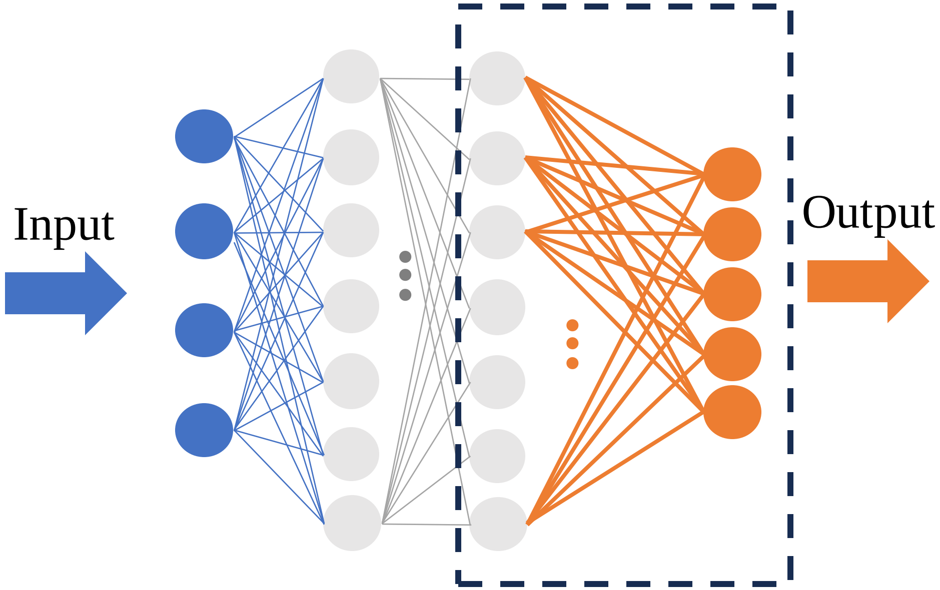
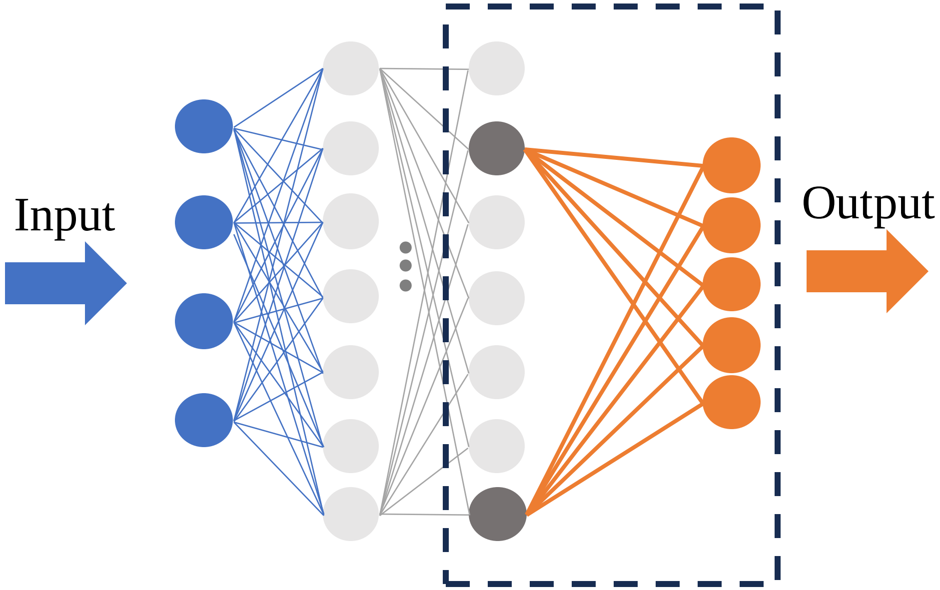
Enhancing Uncertainty Estimation and Interpretability via Bayesian Non-Negative Decision Layer

Motivation

Interpretable via Sparsity

DNN: Deep feature extractor + Decision layer

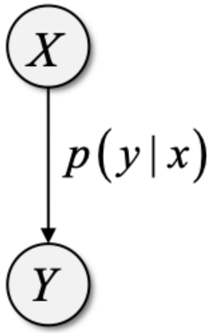




- Dense decision layer: millions of parameters operating on thousands of deep features
- **Sparse** decision later: inspecting only the few linear coefficients and deep features that dictate its predictions

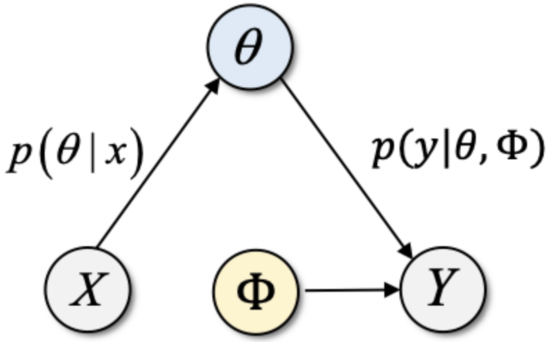
Motivation

Uncertainty Estimation via Bayesian Network

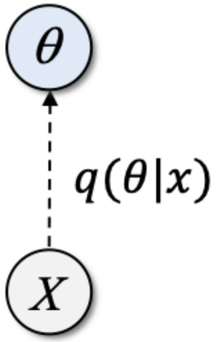


Graphical Model of DNNs

Introducing stochastic latent variables



(c) Generative Model of BNDL



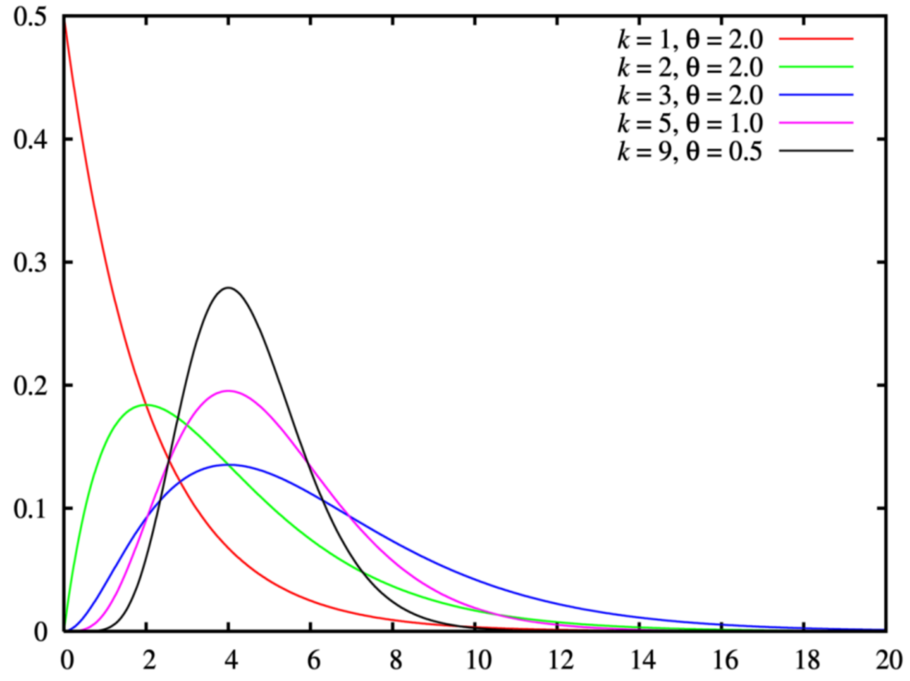
(d) Inference Model of BNDL

$$\boldsymbol{y}_j \mid \boldsymbol{\theta}_j \sim \text{Category}\left(\boldsymbol{\theta}_j \boldsymbol{\Phi}\right), \boldsymbol{\theta}_j \mid \boldsymbol{x}_j \sim \text{Gamma}\left(f_\theta(x_j), 1\right), \boldsymbol{\Phi} \sim \text{Gamma}\left(1, 1\right).$$

Bayesian Non-negative Decision Layer

Theoretical Guarantees

$$\boldsymbol{y}_j \mid \boldsymbol{\theta}_j \sim \text{Category}\left(\boldsymbol{\theta}_j \boldsymbol{\Phi}\right), \boldsymbol{\theta}_j \mid \boldsymbol{x}_j \sim \text{Gamma}\left(f_\theta(x_j), 1\right), \boldsymbol{\Phi} \sim \text{Gamma}\left(1, 1\right).$$



PDF of Gamma Distribution

Properties of Gamma Distribution

- Non-negativity
- Sparsity

Consistent with the **identifiability** proposition

Proposition 1 ((Gillis & Rajkó, 2023)). *The k-th column of $\boldsymbol{\theta}$ is identifiable under the two assumptions:*

- ***Selective Window**: There exists a row of $\boldsymbol{\Phi}$, say the j-th, such that $\boldsymbol{\Phi}(j,:) = \alpha e_{(k)}^T$ for $\alpha > 0$, where $e_{(k)}^T$ represents the k-th standard row vector in vector space.*
- ***Sparsity Constrain**: The k-th column of $\boldsymbol{\Phi}$ contains at least $r-1$ entries equal to zero, where r is the rank of Y.*

BNDL fit the above assumptions, and has proven to be

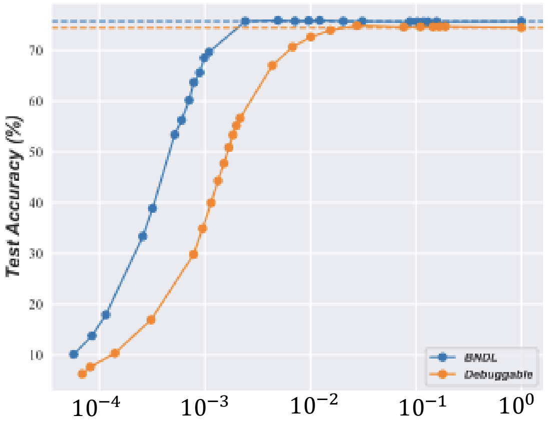
identifiable and interpretable

Experimental Results

Enhancing sparsity while maintaining model performance

Model	CIFAR-10		CIFAR-100		ImageNet-1k	
	ACC	PAvPU	ACC	PAvPU	ACC	PAvPU
ResNet	94.98 ± 0.12	-	74.62 ± 0.23	-	75.33 ± 0.14	-
MC Dropout	94.54 ± 0.03	78.83 ± 0.12	78.12 ± 0.06	64.41 ± 0.22	75.98 ± 0.08	76.50 ± 0.02
BM	94.07 ± 0.07	93.98 ± 0.3	75.81 ± 0.34	77.13 ± 0.67	-	-
CARD	90.93 ± 0.02	91.11 ± 0.04	71.42 ± 0.01	71.48 ± 0.03	76.20 ± 0.00	76.29 ± 0.01
ResNet-BNDL	**95.54 ± 0.08**	**95.58 ± 0.20**	**79.82 ± 0.13**	**81.1 ± 0.21**	**77.01 ± 0.14**	**77.66 ± 0.03**
ViT-Base	95.51 ± 0.03	-	84.15 ± 0.03	-	80.33	-
ViT-BNDL	**96.34 ± 0.04**	**97.01 ± 0.02**	**85.16 ± 0.03**	**86.37 ± 0.11**	**81.29 ± 0.02**	**82.50 ± 0.03**

➢ $ReLU(x-\alpha)$: dynamically pruning small weights



$$1 - Sparsity = \log\left(1 - \frac{WeightNum_{nonsparse}}{WeightNum}\right)$$

- Compared with other sparse decision layer method
  - ✓ Achieving higher accuracy under the same level of sparsity

Experimental Results

Interpretable Features

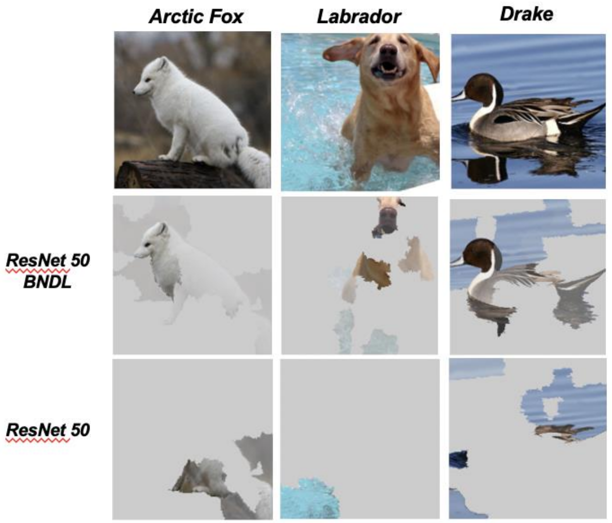
- **Non-negativity Constrain**

$$\boldsymbol{\theta}_j \,|\, \boldsymbol{x}_j \sim \mathrm{Gamma}\left(f_\theta(x_j), 1\right), \boldsymbol{\Phi} \sim \mathrm{Gamma}\left(1, 1\right).$$

Enforce θ and Φ to be non-negative

Ensures that different features will not cancel one another out

→ *Interpretable Features*

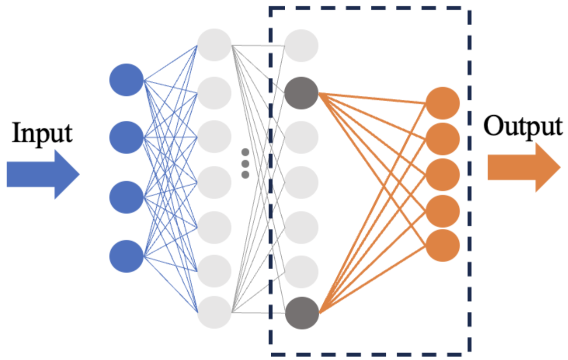


Feature Visualization of BNDL and Baseline Models

- BNDL's features align more closely with the semantic meaning of true labels
- More disentangled and interpretable features

Experimental Results

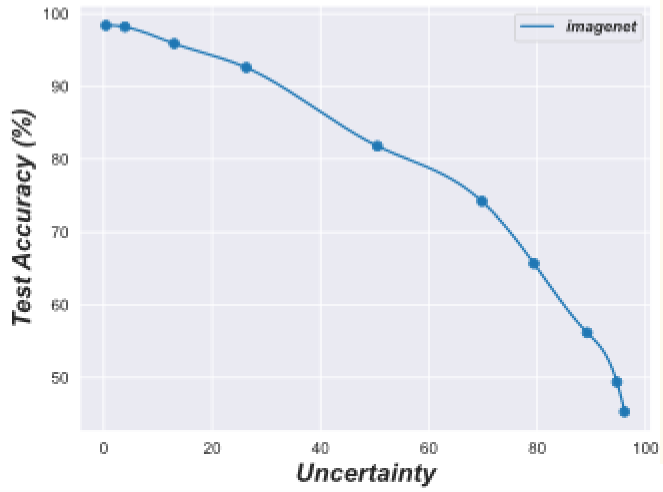
Uncertainty Estimation via Bayesian Last Layer

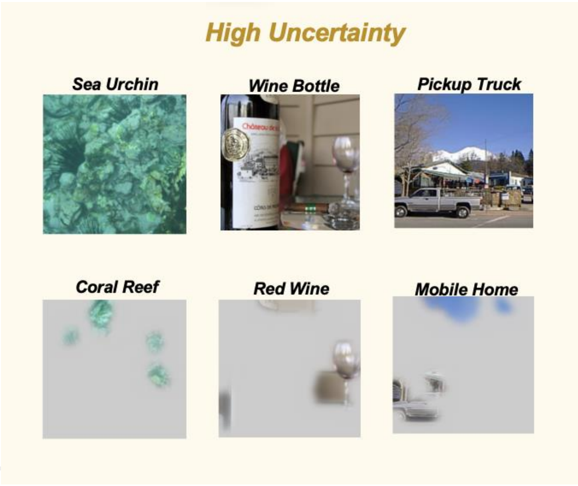


Only Modify the decision layer, Enabling uncertainty Estimation

Benefits: negligible added complexity

Visualizing High Uncertainty Sample





- ➢ negative correlation between uncertainty and accuracy
- ➢ samples with high uncertainty often belong to ambiguous classification boundary cases (e.g., misclassifying "**wine bottle**" as "**wine glass**")

Thank you for Listening