Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection

Nilesh Ahuja, Ibrahima J. Ndiour, Trushant Kalyanpur, and Omesh Tickoo
Intel Labs

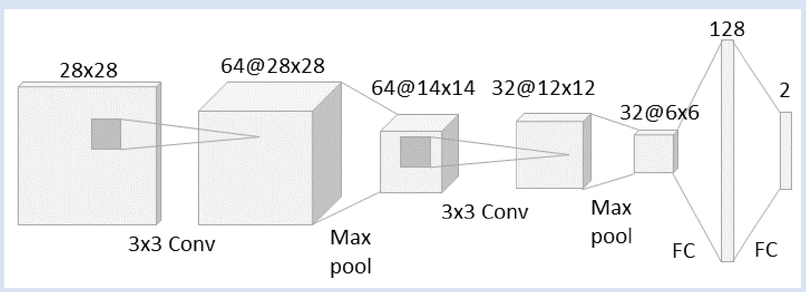
Introduction

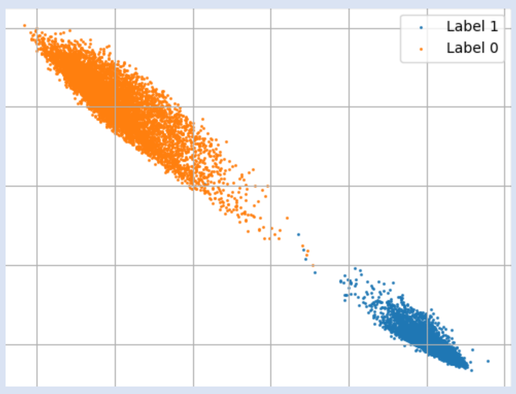
Softmax scores often result in overconfident predictions, especially when the input does not resemble the training data (out-of-distribution), or has been crafted to attack and "fool" the network (adversarial example).

We present an approach for detecting OOD and adversarial samples in DNNs based on probabilistic modeling of the deep-features. Our contributions include:

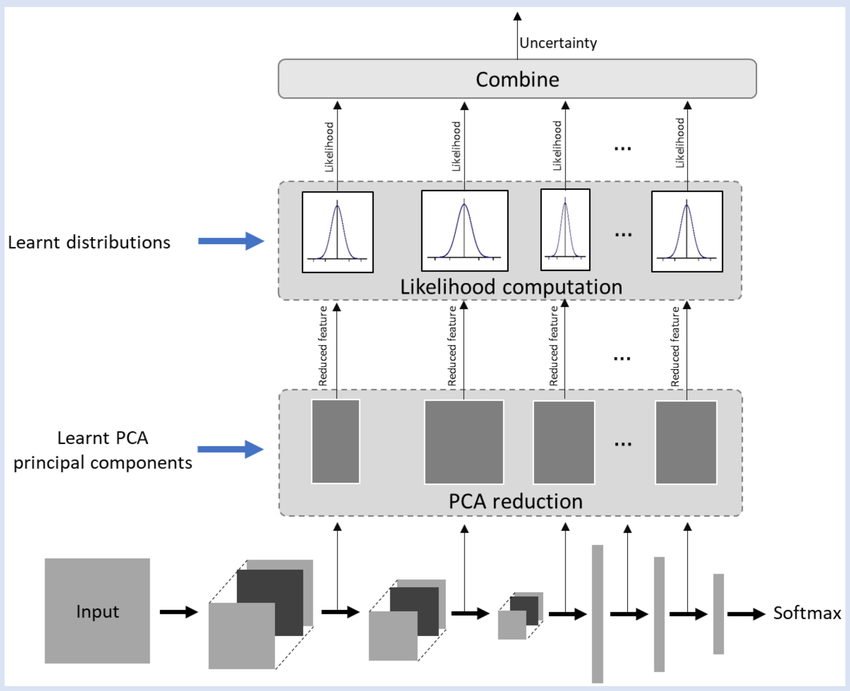
- Demonstrating that the high-dimensional features (of a DNN) actually reside in a low-dimensional subspace, that can accurately be captured with statistical dimensionality reduction techniques such as principal component analysis (PCA).
- Modeling the (embedded) deep features with parametric, class-conditional multivariate distributions (e.g. Gaussian, Gaussian mixture).
- Demonstrating that our method outperforms the state of the art by a substantial margin (up to 13 percentage points in AUROC and AUPR), while incurring negligible computational cost at inference.

Our Approach





Scatterplot of 2D features (logits) for a simple binary classifier example

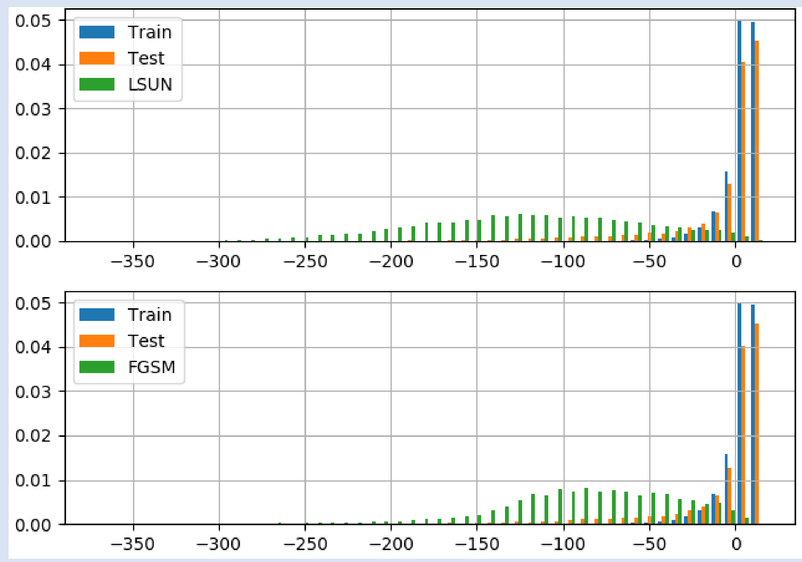


Block diagram of our approach

Experiments and Results

		SVHN			LSUN			FGSM	
	GMM	Separate	Tied	GMM	Separate	Tied	GMM	Separate	Tied
AUPR Layer 2	62.7	62.2	61.6	87.9	88	79.4	92.6	92.5	90.9
AUPR Layer 1	85.9	83.5	77.8	96.3	95.7	95	95	95.1	95.1
AUPR Layer 0	84.3	82.9	80.4	96.3	95.9	95.5	94.8	94.7	94.9
AUROC Layer 2	90.2	90.1	91.6	87.8	87.9	78.1	93.6	93.6	93.1
AUROC Layer 1	94.1	92.3	91.6	95.5	94.7	94.1	93.4	93.2	93.3
AUROC Layer 0	94.9	93.4	92.9	95.7	95.1	94.8	93	92.6	93

AUPR and AUROC scores for OOD and adversarial detection: GMM, Sep (Gaussian with separate covariance per class), Tied (Gaussian with tied covariance). Best values are shown in red.



Histogram of log-likelihood scores

	MNIST			CIFAR		
	GMM	Sep	Tied	GMM	Sep	Tied
Layer 0	98.9	98.6	98.6	90.8	90.8	92.2
Layer 1	98.2	98.6	98.6	95.0	95.3	95.3
Layer 2	86	97.4	98.3	95.3	95.2	95.3
		98.99			95.3	

Classification Accuracies using Log-likelihoods

Conclusions & Future Work

- We show that modeling deep-features with parametric probability distributions provides reliable uncertainty scores, which can enable reliable detection of OOD and adversarial samples as well as classification of in-distribution samples.
- Future work will seek to analyze the evolution of the feature distributions during training.

Fourth Workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada