Density-Softmax: Efficient Test-time Model for Uncertainty Estimation and Robustness under Distribution Shifts

Ha Manh Bui, Angie Liu

Department of Computer Science, Johns Hopkins University

International Conference on Machine Learning, July 2024

Table of contents

Table of contents

Problem & Motivation

Method	Uncertainty quality	Robustness quality	Test-time efficiency	Without prior requirement
Deterministic	✗	✗	✓	✓
Bayesian	✓	✗	✓	✗
Ensembles	✓	✓	✗	✓
Ours	✓	✓	✓	✓

Table: Comparison in uncertainty, robustness quality, test-time efficiency (lightweight & fast), and whether pre-defined prior hyper-parameters are required.

- Deterministic DNN is often non-robust & over-confident, Bayesian NN additionally requires pre-defined prior hyperparams, and Ensembles suffers from high storage & slow speed in inference (test-time).

Problem & Motivation

Method	Uncertainty quality	Robustness quality	Test-time efficiency	Without prior requirement
Deterministic	✗	✗	✓	✓
Bayesian	✓	✗	✓	✗
Ensembles	✓	✓	✗	✓
Ours	✓	✓	✓	✓

Table: Comparison in uncertainty, robustness quality, test-time efficiency (lightweight & fast), and whether pre-defined prior hyper-parameters are required.

- Deterministic DNN is often non-robust & over-confident, Bayesian NN additionally requires pre-defined prior hyperparams, and Ensembles suffers from high storage & slow speed in inference (test-time).
- Density-Softmax: leverages the density function built on a Lipschitz-constrained feature extractor with the softmax layer for better uncertainty estimation, robustness, and fast inference.

Table of contents

Density-Softmax: Setting

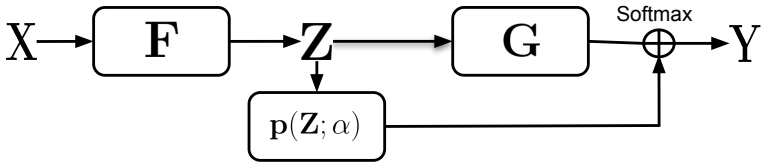


Figure: Architecture includes encoder f, regressor g, and density function $p(Z;\alpha)$.

Setting. Consider we predict a target label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is discrete with K possible categories by using a forecast $h = \sigma(g \circ f)$, which composites a feature extractor $f : \mathcal{X} \to \mathcal{Z}$, a classifier $g : \mathcal{Z} \to \mathbb{R}^K$, and a softmax layer $\sigma : \mathbb{R}^K \to \Delta_y$ which outputs a probability distribution $W(y) : \mathcal{Y} \to [0, 1]$ within the set Δ_y over $\mathcal{Y}$.

Density-Softmax

Robustness. Optimize h by using ERM and gradient-penalty regularization from training data D_s by solving

$$\min_{\theta_{g,f}} \{\mathbb{E}_{(x,y)\sim D_s}[-y \log(\sigma(g(f(x)))) + \lambda(||\nabla_x f(x)||_2 - 1)^2]\}, \quad (1)$$

where $\theta_{g,f}$ is the parameter of encoder f and classifier g, λ is the gradient-penalty coefficient, and $||\nabla_x f(x)||_2$ is the Spectral norm of the Jacobian matrix $\nabla_x f(x)$.

Density-Softmax

Robustness. Optimize h by using ERM and gradient-penalty regularization from training data D_s by solving

$$\min_{\theta_{g,f}} \{\mathbb{E}_{(x,y)\sim D_s}[-y \log(\sigma(g(f(x)))) + \lambda(||\nabla_x f(x)||_2 - 1)^2]\}, \quad (1)$$

where $\theta_{g,f}$ is the parameter of encoder f and classifier g, λ is the gradient-penalty coefficient, and $||\nabla_x f(x)||_2$ is the Spectral norm of the Jacobian matrix $\nabla_x f(x)$.

Remark 1. The gradient-penalty enforces $\sup_{x\in\mathbb{R}^n} ||\nabla_x f(x)||_2 = 1$, suggests $f(x)$ satisfy $||f(x_1) - f(x_2)||_2 \leq ||x_1 - x_2||_2$. This 1-Lipstchiz $f(x)$ is proved to be robust on corruptions by the Local Robustness Certificates.

Density-Softmax

Robustness. Optimize h by using ERM and gradient-penalty regularization from training data D_s by solving

$$\min_{\theta_{g,f}} \{\mathbb{E}_{(x,y)\sim D_s}[-y \log(\sigma(g(f(x)))) + \lambda(||\nabla_x f(x)||_2 - 1)^2]\}, \tag{1}$$

where $\theta_{g,f}$ is the parameter of encoder f and classifier g, λ is the gradient-penalty coefficient, and $||\nabla_x f(x)||_2$ is the Spectral norm of the Jacobian matrix $\nabla_x f(x)$.

Remark 1. The gradient-penalty enforces $\sup_{x\in\mathbb{R}^n} ||\nabla_x f(x)||_2 = 1$, suggests $f(x)$ satisfy $||f(x_1) - f(x_2)||_2 \leq ||x_1 - x_2||_2$. This 1-Lipstchiz $f(x)$ is proved to be robust on corruptions by the Local Robustness Certificates.

Uncertainty. Integrate density function $p(Z;\alpha)$ with classifier g by

$$p(y = i|x_t) = \frac{\exp(p(z_t;\alpha)\cdot(z_t^\top \theta_{g_i}))}{\sum_{j=1}^{K}\exp(p(z_t;\alpha)\cdot(z_t^\top \theta_{g_j}))}, \forall i \in \mathcal{Y}, \tag{2}$$

where $z_t = f(x_t)$ is the feature of test sample x_t.

Density-Softmax

Robustness. Optimize h by using ERM and gradient-penalty regularization from training data D_s by solving

$$\min_{\theta_{g,f}} \{\mathbb{E}_{(x,y)\sim D_s}[-y \log(\sigma(g(f(x)))) + \lambda(||\nabla_x f(x)||_2 - 1)^2]\}, \quad (1)$$

where $\theta_{g,f}$ is the parameter of encoder f and classifier g, λ is the gradient-penalty coefficient, and $||\nabla_x f(x)||_2$ is the Spectral norm of the Jacobian matrix $\nabla_x f(x)$.

Remark 1. The gradient-penalty enforces $\sup_{x\in\mathbb{R}^n} ||\nabla_x f(x)||_2 = 1$, suggests $f(x)$ satisfy $||f(x_1) - f(x_2)||_2 \leq ||x_1 - x_2||_2$. This 1-Lipstchiz $f(x)$ is proved to be robust on corruptions by the Local Robustness Certificates.

Uncertainty. Integrate density function $p(Z;\alpha)$ with classifier g by

$$p(y = i|x_t) = \frac{\exp(p(z_t;\alpha)\cdot(z_t^\top \theta_{g_i}))}{\sum_{j=1}^{K}\exp(p(z_t;\alpha)\cdot(z_t^\top \theta_{g_j}))}, \forall i \in \mathcal{Y}, \quad (2)$$

where $z_t = f(x_t)$ is the feature of test sample x_t.

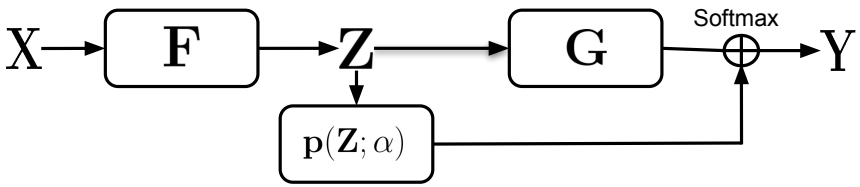


Figure: Architecture includes encoder f, regressor g, and density function $p(Z;\alpha)$.

Algorithm 4 Sketch Algorithm

1: **Train-time:**
2: Pre-train h by ERM and gradient-penalty regularization in Eq. 1.
3: Freeze f, then estimate $p(z;\alpha)$ on feature space $\mathcal{Z}$.
4: Re-update g by ERM in Eq. 1 with $p(z;\alpha)$.
5: **Test-time:**
6: **Infer** $\hat{y}_t$ for x_t by Eq. 2 with only a single forward pass.

Density-Softmax

Robustness. Optimize h by using ERM and gradient-penalty regularization from training data D_s by solving

$$\min_{\theta_{g,f}} \{\mathbb{E}_{(x,y)\sim D_s}[-y \log(\sigma(g(f(x)))) + \lambda(||\nabla_x f(x)||_2 - 1)^2]\}, \quad (1)$$

where $\theta_{g,f}$ is the parameter of encoder f and classifier g, λ is the gradient-penalty coefficient, and $||\nabla_x f(x)||_2$ is the Spectral norm of the Jacobian matrix $\nabla_x f(x)$.

Remark 1. The gradient-penalty enforces $\sup_{x\in\mathbb{R}^n} ||\nabla_x f(x)||_2 = 1$, suggests $f(x)$ satisfy $||f(x_1) - f(x_2)||_2 \leq ||x_1 - x_2||_2$. This 1-Lipstchiz $f(x)$ is proved to be robust on corruptions by the Local Robustness Certificates.

Uncertainty. Integrate density function $p(Z;\alpha)$ with classifier g by

$$p(y = i|x_t) = \frac{\exp(p(z_t;\alpha)\cdot(z_t^\top \theta_{g_i}))}{\sum_{j=1}^{K}\exp(p(z_t;\alpha)\cdot(z_t^\top \theta_{g_j}))}, \forall i \in \mathcal{Y}, \quad (2)$$

where $z_t = f(x_t)$ is the feature of test sample x_t.

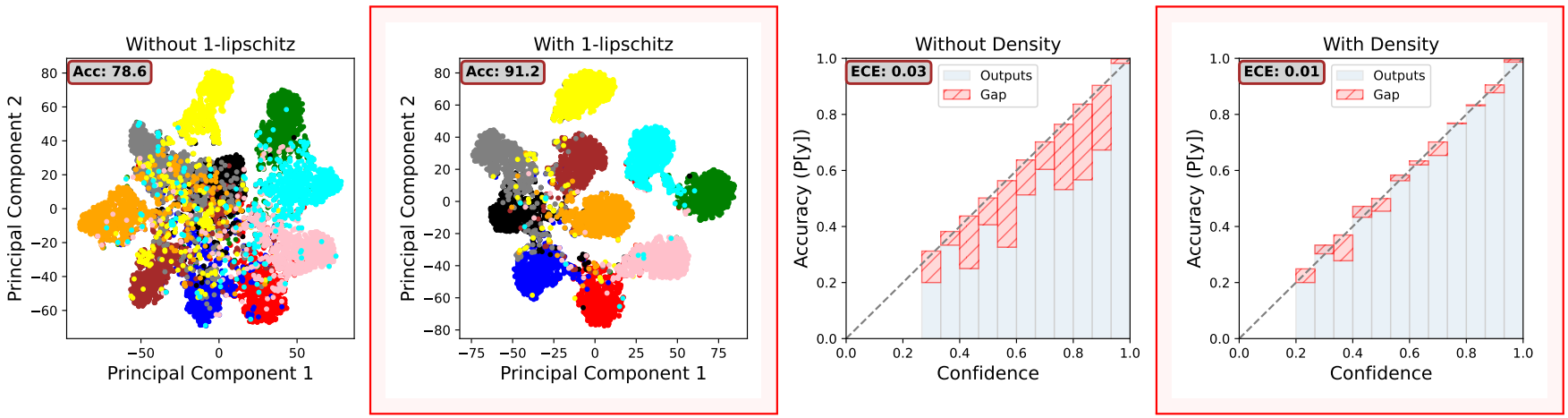


Figure: Feature visualizations between models w/o & w 1-Lipschitz constraint (2 lefts), reliability diagrams between models w/o & w the density-function (2 rights).

Table of contents

Theoretical results

Theorem 1. Density-Softmax's prediction is the **optimal solution of the minimax uncertainty risk**, i.e., $\sigma(p(f(X);\alpha)$ $g(f(X)) = \underset{\mathbb{P}(Y|X)\in\mathcal{P}}{\arg\inf} \left[\sup_{\mathbb{P}^*(Y|X)\in\mathcal{P}^*} S(\mathbb{P}(Y|X), \mathbb{P}^*(Y|X))\right]$.

Theorem 2. The predictive distribution of Density-Softmax $\sigma(p(z = f(x);\alpha) \cdot (g \circ f(x)))$ is **distance aware on the feature space** $\mathcal{Z}$, i.e., $\exists$ a summary statistic $u(z_t)$ of $\sigma(p(z_t;\alpha) \cdot (g \circ z_t))$ on the new test feature $z_t = f(x_t)$ s.t., $u(z_t) = v(d(z_t, Z_s))$, where v is a monotonic function and $d(z_t, Z_s) = \mathbb{E}\,\|z_t - Z_s\|_{\mathcal{Z}}$ is the distance between z_t and the training features random variable Z_s.

Theoretical results

Theorem 1. Density-Softmax's prediction is the **optimal solution of the minimax uncertainty risk**, i.e., $\sigma(p(f(X);\alpha)$ $g(f(X)) = \underset{\mathbb{P}(Y|X)\in\mathcal{P}}{\arg\inf} \left[\sup_{\mathbb{P}^*(Y|X)\in\mathcal{P}^*} S(\mathbb{P}(Y|X), \mathbb{P}^*(Y|X))\right]$.

Theorem 2. The predictive distribution of Density-Softmax $\sigma(p(z = f(x);\alpha) \cdot (g \circ f(x)))$ is **distance aware on the feature space** $\mathcal{Z}$, i.e., $\exists$ a summary statistic $u(z_t)$ of $\sigma(p(z_t;\alpha) \cdot (g \circ z_t))$ on the new test feature $z_t = f(x_t)$ s.t., $u(z_t) = v(d(z_t, Z_s))$, where v is a monotonic function and $d(z_t, Z_s) = \mathbb{E}\,\|z_t - Z_s\|_{\mathcal{Z}}$ is the distance between z_t and the training features random variable Z_s.

Remark 2. When the likelihood of $p(Z;\alpha)$ is high, our model is certain on IID data, and when the likelihood of $p(Z;\alpha)$ decreases on OOD data, the certainty will decrease correspondingly, $\cdots$

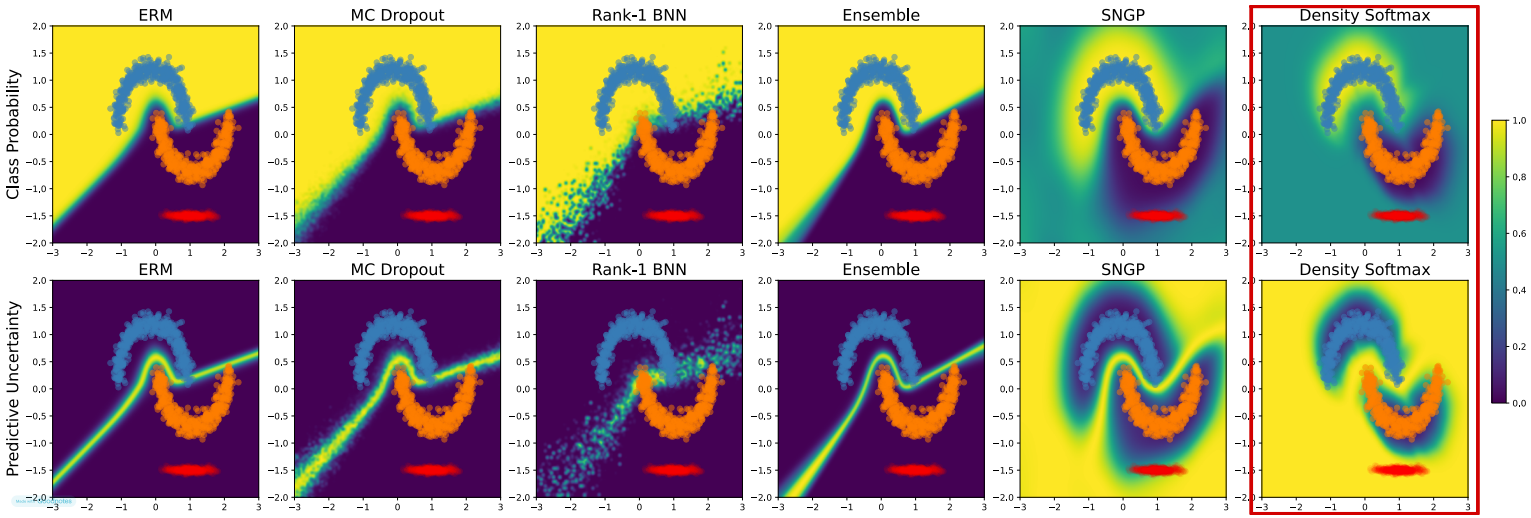


Figure: Density-Softmax achieves **distance awareness**, has confident predictions on IID data for positive (Orange) and negative classes (Blue), and decreases certainty to a uniform class probability when OOD data is far from the training set.

Theoretical results

Theorem 1. Density-Softmax's prediction is the **optimal solution of the minimax uncertainty risk**, i.e., $\sigma(p(f(X);\alpha)g(f(X)) = \underset{\mathbb{P}(Y|X)\in\mathcal{P}}{\arg\inf}\left[\sup_{\mathbb{P}^*(Y|X)\in\mathcal{P}^*} S(\mathbb{P}(Y|X),\mathbb{P}^*(Y|X))\right]$.

Theorem 2. The predictive distribution of Density-Softmax $\sigma(p(z = f(x);\alpha)\cdot(g\circ f(x)))$ is **distance aware on the feature space** $\mathcal{Z}$, i.e., $\exists$ a summary statistic $u(z_t)$ of $\sigma(p(z_t;\alpha)\cdot(g\circ z_t))$ on the new test feature $z_t = f(x_t)$ s.t., $u(z_t) = v(d(z_t, Z_s))$, where v is a monotonic function and $d(z_t, Z_s) = \mathbb{E}\,\|z_t - Z_s\|_{\mathcal{Z}}$ is the distance between z_t and the training features random variable Z_s.

Remark 2. When the likelihood of $p(Z;\alpha)$ is high, our model is certain on IID data, and when the likelihood of $p(Z;\alpha)$ decreases on OOD data, the certainty will decrease correspondingly, $\cdots$ improving calibration of the standard softmax by reducing its over-confidence as follows:

Proposition 1. If the predictive distribution of the standard softmax $\sigma(g\circ f)$ makes $\text{acc}(B_m) \leq \text{conf}(B_m), \forall B_m, m \in [M]$, where B_m is the set of sample indices whose confidence falls into $\left(\frac{m-1}{M}, \frac{m}{M}\right]$ in M bins, then Density-Softmax $\sigma((p(f;\alpha)\cdot g)\circ f)$ can **improve calibrated-uncertainty** by $\text{ECE}(\sigma((p(f;\alpha)\cdot g)\circ f) \leq \text{ECE}(\sigma(g\circ f))$.

Empirical results

Density-Softmax achieves a competitive robust generalization and uncertainty estimation performance with SOTA across different datasets and modern DNN architectures.

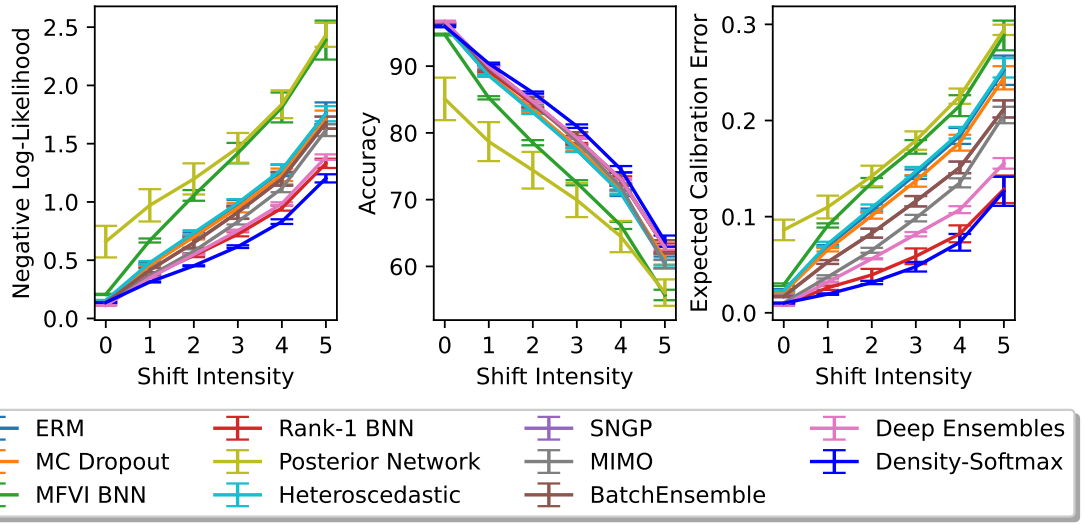


Figure: Benchmark performance on CIFAR-10 with Wide Resnet-28-10.

Method	NLL(↓)	Acc(↑)	ECE(↓)	cNLL(↓)	cAcc(↑)	cECE(↓)	AUPR-S(↑)	AUPR-C(↑)
Rank-1 BNN	0.692	81.3	**0.018**	2.24	53.8	0.117	0.884	0.797
SNGP	0.805	80.2	0.020	2.02	54.6	0.092	**0.923**	0.801
Deep Ensembles	**0.666**	**82.7**	0.021	2.27	54.1	0.138	0.888	0.780
Density-Softmax	0.780	80.8	0.038	**1.96**	**54.7**	**0.089**	0.910	**0.804**

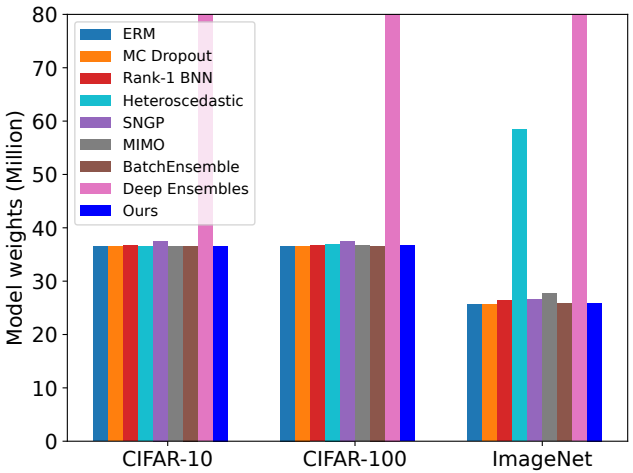
Table: Results for Wide Resnet-28-10 on CIFAR-100.

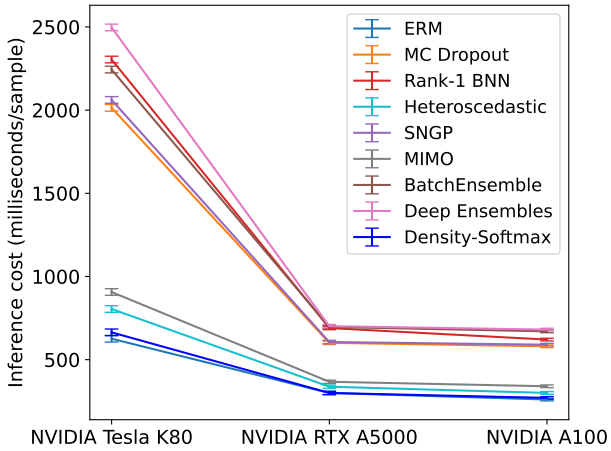
Method	NLL(↓)	Acc(↑)	ECE(↓)	cNLL(↓)	cAcc(↑)	cECE(↓)	#Params(↓)	Latency(↓)
Deterministic ERM	0.939	76.2	0.032	3.21	40.5	0.103	**25.61M**	**299.81**
Rank-1 BNN	0.886	77.3	0.017	2.95	42.9	0.054	26.35M	690.14
Heteroscedastic	0.898	77.5	0.033	3.20	42.4	0.111	58.39M	337.50
SNGP	0.931	76.1	**0.013**	3.03	41.1	0.045	26.60M	606.11
MIMO	0.887	77.5	0.037	3.03	43.3	0.106	27.67M	367.17
BatchEnsemble	0.922	76.8	0.037	3.09	41.9	0.089	25.82M	696.81
Deep Ensembles	**0.857**	**77.9**	0.017	2.82	**44.9**	0.047	102.44M	701.34
Density-Softmax	0.885	77.5	0.019	**2.81**	44.6	**0.042**	25.88M	**299.90**

Table: Results for Resnet-50 on ImageNet.

Empirical results

Importantly, our method has fewer parameters and is much faster than other baselines.





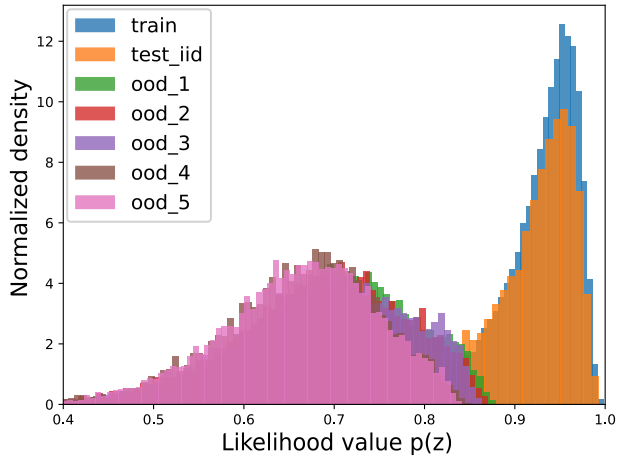


Figure: Storage requirement & inference cost comparison at test time (2 lefts); Histogram of $p(z; \alpha)$'s likelihood, train on CIFAR-10, test on CIFAR-10-C.

Table of contents

Summary

Contributions:

- Introduce Density-Softmax, a reliable, sampling-free, and single DNN framework via a direct combination of a density function built on a Lipschitz-constrained feature extractor with the softmax layer. It is fast & lightweight and can be implemented efficiently and easily across DNN architectures.
- Formally prove that our model is the solution to the minimax uncertainty risk, distance awareness on the feature space, and can reduce over-confidence of the standard softmax when the test feature is far from the training set.
- Empirically shows it achieves robust generalization and uncertainty estimation performance with SOTA across different datasets and modern DNN architectures. Importantly, it has fewer parameters and is much faster than other baselines at test time.

For more information:

- PDF, code available at *https://openreview.net/forum?id=lon750Kf7n*
- **Come see our poster!**

See you at the ICML 2024 conference!