MIT LINCOLN LABORATORY

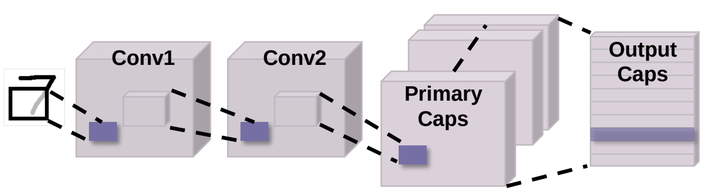
Kernelized Capsule Networks

Taylor Killian*, Justin Goodwin*, Olivia Brown, and Sung-Hyun Son

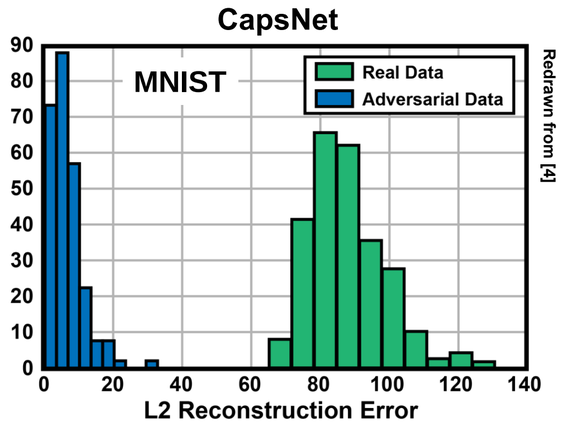
MIT Lincoln Laboratory

Background

Capsule Networks

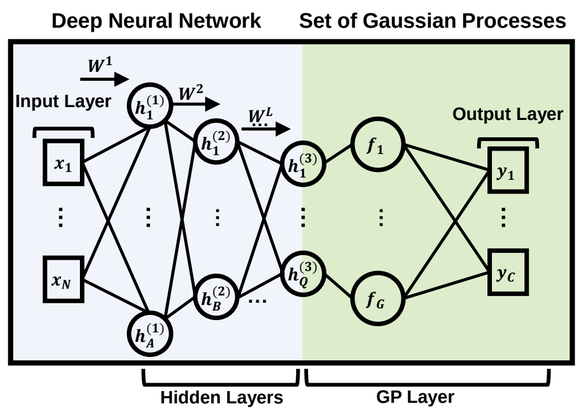


Capsule Networks [1], replace scalar neurons in DNNs with vector capsules, which encode spatial relationships between the learned features. CapsNets have shown promise in simple image classification.

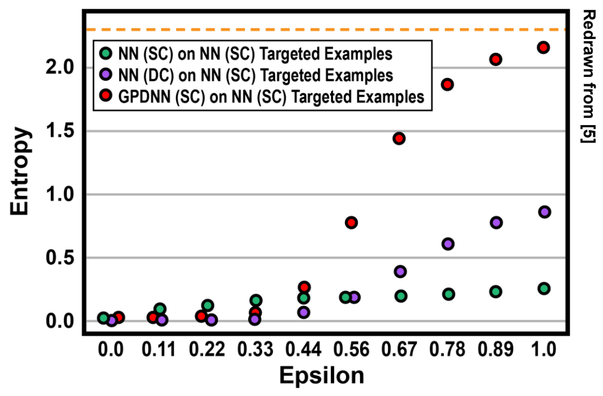


Frosst, et. al [4] demonstrated that a reconstruction network, originally used to regularize the CapsNet, can detect corrupted inputs by measuring the distance between the input image and reconstruction of the output capsule.

Deep Kernel Learning

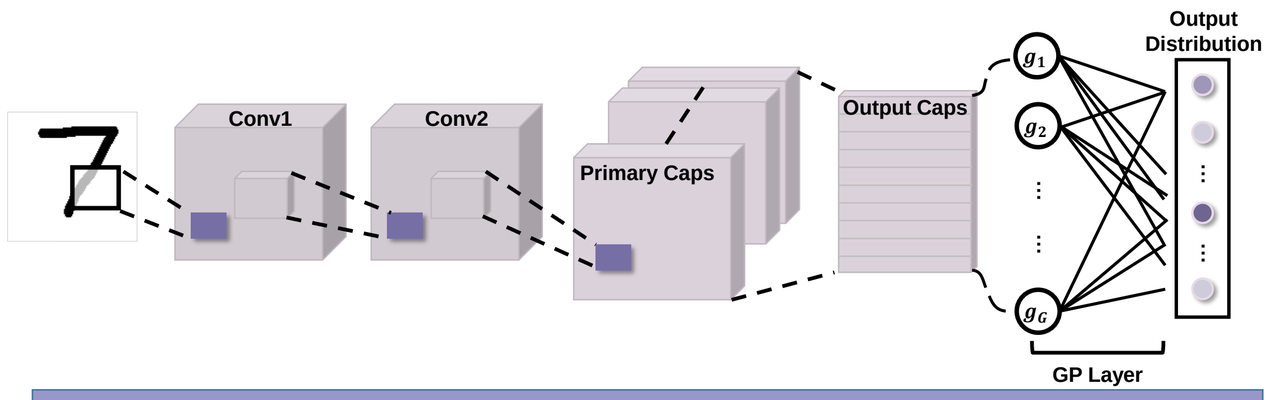


Building from [2], Wilson and Hu, et. al [3] introduced Deep Kernel Learning (DKL), an innovative hybrid GP and DNN architecture that leverages the rich features of a DNN to flexibly construct a GP kernel function.



Bradshaw, et. al [5] showed that these hybrid DKL architectures can provide a means to detect Adversarial Examples and improve robustness through the entropy of the Multivariate Normal Distribution constructed by the GP.

Kernelized Capsule Network (KCN)



Objective:

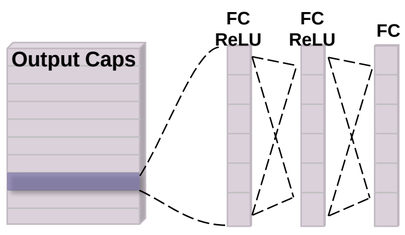
Utilize DKL to construct a kernel covariance function from the output capsule feature representations to provide:

1) Robustness to Adversarial Examples

2) Detection of Adversarial Examples

3) End-to-end training of CapsNet with marginal likelihood

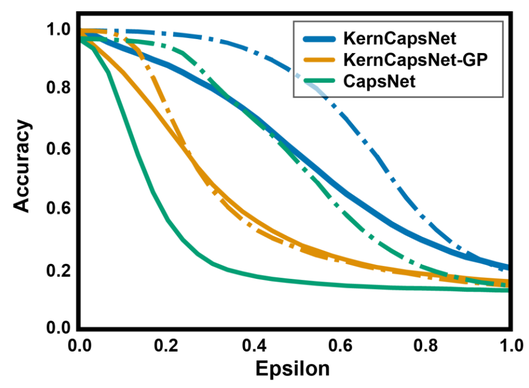
Reconstruction Network

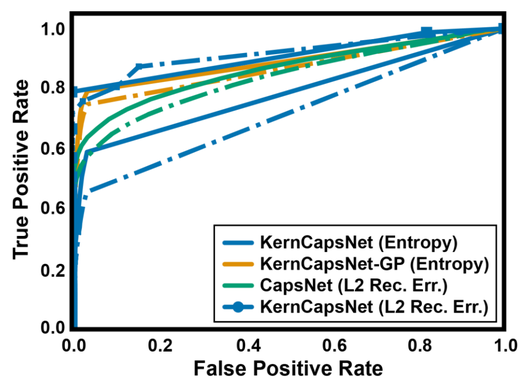


- **Robust architectures should have the capability to identify when observations have been corrupted**
- **Reconstruction error provides a signal that features are unreliable for classification and possibly corrupt**
- **Entropy of the posterior distribution derived from the learned GP can also signal uncertainty in classification decision**

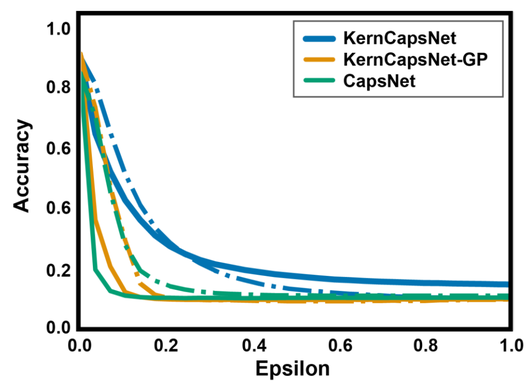
Experimental Results

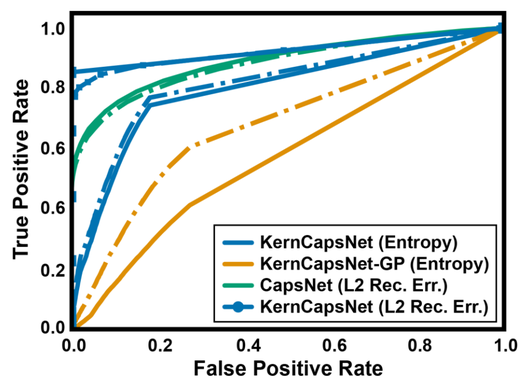
MNIST



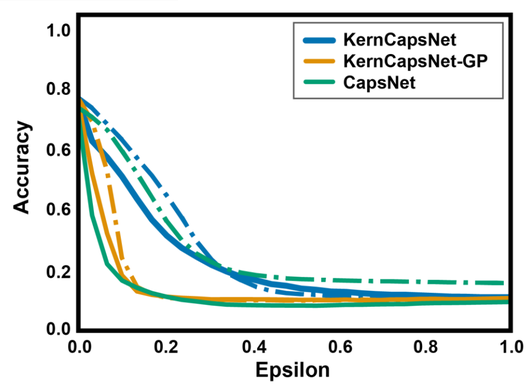


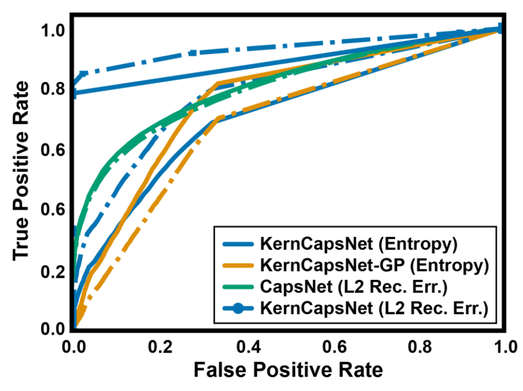
SVHN





CIFAR10





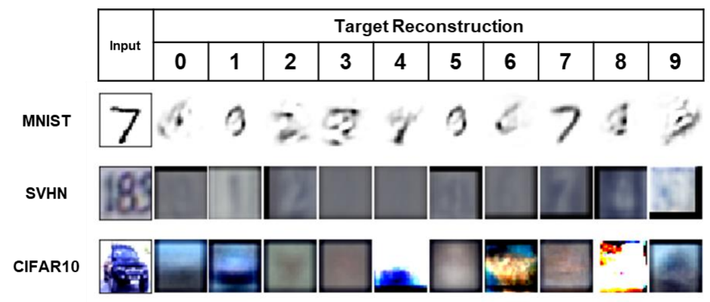
When comparing the KCN with the CapsNet and an ablated KCN that forgoes the reconstruction network (KCN-GP), it is clear that the KCN is more robust to Adversarial Examples (left column) and provides the most effective mechanism to detect when inputs are corrupted [right column; KCN (L2)]. Solid and dashed lines represent white and black box attacks, respectively.

Adversarial Example Detection Results (AUC)

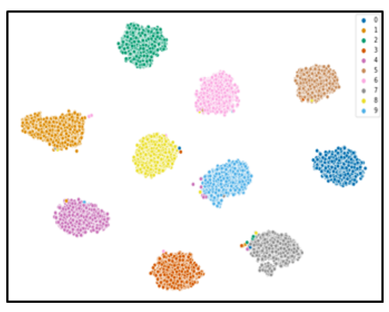
	MNIST		SVHN		CIFAR10	
	White Box	Black Box	White Box	Black Box	White Box	Black Box
CapsNet (L2)	0.8616	0.8344	0.8907	0.8813	0.8073	0.7992
KCN (L2)	0.9072	0.9160	0.9266	0.9229	0.8915	0.9350
KCN (Entropy)	0.7806	0.7134	0.7935	0.8132	0.7062	0.7843
KCN-GP (Entropy)	0.8860	0.8631	0.5688	0.6758	0.7580	0.6844

Reconstruction

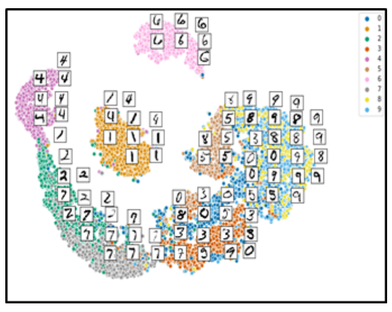
Capsules display "attention" behavior and key in on certain invariant properties such as "stroke" or "curvature" for MNIST, while focusing on color and gradients for SVHN and CIFAR10.



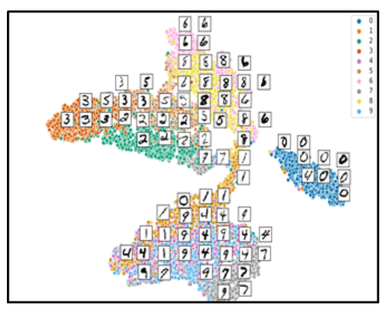
Visualizing Learned Capsules



Full Output Capsule Feature



1st Output Capsule Feature



3rd Output Capsule Feature

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

References:

[1] Sabour, S., et. al, "Dynamic routing between capsules", [NeurIPS, 2017]

[2] Hensman, J., et. al, "Scalable Variational Gaussian Process Classification", [AISTATS, 2015]

[3] Wilson, A., Hu, Z., et. al, "Stochastic Variational Deep Kernel Learning", [NeurIPS, 2016]

[4] Frosst, N., et. al, "DARCCC: Detecting Adversaries…", [arXiv:1811.06969]

[5] Bradshaw, J., et. al, "Adversarial examples, uncertainty, and transfer…", [arXiv:1707.02476]