MP5: Robustness of Single Pass Deterministic Uncertainty Estimation Methods

Asher Mai (hanlinm2) **Rasmus Larsen (larsen8)** **Ayush Sarkar (ayushs2)**

Anirudh Choudhary (ac67)

1 Introduction

Neural networks have demonstrated outstanding performance across a range of tasks including computer vision, natural language processing, speech recognition, and clinical diagnosis. Despite their notable achievements, recent research on robustness and generalization has underscored the susceptibility of deep neural networks to robustness challenges. While various strategies such as adversarial training, data augmentation, weight regularization, and noise-based smoothing have been proposed to enhance robustness, deep neural networks remain prone to adversarial and universal perturbations [1].

Additionally, deep networks often exhibit overconfident predictions when confronted with out-of-distribution data, resulting in suboptimal generalization performance. The deployment of safety-critical systems such as autonomous driving and medical diagnosis necessitates reliable uncertainty estimation. Uncertainty quantification assumes significance due to the possibility of encountering out-of-distribution (OOD) data post-training, and predictions accompanied by uncertainty estimates facilitate informed decision-making and foster increased trust in neural networks.

The sources of uncertainty can be categorized into two distinct types: 1) aleatoric uncertainty, stemming from inherent noise within the data samples, and 2) epistemic uncertainty, which pertains to uncertainty in model parameters arising from model complexity and the selection of hypothesis classes. Various methods for uncertainty quantification exist, including Bayesian parameter modeling techniques such as MD-dropout, Bayes-by-backprop, Stochastic Gradient Langevin dynamics, and ensembling-based methods [2]. Although these approaches have demonstrated commendable accuracy alongside reasonable uncertainty estimation, their practical application is hindered by costliness. Typically, these methods necessitate either training additional model instances with an associated increase in parameters or require meticulous prior specification and multiple approximations for posterior weight distribution estimation (as seen in Bayesian neural networks). An alternative avenue, Gaussian processes, faces scalability issues with large datasets and mandates covariance sample approximation. Recently, a promising class of single pass uncertainty estimation models [3, 4, 5] has emerged which enable efficient uncertainty estimation using a single forward pass through the network.

In this work, we focus on single pass uncertainty estimation methods for uncertainty estimation and detecting out-of-distribution inputs in classification problems. These methods consider neural networks weight as deterministic and estimate uncertainty by performing a single forward pass, thus making them parameter-wise and computationally efficient. We consider two methods for quantifying distributional uncertainty due to dataset shift. : 1) Dirichlet networks based on evidential learning which explicitly parameterize the distribution over the predictive categorical with a Dirichlet distribution; 2) Feature-space density or distance aware methods which learn a well-regularized feature space and leverage the distance between points in the representation space to identify OOD inputs. Our reasoning behind selecting these networks is two-fold: 1) Minimal dependency on OOD data for training, making them ideal for deployment in any real-world setting with in-distribution(ID);

Preprint. Under review.

2) Resemble the most common class of deterministic neural networks and do not require expensive ensembling or estimation of underlying data manifold (generative).

Dirichlet networks can differentiate between uncertainty estimates from lack of knowledge(evidence) about OOD data vs noisy in-distribution data (data uncertainty). These networks replace the softmax function at the output of classifier network with dirichlet distribution and learn its parameters using the neural network defining a prior. We focus on a particular class of Dirichlet networks named posterior networks, which try to minimize the difference between a prior and posterior distribution. Feature-space density methods leverage spectral normalization layers restricting the Lipschitz constant of the model so that large changes in the input image lead to large distances in feature space. Here, we focus on a density estimation approach (DDU [4]) which fits a class-conditioned Gaussian distribution over the regularized feature space and leverage log-likelihood as a measure of epistemic uncertainty. Existing approaches based on bayesian neural networks or ensembles hide distributional uncertainty with data uncertainty or implicitly model it through model uncertainty. By separating different types of uncertainty, both our considered methods can allow appropriate action to be taken to re-train or finetune the trained network. Our analysis involves one of the first comparisons between these two different approaches over diverse domain generalization settings such as image corruptions and domain shifts. In line with [6], we focus on analyzing the robustness properties of these two approaches beyond classification accuracy. We first introduce the Dirichlet multinomial model and relate it to posterior Dirichlet network. Subsequently, we highlight a very simple formulation of feature-distance aware method which estimates feature space distribution using Gaussian Mixture Model. We evaluate their robustness against the deterministic softmax classifier. Our results highlight that Dirichlet networks achieve higher accuracy on natural images and are more robust to image corruptions but tend to under-perform on domain-shift settings.

2 Background

2.1 Dirichlet Distribution

In a binary classification problem, the conjugate prior for Bernoulli likelihood is the beta distribution. Dirichlet distribution is a multivariate extension of Beta distribution, which is a conjugate prior for categorical distribution and is suitable for multiclass classification. The Dirichlet distribution, characterized by its concentration parameters, α is given as:

$$p(\pi) = Dir(\pi; \alpha) = \frac{\mathcal{G}(\beta_0)}{\prod_{k=1}^{K} \mathcal{G}(\beta_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{1}$$

where $\mathcal{G}(\cdot)$ is the gamma function, K is the number of classes. The sum of α_i is denoted as precision α_0. In a classification problem, typically softmax function is applied on the last layer to derive the categorical distribution. Given a dataset $D = \{x_i, y_i\}_{i=1}^{n}$, multiplying the prior Dirichlet distribution with the categorical likelihood π_i leads to Dirichlet posterior (due to conjugacy) with parameters β corresponding to the K classes.

$$Dir(\pi, \beta) = Dir(\pi | \alpha_1 + N_1, ..., \beta_k + N_k) \tag{2}$$

where N_k is the number of samples belonging to class K in the dataset. Thus, the prior distribution acts as a Bayesian smoothing mechanism by adding pseudo-count α to the true count.

2.2 Dirichlet Networks

Consider a classification problem with K classes, a standard softmax neural network defines a categorical distribution over classes given by $p = p(y = k|x, \theta)$. Typically, the aleatoric uncertainty is computed using categorical entropy. Existing approaches such as ensembling and bayesian neural networks leverage Monte-carlo sampling to estimate an expected distribution and used predictive entropy of expected distribution for aleatoric uncertainty estimation. However, it is difficult to determine whether the uncertainty is due to data uncertainty or because of the sample being from out-of-distribution data. Dirichlet networks extend the softmax framework by estimating the parameters of a Dirichlet distribution using the neural network. A categorical distribution is then created from the predicted concentration parameters of the Dirichlet distribution. Parameterizing output distribution using Dirichlet allows us to achieve the following advantages: 1) When the input has high degree

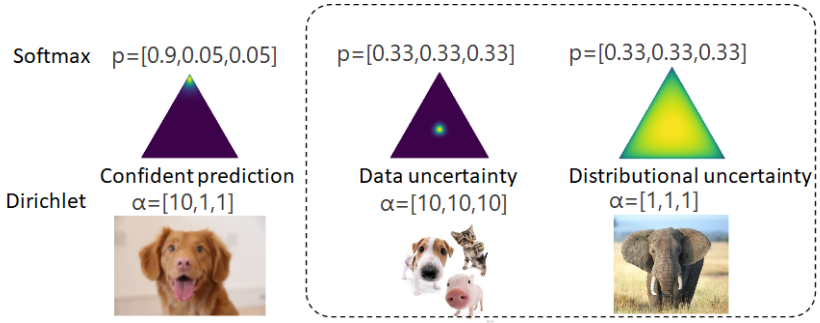

Figure 1: Behaviour of Dirichlet vs Softmax in different scenarios: Probability simplex for a 3-class classification problem. Every corner corresponds to a class and every point to a categorical distribution. Brighter colors represent higher density

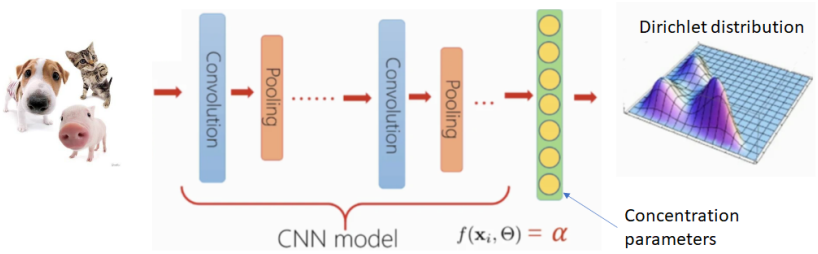

Figure 2: Basic framework of a Dirichlet network with the concentration parameters of Dirichlet distribution estimated using the output of last activation layer

of noise or class overlap, it will lead to a sharp distribution focused on the center of the simplex, which means that the networks is confident in predicting a flat distribution over classes. For out-of-distribution data, the simplex obtained from the Dirichlet network will be a flat distribution, indicating high epistemic uncertainty (Figure 1).

2.3 Feature Space Distance Approach

Feature-Space Distance based method leverage distances and densities within the feature space to provide uncertainty estimates (Figure 3). This approach estimates the density or distance of a data point to the training data within the model's feature space. An OOD point is expected to have a high distance to the learned feature representations and reside in a low-density region. In contrast, in-distribution data points exhibit low distances and reside in high-density areas.

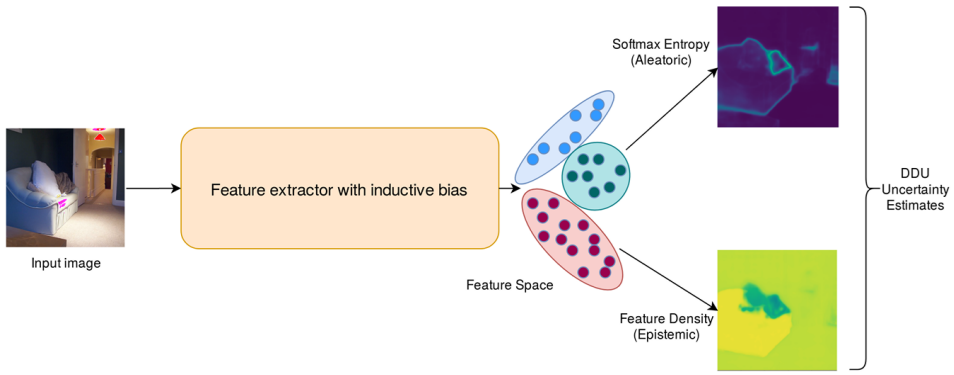

Figure 3: Basic framework of Feature Space Distance approach [4]

However, a key challenge arises from the tendency of discriminative classifiers to learn invariant representations by exploiting spurious correlations within the training data. This is particularly problematic for high-dimensional representations, leading to a phenomenon known as feature collapse [5]. Feature collapse poses a significant hurdle for OOD detection using distances, as OOD embeddings can collapse onto in-distribution features, making them indistinguishable.

Recent advancements address feature collapse by incorporating distance-aware representations that establish a relationship between the distances in the latent representation space and those in the input space. These methods often impose constraints on the bi-Lipschitz constant. A lower bound on this constant guarantees that distinct input points map to distinct representations, effectively mitigating feature collapse. Conversely, an upper limit promotes smoothness, ensuring that small changes in the input lead to small changes in the latent space. Spectral Normalization (SN) offers a computationally efficient approach to enforce distance-awareness in the hidden representation space. SN normalizes the weights of each layer based on their spectral norm, making it particularly well-suited for residual networks.

For our investigation, we focus on a recent distance-based uncertainty model named Deep Density Uncertainty (DDU) [Citation Needed] due to its straightforward formulation involving fitting a Gaussian Mixture Model (GMM) over the learnt feature spaces.

3 Existing Approach: Dirichlet Networks

We consider a specific type of Dirichlet Network called posterior networks. These networks leverage the idea that choosing a Dirichlet prior induces a Dirichlet posterior distribution. In our analysis, we focus on Bayesian Matching Networks [7].

3.1 Belief Matching Network (Dirichlet-BM)

Belief Matching networks estimate the true posterior distribution over the categorical distribution, that is, consider categorical probability as a random variable following the Dirichlet distribution instead of representing it as a one-hot vector over K classes. Defining a prior as a Dirichlet distribution with concentration parameters β_i, the posterior distribution after observing the data set $D = \{x_i, y_i\}$ is given as:

$$p(\pi|x, y) = Dir(\pi|\beta_1 + \tilde{N}_1(x), ..., \beta_1 + \tilde{N}_K(x)) \tag{3}$$

where K is the total number of classes and $\tilde{N}_k(x)$ represents the empirical frequency of the label corresponding to sample x. The posterior mean is smoothed by the prior belief β_i. Joo et al [7] proposed a belief matching framework (Dirichlet-BM) wherein they apply the above Bayesian approach to construct the target distribution for learning classifiers. This target distribution is then approximated by neural network f_θ with a variational distribution $q_\theta(\pi|x)$. $q_\theta(\pi|x)$ is chosen as the Dirichlet distribution with concentration parameters given by $\alpha = \exp^{f_\theta(x)}$ and the KL divergence between the two Dirichlet distributions $p(\pi|x, y)$ and $q_\theta(\pi|x)$ is minimized. The approach leverages variational inference to optimize the neural network, arriving at the following loss function:

$$L_{VI}(\theta) = \mathbb{E}_{p(x,y)}[\mathbb{E}_{q_\theta(\pi|x)}[-logp(y|\pi, x) + KL(q_\theta(\pi|x)||p(u|x))] \tag{4}$$

where the first term is $\propto \psi(\alpha_y) - \psi(\alpha_0)$ where α_0 is the precision term highlighted earlier. The authors choose a uniform Dirichlet prior setting prior concentrations $\beta = 1$ which corresponds to an agnostic belief in evidential theory.

4 Existing Approach: Feature Space Distance

We select Deep Deterministic Uncertainty (DDU) [8] which learns a Gaussian Mixture Model over distance-aware feature space to quantify uncertainty. The reason behind selecting DDU is its simple formulation compared to other methods such as DUQ [9] and SNGP [5] which require significant hyperparameter tuning and do not scale well to large number of classes.

4.1 Deep Deterministic Uncertainty (DDU)

DDU [4] uses the feature space density of a determinisitic neural network trained with proper inductive biases (bi-Lipschitz constraint). DDU performs Gaussian Discriminant Analysis by fitting

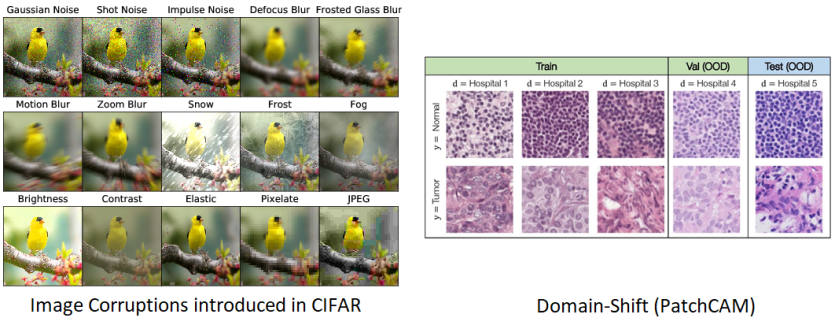

Figure 4: Generalization settings used for evaluating robustness of Dirichlet networks

a class-conditioned GMM over the feature space. The GMM $q(y, z)$ is learned by fitting a single Gaussian mixture component per class using the empirical mean and covariance of feature vectors $z = f_\theta(x)$ belonging to that class. DDU estimates the epistemic uncertainty by marginalizing the feature density over all classes as $p(z)p(z|c)p(c)$ where $p(c|z)$ is parameterized by GDA while $p(c)$ is computed from training data. The entropy $H(y|x, \theta)$ of the softmax distribution $p(y|x, \theta)$ is used to estimate the aleatoric uncertainty. DDU leverages spectral normalization to ensure smoothness and sensitivity of model's latent space to input data.

5 Robustness of Single Pass Uncertainty Estimation Methods

We now analyze the robustness of the two selected single pass uncertainty estimation methods in two generalization settings involving out-of-distribution data : 1) common image corruption; 2) domain shift.

Experimental Setup: We analyze the performance of three models: DDU, Belief Matching Network, and softmax-based deterministic networks across accuracy, confidence calibration, and out-of-distribution (OOD) uncertainty. We pose the following questions and assess them using uncertainty and classification metrics:

1. Does accounting for data and model uncertainty enhance generalization over OOD data?
2. Does spectral normalization in the feature space distance-based approach improve confidence calibration compared to Dirichlet networks?
3. Do feature space distance-aware methods achieve superior out-of-distribution uncertainty compared to Dirichlet networks? Can methods that consider uncertainty better differentiate between OOD and in-distribution (ID) data compared to deterministic softmax-based networks?

We employ specific uncertainty quantification metrics for each model: 1) differential entropy for continuous Dirichlet distribution, 2) softmax entropy for deterministic softmax network, and 3) feature space density for DDU, to measure OOD uncertainty.

To evaluate calibration performance on ID test data, we label the correctly classified samples as '1' and wrongly classified samples as '0', and calculate the area under the precision-recall curve (AUPRC) using the entropy of their output distribution.

We anticipate that Dirichlet-BM and DDU should perform better in differentiating between OOD and ID inputs than deterministic softmax methods. These models should exhibit higher uncertainty for OOD data points (e.g., corrupted images). By labeling ID data points from the default test set as '0' and OOD data points from corrupted test images as '1', we utilize the epistemic uncertainty quantification metric for each model to distinguish between these points. We assess the model's OOD detection performance using AUROC score.

Datasets: We employ two image datasets for robustness evaluation: CIFAR10/CIFAR100 and PatchCAM [10]. To evaluate against image corruptions, we leverage CIFAR-C [11] which is artificially generated by applying 13 types of corruptions to CIFAR dataset (Figure 4). Performance in a domain-shift setting is evaluated using the PatchCAM data set, which comprises medical

Table 1: Performance of single-pass uncertainty estimation methods on image corruption and domain generalization

Datasets	Softmax				Dirichlet-BM				DDU			
	Prec@1	**AUPRC**	**AUROC**	**Entropy**	**Prec@1**	**AUPRC**	**AUROC**	**Entropy**	**Prec@1**	**AUPRC**	**AUROC**	**Entropy**
Image Corruptions												
CIFAR 10	**89.50**	76.85	-	0.14	88.45	**90.31**	-	0.63	88.11	89.32	-	0.84
CIFAR 10-C	41.05	31.42	73.65	0.36	**43.47**	**46.19**	**81.47**	1.03	28.26	22.40	55.41	0.37
CIFAR 100	63.05	45.00	-	0.52	**67.48**	**67.97**	-	3.61	65.36	47.10	-	2.93
CIFAR 100-C	10.93	7.55	74.32	1.14	**12.22**	**12.43**	**83.51**	4.29	6.91	4.72	40.90	1.30
Domain Shift												
PatchCAM	55.66	**58.11**	-	0.23	53.46	56.55	-	0.42	**59.90**	55.02	-	0.18

histopathology images across 5 hospitals comprising of tumor and normal classes (Figure 4). This domain-shift is due to procedural variations (tissue slide staining and scanner) between different hospitals. All models are trained using in-distribution data.

6 Results

We highlight the performance of the two uncertainty estomation methods for evaluating the three hypotheses in Table 1.

6.1 Classification Performance

We analyse the classification accuracy of the models using Prec@1 metric. Dirichlet network (Dirichlet-BM) outperforms DDU and vanilla softmax network on corrupted image datasets (CIFAR 10-C, CIFAR 100-C) (Table 1). While Dirichlet network achieves slightly lower classification accuracy on CIFAR10, Dirichlet-BM outperforms all approaches on clean and corrupted CIFAR100 dataset. Moreover, while the accuracy of each approach drops in the presence of image corruptions (CIFAR10-C, CIFAR100-C), the reduction for Dirichlet-BM is lesser than DDU and softmax approach. This highlights that principled formulation of Dirichlet-BM enforcing a distributional prior over classifier's output and leveraging KL-divergence to align the posterior and prior distributions plays an important role as a regularizer during training. DDU performs suboptimally across both CIFAR datasets compared to Dirichlet-BM and softmax approach. In contrast to the claims made by DDU's paper [8], we find that DDU does not generalize well on OoD data. Interestingly in the domain shift setting (PatchCAM dataset), Dirichlet-BM underperforms softmax approach which might be due to the KL-divergence based regularization in the variational inference formulation. DDU outperforms both approaches on PatchCAM dataset, highlighting the impact of spectral normalization on learning an improved feature space.

6.2 Model Confidence Calibration

While deterministic neural networks are expected to suffer from suboptimal performance on OOD data, the output uncertainty is expected to be better calibrated for uncertainty-aware methods. Our hypothesis is that predictions with lower uncertainty have a higher likelihood of being correct than predictions with high uncertainty. We computed the epistemic uncertainty scores for test samples (both OOD and ID data) and use them to distinguish between correctly classified samples (low entropy, class = 0) and wrongly classified samples (high entropy, class = 1). In line with previous study [6], we leverage Area under Precision Recall curve (AUPRC) scores to assess model calibration. We observe that while both uncertainty estimation-based methods are better calibrated than softmax networks on ID data (CIFAR10, CIFAR100). While the calibration performance is suboptimal for Dirichlet-BM and DDU on OOD data (CIFAR10-C, CIFAR100-C), Dirichlet-BM achieves higher AUPRC than softmax approach and DDU. DDU has the worse calibration on OOD data in spite of achieving higher Prec@1 than softmax approach. This highlights that improved classification performance on ID data does not imply that the model is well calibrated. In domain-shift setting, we observe that softmax network is better calibrated although it is slightly overconfident (lower average entropy) compared to Dirichlet network.

6.3 Out-of-distribution Uncertainty

To quantify OOD uncertainty, we leverage epistemic uncertainty, implying that in case the neural network is not confident about the output, it will predict a distribution with higher entropy vs low entropy when the neural network is confident about its predictions. When the Dirichlet network is evaluated on OOD data, it should exhibit peaks of entropy distribution in high uncertainty regions. On the other hand, softmax should predict relatively flat uncertainty for OOD samples. DDU leverages distances from GMM class centroids in the ID feature space to compute the entropy for OOD data. From Table 1, we observe that while softmax network and Dirichlet-BM predict higher average entropy for OOD data(CIFAR10-C, CIFAR100-C), DDU surprisingly has lower entropy for OOD data. The increase in average entropy for Dirichlet-BM is higher than softmax network for CIFAR10. For CIFAR100, Dirichlet networks have higher uncertainty on clean ID data and the increase in entropy on CIFAR100-C is slightly lesser than softmax-based models. Both uncertainty estimation methods have higher entropy than softmax approach on ID test data (CIFAR10, CIFAR100). To evaluate the distribution-shift in entropy with OOD samples, we plan to incorporate the density plots of epistemic uncertainty scores for Dirichlet network and DDU compared to softmax network.

We evaluate the OoD detection accuracy using AUROC scores. By labeling ID inputs as '0' and OoD inputs as '1', we compute AUROC using the epistemic uncertainty score for each model. We find that Dirichlet network significantly outperforms all approaches on OoD detection while DDU performs worse than the softmax-based approach. This is in contrast to DDU's paper which claims better OoD detection performance and we plan to investigate this further.

7 Conclusion

Our work analyzes the robustness of uncertainty estimation by Dirichlet networks and Feature space distance aware methods and their generalization performance in two common settings (image corruption and domain-shift). Our results show that Dirichlet networks achieve better ID and OOD accuracy on natural images but underperform on clinical images in domain-shift setting. Moreover, the uncertainty estimates in Dirichlet networks are better calibrated compared to softmax classifier as well as the feature-space distance approach. We surprisingly find that ensuring distance awareness on the feature space and fitting a GMM to model the learnt feature density leads to worse calibration and classification accuracy on OOD dataset. This is in contrast to DDU's paper findings and we hypothesize it could be because fitting a class-conditioned GMM on the training data assumes translation-invariant distribution which factories along the spatial dimensions in the feature space. Our next steps include analyzing DDU's feature space in detail and exploring fitting GMM on CIFAR10/CIFAR100 feature space with lesser number of classes. Moreover, we plan to evaluate SNGP [5] which is a more principled approach and involves learning a Gaussian process layer at classifier's output along with enforcing smoothness in feature space. We also plan to benchmark various Dirichlet networks and distance-aware methods across more real-world domain generalization scenarios.

References

[1] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2755–2764, 2017.

[2] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[3] Dennis Ulmer. A survey on evidential deep learning for single-pass uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

[4] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv preprint arXiv:2102.11582*, 2021.

[5] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

[6] Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *International Conference on Machine Learning*, pages 5707–5718. PMLR, 2021.

[7] Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being bayesian about categorical probability. In *International Conference on Machine Learning*, pages 4950–4961. PMLR, 2020.

[8] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.

[9] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

[10] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.