1e-4. As KLoS* ranges from zero to large positive values (>1000), one may encounter some issues when training KLoSNet. Consequently, we apply a sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$, after computing the KL-divergence between NN's output and γ_y. To prevent over-fitting, training is stopped when validation AUC metric for misclassification detection starts decreasing. Then, a second training step is performed by initializing new encoder E' such that $\theta_{E'} = \theta_E$ and by optimizing weights $(\theta_{E'}, \omega)$ for 30 epochs with Adam optimizer with learning rate 1e-6. We stop training once again based on validation AUC metric.

B. Additional Results

B.1. Results on CIFAR-100 and with ResNet-18

In Table 3, we extend our comparative experiments on simultaneous detection to CIFAR-100 dataset and to mod-

Table 3. Comparative experiments on CIFAR-10 and CIFAR-100 with ResNet-18 architectures. Misclassification (Mis.), out-of-distribution (OOD) and simultaneous (Mis+OOD) detection results (mean % AUROC and std. over 5 runs). Bold type indicates significantly best performance ($p < 0.05$) according to paired t-test

	Method	Mis.	LSUN OOD	LSUN Mis+OOD	TinyImageNet OOD	TinyImageNet Mis+OOD	STL-10 OOD	STL-10 Mis+OOD
CIFAR-10 ResNet-18	MCP	84.9 ±0.8	79.6 ±1.0	83.0 ±0.9	77.2 ±0.7	81.8 ±0.7	58.5 ±1.2	72.5 ±0.4
	Entropy	84.6 ±0.8	79.6 ±1.1	82.8 ±0.9	77.2 ±0.7	81.6 ±0.7	58.4 ±1.2	72.2 ±0.4
	ConfidNet	90.7 ±0.4	84.6 ±1.1	88.6 ±0.6	83.5 ±1.1	88.0 ±0.6	63.2 ±1.2	77.9 ±0.5
	Mut. Inf	80.6 ±0.6	77.0 ±1.2	79.4 ±0.9	74.3 ±0.8	78.0 ±0.7	56.4 ±1.0	69.1 ±0.2
	Diff. Ent	82.7 ±0.6	78.3 ±1.2	81.1 ±0.9	75.9 ±0.8	79.9 ±0.7	57.5 ±1.1	70.8 ±0.3
	EPKL	80.2 ±0.6	76.8 ±1.3	79.0 ±0.9	74.1 ±0.8	77.7 ±0.7	56.2 ±1.0	68.9 ±0.3
	ODIN	83.7 ±0.7	78.9 ±1.0	81.9 ±0.9	76.5 ±0.7	80.7 ±0.7	57.9 ±1.2	71.5 ±0.4
	Mahalanobis	91.2 ±0.4	90.7 ±0.4	91.8 ±0.3	87.6 ±0.4	90.3 ±0.4	66.8 ±0.5	80.0 ±0.3
	KLoSNet (Ours)	**93.9** ±0.4	**93.1** ±1.1	**94.4** ±0.3	**90.6** ±0.6	**93.2** ±0.2	**68.5** ±0.3	**82.3** ±0.2
CIFAR-100 VGG-16	MCP	82.9 ±0.8	62.8 ±1.3	77.6 ±0.9	72.0 ±0.5	81.8 ±0.7	69.7 ±0.7	80.9 ±0.7
	Entropy	82.2 ±0.8	63.2 ±1.4	77.2 ±1.0	72.5 ±0.6	81.5 ±0.8	70.1 ±0.8	80.6 ±0.7
	ConfidNet	84.4 ±0.6	65.3 ±2.0	80.0 ±1.3	73.8 ±0.6	83.7 ±0.7	71.5 ±0.6	82.7 ±0.3
	Mut. Inf.	78.9 ±0.8	65.6 ±0.7	76.2 ±0.9	71.8 ±0.2	79.1 ±0.4	70.1 ±0.6	78.5 ±0.6
	Diff. Ent.	80.2 ±0.8	65.6 ±0.9	77.2 ±0.8	72.7 ±0.3	80.4 ±0.4	71.0 ±0.5	79.7 ±0.5
	EPKL	78.8 ±0.8	65.2 ±1.0	76.1 ±0.9	71.6 ±0.2	78.9 ±0.4	70.0 ±0.6	78.3 ±0.6
	ODIN	82.1 ±0.8	62.9 ±1.4	77.1 ±1.0	71.9 ±0.6	81.3 ±0.8	69.6 ±0.8	80.3 ±0.7
	Mahalanobis	84.0 ±0.2	**71.1** ±1.0	82.4 ±0.5	**77.0** ±0.5	84.9 ±0.3	**75.4** ±0.3	84.3 ±0.5
	KLoSNet (Ours)	**86.7** ±0.4	68.4 ±1.1	**83.0** ±0.6	76.4 ±0.4	**86.4** ±0.4	**75.0** ±0.5	**86.0** ±0.4
CIFAR-100 ResNet-18	MCP	84.0 ±0.4	70.4 ±0.9	81.0 ±0.3	76.6 ±0.5	83.6 ±0.4	75.4 ±0.5	83.1 ±0.2
	Entropy	83.7 ±0.4	70.4 ±0.9	80.8 ±0.3	76.9 ±0.5	83.5 ±0.3	75.7 ±0.5	83.0 ±0.3
	ConfidNet	**87.1** ±0.2	73.0 ±1.4	**84.5** ±0.6	79.1 ±0.3	86.8 ±0.3	**78.5** ±0.8	**86.6** ±0.5
	Mut. Inf	82.6 ±0.4	70.2 ±1.1	80.0 ±0.4	76.4 ±0.6	82.6 ±0.3	75.1 ±0.5	82.1 ±0.3
	Diff. Ent	83.0 ±0.4	70.1 ±1.1	80.2 ±0.4	76.8 ±0.5	83.0 ±0.3	75.6 ±0.5	82.5 ±0.3
	EPKL	82.5 ±0.4	70.2 ±1.1	80.0 ±0.4	76.3 ±0.6	82.5 ±0.3	75.0 ±0.5	82.0 ±0.2
	ODIN	83.7 ±0.4	70.3 ±0.9	80.8 ±0.3	76.6 ±0.5	83.5 ±0.3	75.4 ±0.5	83.0 ±0.3
	Mahalanobis	85.9 ±0.4	**75.2** ±0.6	**84.5** ±0.1	78.4 ±0.5	85.9 ±0.3	77.5 ±0.4	85.6 ±0.3
	KLoSNet (Ours)	86.9 ±0.3	73.1 ±0.4	**84.4** ±0.1	**80.8** ±0.2	**87.3** ±0.2	**79.0** ±0.2	**86.7** ±0.3

Table 4. **Impact of confidence learning.** Comparison of detection performances between KLoS and KLoSNet for CIFAR-10 and CIFAR-100 experiments with VGG-16 architecture.

	Method	Mis.	LSUN OOD	LSUN Mis+OOD	TinyImageNet OOD	TinyImageNet Mis+OOD	STL-10 OOD	STL-10 Mis+OOD
CIFAR-10 VGG-16	KLoS	92.1 ±0.3	86.5 ±0.3	91.2 ±0.2	85.4 ±0.3	90.4 ±0.2	64.1 ±0.3	79.6 ±0.3
	KLoSNet	**92.5** ±0.6	**87.6** ±0.9	**91.7** ±0.9	**86.6** ±0.9	**91.2** ±0.8	**67.7** ±1.4	**81.8** ±0.9
CIFAR-100 VGG-16	KLoS	85.4 ±0.2	65.1 ±1.1	81.3 ±0.6	74.5 ±0.4	85.4 ±0.4	72.7 ±0.3	84.8 ±0.4
	KLoSNet	**86.7** ±0.4	**68.4** ±1.1	**83.0** ±0.6	**76.4** ±0.4	**86.4** ±0.4	**75.0** ±0.5	**86.0** ±0.4

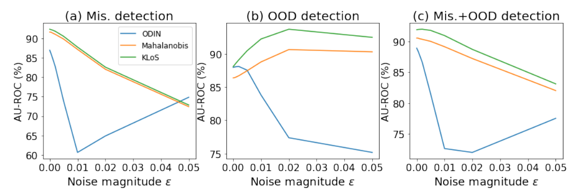
els with ResNet-18 architecture. We can observe that density estimation-based methods, Mahalanobis and KLoSNet, still outperform second-order measures in OOD detection. KLoSNet improves also misclassification detection, even compared to dedicated methods such as ConfidNet. These results confirm that simultaneous detection of misclassifications and OOD samples can be significantly improved by KLoSNet in settings without OOD training data.

B.2. Impact of Confidence Learning

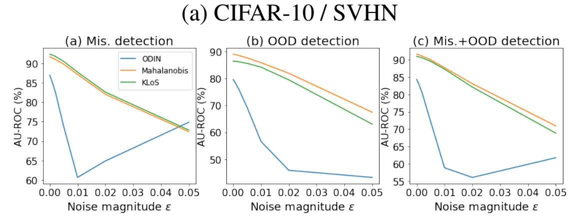
To evaluate the effect of the uncertainty measure KLoS and of the auxiliary confidence network KLoSNet, we report a detailed ablation study in Table 4. We can notice that KLoSNet improves misclassification over KLoS but also OOD detection in every benchmark. We intuit that learning to improve misclassification detection also helps to spot some OOD inputs that share similar characteristics.

B.3. Impact of Adversarial Perturbations

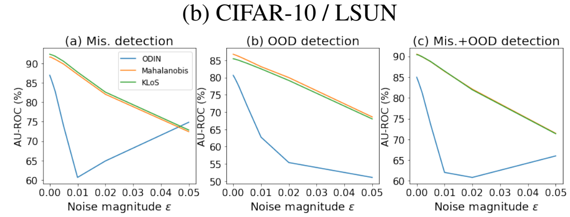
In the original papers, ODIN and Mahalanobis preprocess inputs by adding small inverse adversarial perturbations to reinforce networks in their prediction; this has also the observed benefit to make in-distribution and out-of-distribution samples more separable. The tuning of the adversarial noise's magnitude depends on the evaluated OOD data. In Figure 7a, we plot the AUC of each detection task with different values of perturbation magnitude ε with ODIN, Mahalanobis and our criterion KLoS, using SVHN as OOD dataset. Even though there exists a particular noise value for improved OOD detection (Fig. 7a, middle), increasing noise magnitude deteriorates performances in misclassification detection (Fig. 7a, left) for each method. Best results on the simultaneous detection task (Fig. 7a, right) correspond to $\varepsilon = 0$, as done in previous experiments.



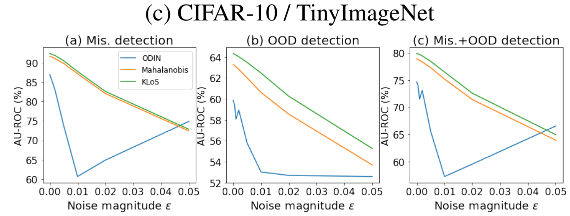
(a) CIFAR-10 / SVHN



(b) CIFAR-10 / LSUN



(c) CIFAR-10 / TinyImageNet



(d) CIFAR-10 / STL-10

Figure 7. Effect of inverse adversarial perturbations on OOD-designed measures and KLoS for misclassification detection, OOD detection and simultaneous detection with VGG16 architecture.

Worse, except with SVHN, adversarial perturbations are detrimental even to OOD detection. We report the AUC results of varying adversarial perturbations on CIFAR-10 dataset when using LSUN (Fig. 7b), TinyImageNet (Fig. 7c) and STL-10 (Fig. 7d) as OOD datasets. Best results on each considered task correspond to $\varepsilon = 0$ and KLoS outperforms both Mahalanobis and ODIN. As opposed to results with SVHN as OOD dataset, we did not observe improvements on any method (ODIN, Mahalanobis and KLoS) when using inverse adversarial perturbations for OOD detection with LSUN, TinyImageNet and STL-10 datasets. Similar results are observed in (Liang et al., 2018) (Appendix B, Fig. 8) when using WideResNet architectures. Regarding Mahalanobis (Lee et al., 2018), the authors only provided ablation for SVHN dataset.

C. Link between KLoS* and Evidential Training Objective

Let us remind the definition of KLoS as a KL divergence between the model's output and a sharp Dirichlet distribution with concentrations $\boldsymbol{\gamma}_{\hat{y}}$ focused on the *predicted* class $\hat{y}$:

$$\text{KLoS}^*(\boldsymbol{x}, y) \triangleq \text{KL}\Big(\text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}\big) \,||\, \text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\gamma}_y\big)\Big), \tag{7}$$

where $\boldsymbol{\alpha} = \exp f(\boldsymbol{x}, \boldsymbol{\theta})$ is model's output and $\boldsymbol{\gamma}_y = (1, \dots, 1, \tau, 1, \dots, 1)$ are the uniform concentration parameters except for the true class with $\tau = 1/\lambda + 1$.

The KL-divergence between two Dirichlet distributions can be obtained in closed form and KLoS* can be calculated as:

$$\begin{aligned}
\text{KLoS}^*(\boldsymbol{x}, y) &= \text{KL}\Big(\text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}\big) \,||\, \text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\gamma}_y\big)\Big) && (8)\\
&= \log\Gamma(\alpha_0) - \log\Gamma(C - 1 + 1/\lambda)\\
&\quad + \log\Gamma(1 + 1/\lambda) - \sum_{c=1}^{C} \log\Gamma(\alpha_c)\\
&\quad + \sum_{c \neq y} \big(\alpha_c - 1\big)\big(\psi(\alpha_c) - \psi(\alpha_0)\big)\\
&\quad + \big(\alpha_y - (1 + 1/\lambda)\big)\big(\psi(\alpha_y) - \psi(\alpha_0)\big). && (9)
\end{aligned}$$

On the other hand, the KL-divergence between the model's output and a uniform Dirichlet distribution $\text{Dir}\big(\boldsymbol{\pi}|\mathbf{1}\big)$ reads:

$$\begin{aligned}
&\text{KL}\Big(\text{Dir}\big(\boldsymbol{\pi}|\boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{\theta})\big) \,||\, \text{Dir}\big(\boldsymbol{\pi}|\mathbf{1}\big)\Big)\\
&\quad = \log\Gamma(\alpha_0) - \log\Gamma(C) - \sum_{c=1}^{C} \log\Gamma(\alpha_c)\\
&\quad\quad + \sum_{c=1}^{C} \big(\alpha_c - 1\big)\big(\psi(\alpha_c) - \psi(\alpha_0)\big). \qquad (10)
\end{aligned}$$

Hence, KLoS* can be written as:

$$\begin{aligned}
\text{KLoS}^*(\boldsymbol{x}, y) = \frac{1}{\lambda}\big(\psi(\alpha_y) - \psi(\alpha_0)\big)&\\
+ \text{KL}\big(\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \,||\, \text{Dir}(\boldsymbol{\pi}|\mathbf{1})\big)&\\
+ \big(\log\Gamma(1 + 1/\lambda) - \log\Gamma(C - 1 + 1/\lambda)&\\
- \log\Gamma(C)\big).& \qquad (11)
\end{aligned}$$

Let us decompose $\mathcal{L}_{\text{var}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{N}\sum_{(\boldsymbol{x}, y) \in \mathcal{D}} l_{\text{var}}(\boldsymbol{x}, y, \boldsymbol{\theta})$. We retrieve that $\text{KLoS}^*(\boldsymbol{x}) \propto l_{\text{var}}(\boldsymbol{x}, y, \boldsymbol{\theta}) + r$ where $r = -\big(\log\Gamma(1 + 1/\lambda) - \log\Gamma(C - 1 + 1/\lambda) - \log\Gamma(C)\big)$ does not depend on the model parameters $\boldsymbol{\theta}$.

In summary, minimizing the evidential training objective $\mathcal{L}_{\text{var}}(\boldsymbol{\theta}; \mathcal{D})$ is equivalent to minimizing the KLoS* value of each training point $\boldsymbol{x}$.