cases	doc_1	doc_2	decision	id
			DUPLICATES	158

doc_1

authors	• Nilesh A. Ahuja • Ibrahima J. Ndiour • Trushant Kalyanpur • Omesh Tickoo
title	Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection
publication_date	2019-09-25 00:00:00
source	SupportedSources.OPENALEX
journal	arXiv (Cornell University)
volume	
doi	None
urls	• https://openalex.org/W2975050087
id	id-8541201073306179650
abstract	
versions	

doc_2

authors	• Nilesh A. Ahuja • Ibrahima Ndiour • Trushant Kalyanpur • Omesh Tickoo
title	Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection
publication_date	2019-09-25 21:41:56+00:00
source	SupportedSources.ARXIV
journal	None
volume	
doi	
urls	• http://arxiv.org/pdf/1909.11786v1 • http://arxiv.org/abs/1909.11786v1 • http://arxiv.org/pdf/1909.11786v1
id	id5055410132234189327
abstract	We present a principled approach for detecting out-of-distribution (OOD) and adversarial samples in deep neural networks. Our approach consists in modeling the outputs of the various layers (deep features) with parametric probability distributions once training is completed. At inference, the likelihoods of the deep features w.r.t the previously learnt distributions are calculated and used to derive uncertainty estimates that can discriminate in-distribution samples from OOD samples. We explore the use of two classes of multivariate distributions for modeling the deep features - Gaussian and Gaussian mixture - and study the trade-off between accuracy and computational complexity. We demonstrate benefits of our approach on image features by detecting OOD images and adversarially-generated images, using popular DNN architectures on MNIST and CIFAR10 datasets. We show that more precise modeling of the feature distributions result in significantly improved detection of OOD and adversarial samples; up to 12 percentage points in AUPR and AUROC metrics. We further show that our approach remains extremely effective when applied to video data and associated spatio-temporal features by detecting adversarial samples on activity classification tasks using UCF101 dataset, and the C3D network. To our knowledge, our methodology is the first one reported for reliably detecting white-box adversarial framing, a state-of-the-art adversarial attack for video classifiers.
versions	