

cases	doc_1		doc_2				decision	id
							DUPLICATES	157
	authors	<ul style="list-style-type: none"><li>Nilesh A. Ahuja</li><li>Ibrahima J. Ndiour</li><li>Trushant Kalyanpur</li><li>Omesh Tickoo</li></ul>	authors	<ul style="list-style-type: none"><li>Nilesh A. Ahuja</li><li>Ibrahima Ndiour</li><li>Trushant Kalyanpur</li><li>Omesh Tickoo</li></ul>				
	title	Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection	title	Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection				
	publication_date	2019-09-25 00:00:00	publication_date	2019-09-25 00:00:00				
	source	SupportedSources.OPENALEX	source	SupportedSources.INTERNET_ARCHIVE				
	journal	arXiv (Cornell University)	journal					
	volume		volume					
	doi	None	doi					
	urls	<ul style="list-style-type: none"><li>https://openalex.org/W2975050087</li></ul>	urls	<ul style="list-style-type: none"><li>https://web.archive.org/web/20200910104810/https://arxiv.org/pdf/1909.11786v1.pdf</li></ul>				
	id	id-8541201073306179650	id	id-1570997348619314032				
	abstract		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		versions					