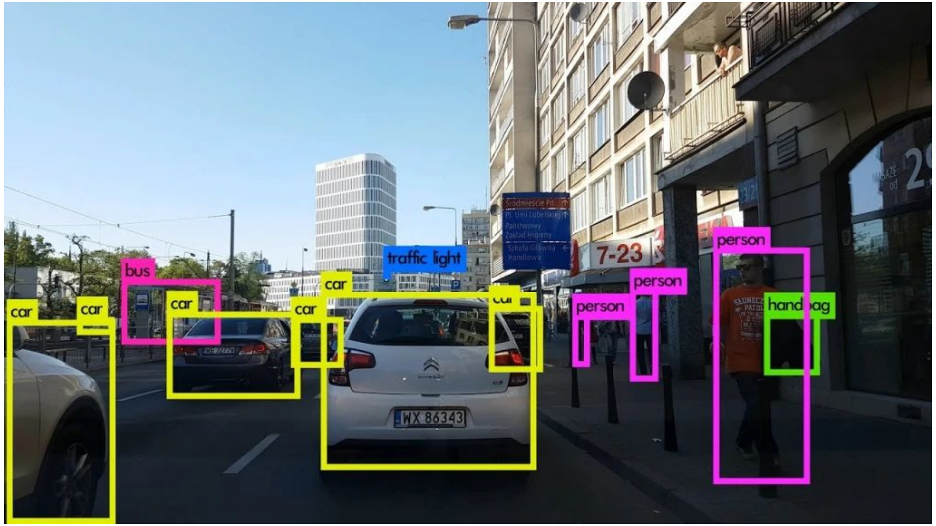
ReAct: OOD Detection with Rectified Activations

Yiyou Sun, Chuan Guo, Yixuan Li (Neurips 2021)

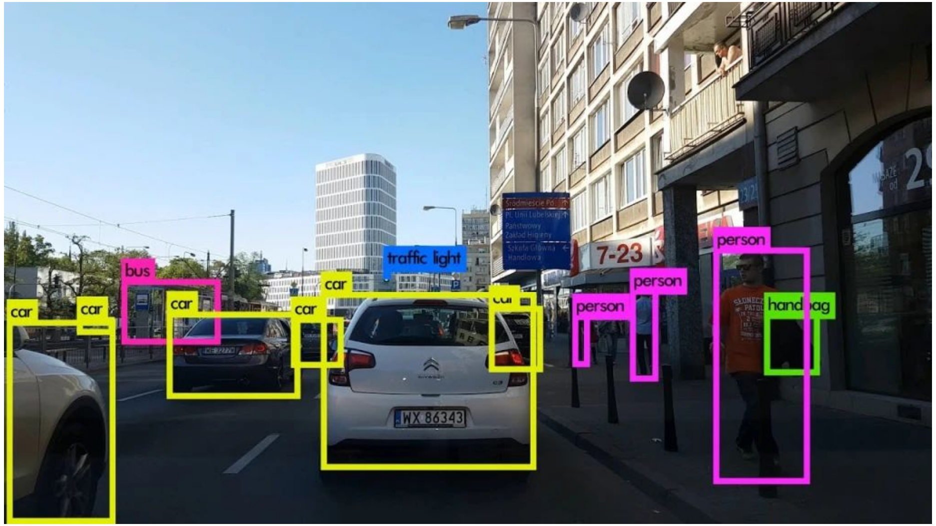
Presented by: **Group 4**
{Logan Crowl, Abishek Sridhar, Deying Song, Prince Wang, Shuaiqi Wang}

ML-715 (Fall 2022)

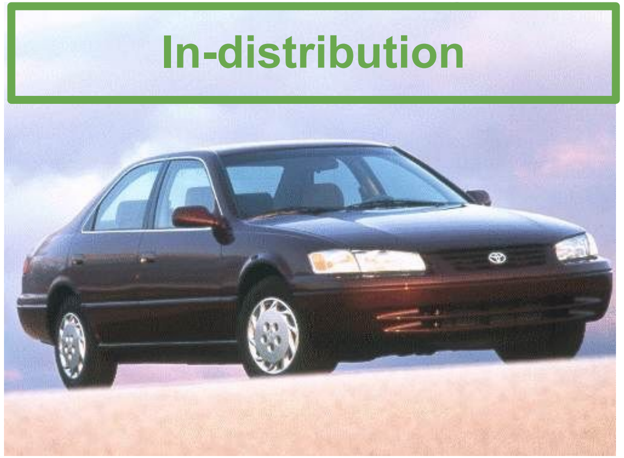
Carnegie Mellon University

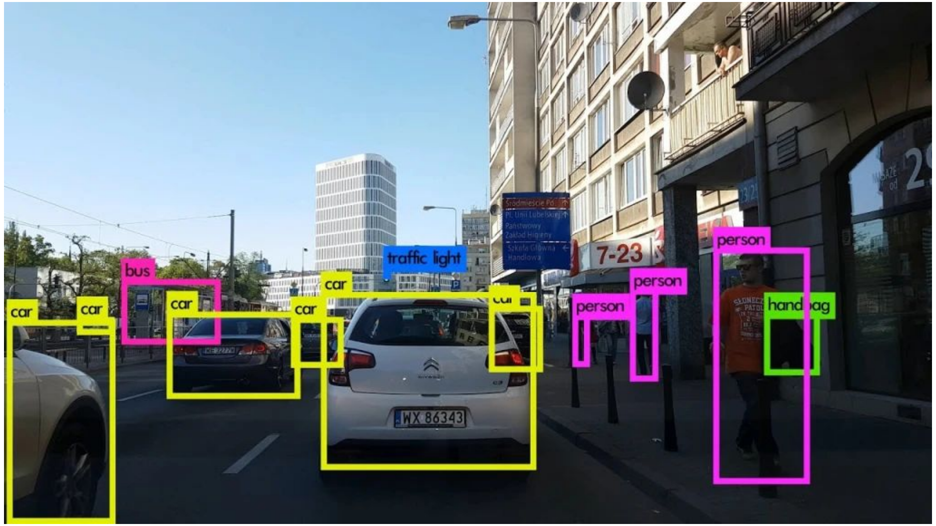


Autonomous driving

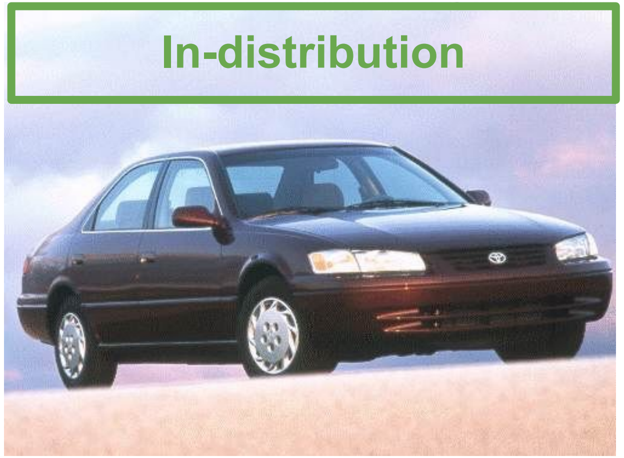


Autonomous driving

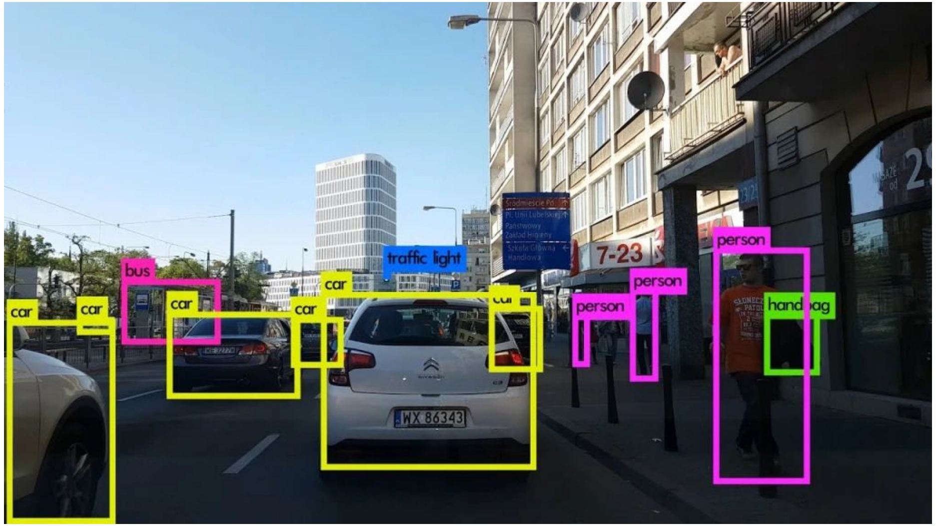




Autonomous driving



How can we detect these OOD inputs at test time?

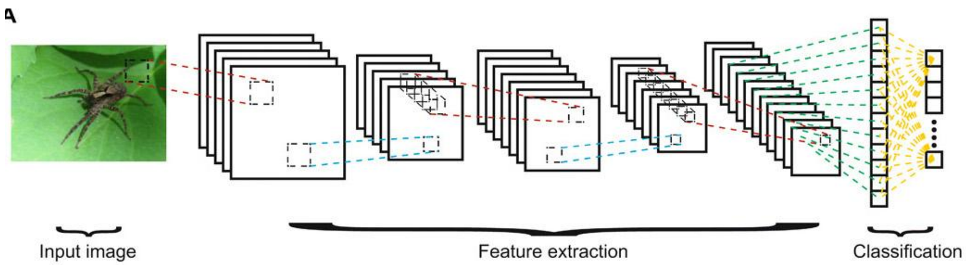


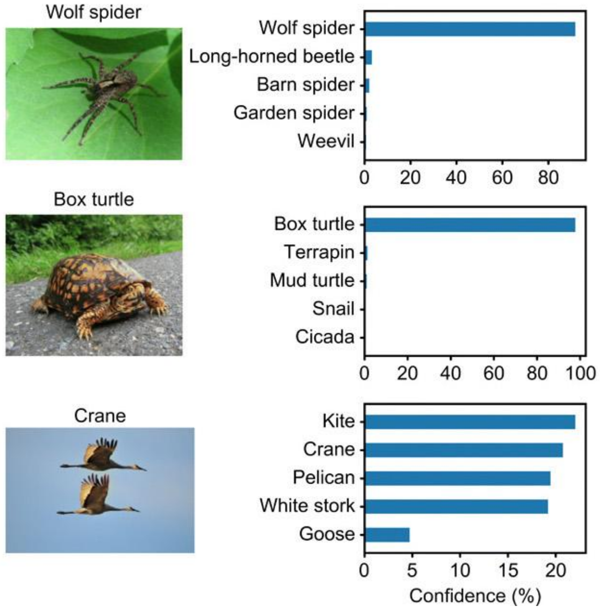
Autonomous driving

In-distribution

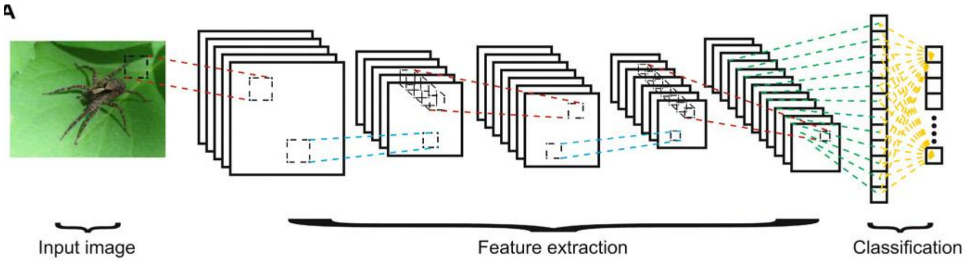
Out-of-distribution

Typical approach: interpret magnitude of logits as “confidence”



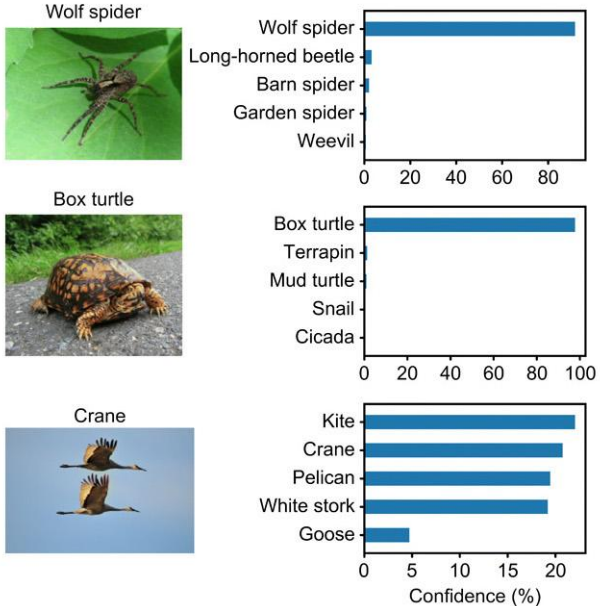


Typical approach: interpret magnitude of logits as "confidence"

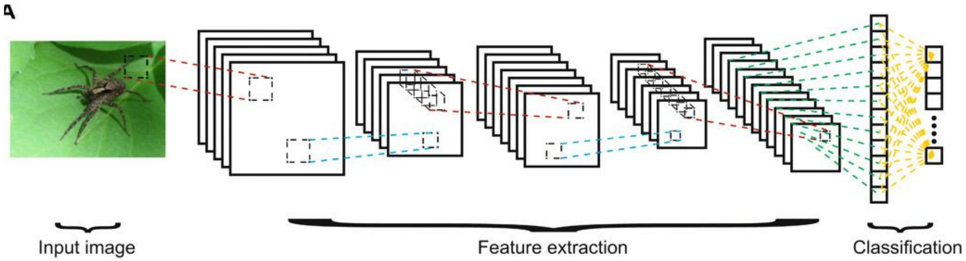


$$S_{\mathrm{MSP}}(\mathbf{x}; f) := \max_k \mathrm{Softmax}(Wh(\mathbf{x}) + \mathbf{b})_k$$

$$S_{\mathrm{Energy}}(\mathbf{x}; f) = -\log \sum_{k=1}^{K} \exp(\mathbf{w}_i^\top h(\mathbf{x}) + b_i)$$

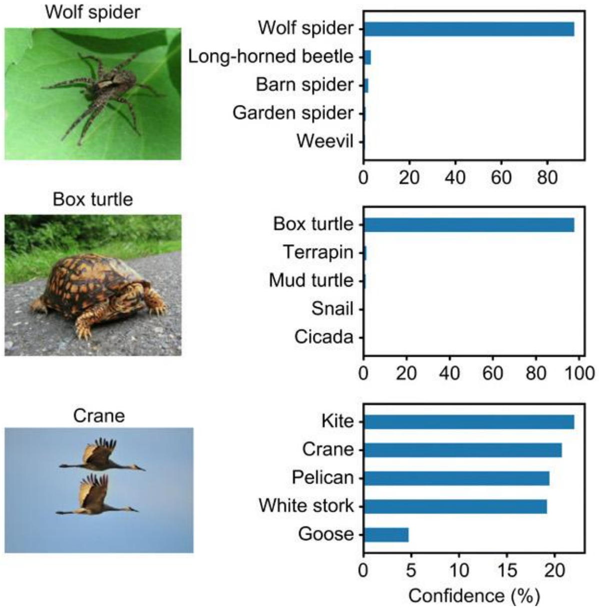


Typical approach: interpret magnitude of logits as "confidence"



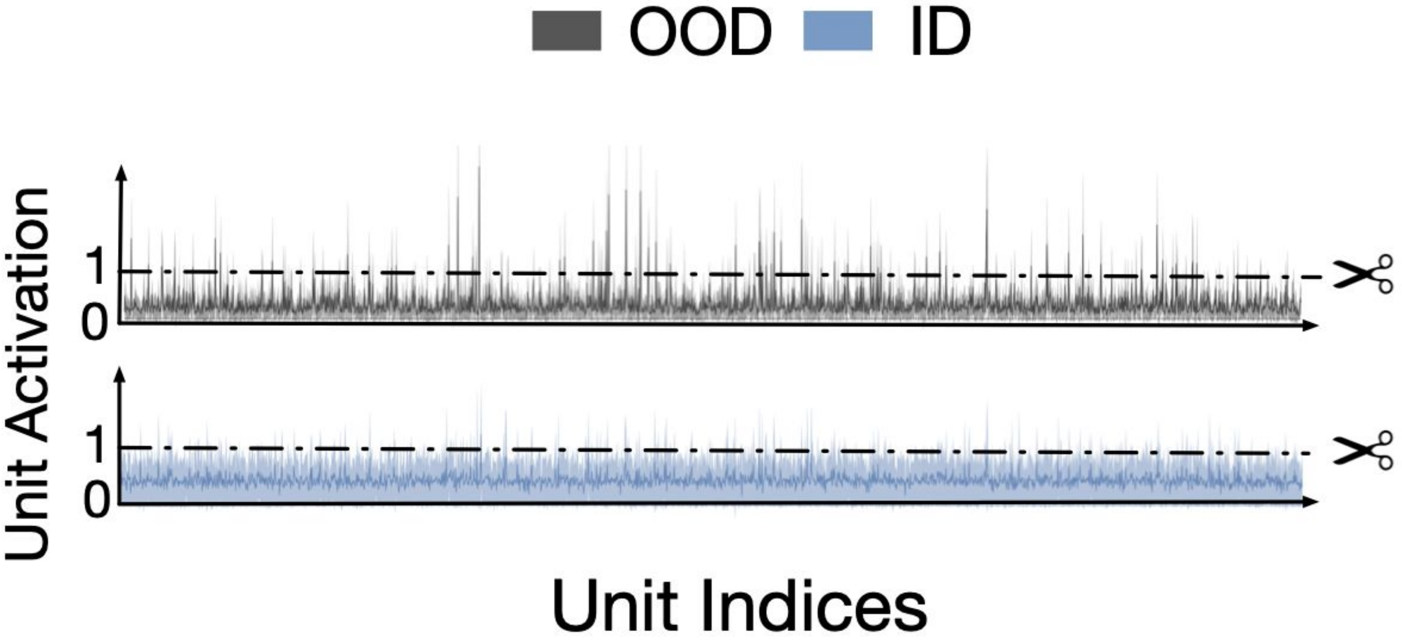
$$S_{\mathrm{MSP}}(\mathbf{x}; f) := \max_k \mathrm{Softmax}(Wh(\mathbf{x}) + \mathbf{b})_k$$

$$S_{\mathrm{Energy}}(\mathbf{x}; f) = -\log \sum_{k=1}^{K} \exp(\mathbf{w}_i^\top h(\mathbf{x}) + b_i)$$

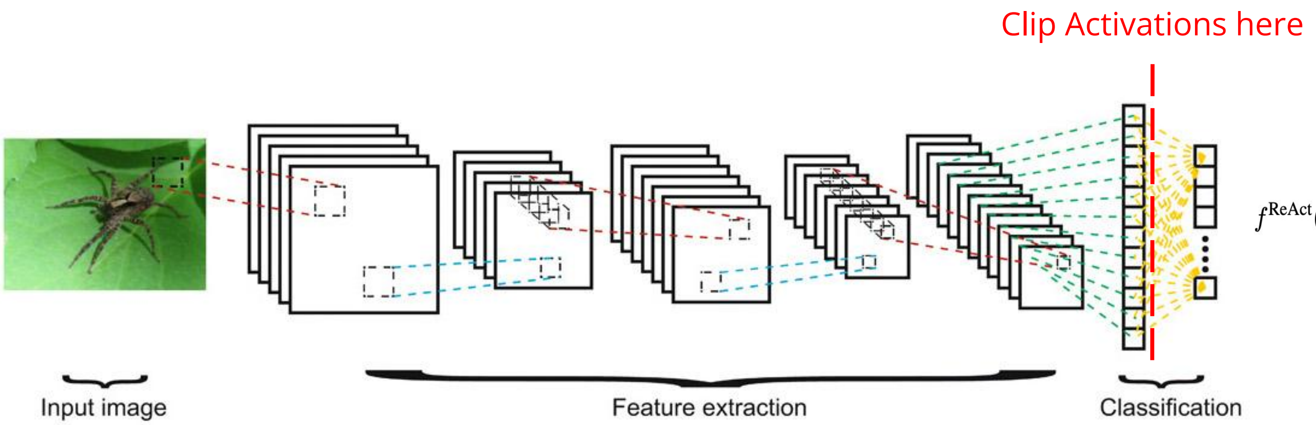


Problem: model is often too "confident" on OOD data (Nguyen et al.)

Key insight: OOD data have a distinctive activation pattern



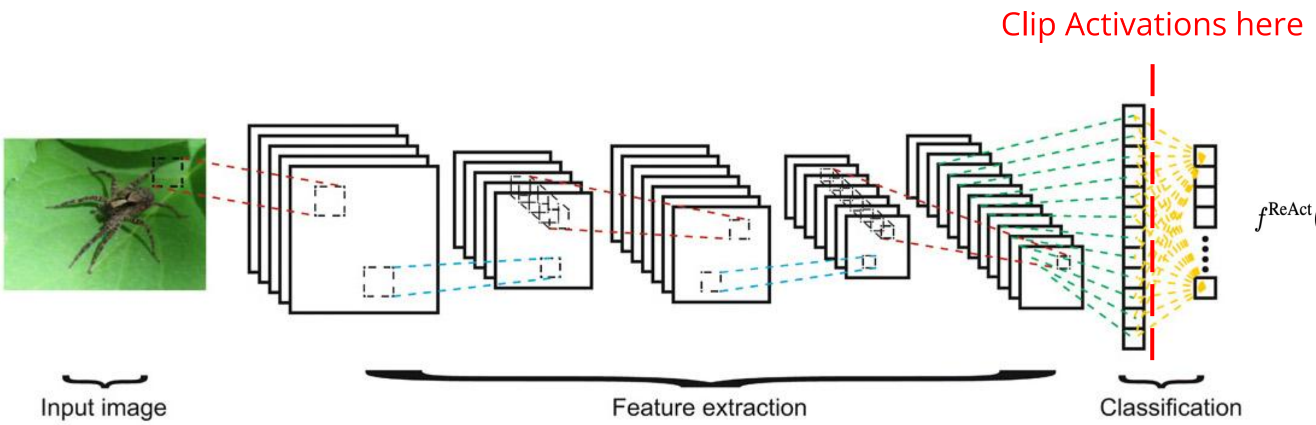
Proposed solution: clip activations in the penultimate layer



$$f^{\text{ReAct}}(\mathbf{x};\theta) = \mathbf{W}^\top \bar{h}(\mathbf{x}) + \mathbf{b},$$

$$\texttt{ReAct}(x;c) = \min(x,c)$$

Proposed solution: clip activations in the penultimate layer



$$f^{\text{ReAct}}(\mathbf{x};\theta) = \mathbf{W}^\top \bar{h}(\mathbf{x}) + \mathbf{b},$$

$$\texttt{ReAct}(x;c) = \min(x,c)$$

$$G_\lambda(\mathbf{x}; f^{\text{ReAct}}) = \begin{cases} \text{in} & S(\mathbf{x}; f^{\text{ReAct}}) \geq \lambda \\ \text{out} & S(\mathbf{x}; f^{\text{ReAct}}) < \lambda \end{cases},$$

ReAct beats other post-hoc OOD detection methods

Model	Methods	OOD Datasets								Average	
		iNaturalist		SUN		Places		Textures			
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
ResNet	MSP [15]	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
	ODIN [31]	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
	Mahalanobis [29]	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
	Energy [33]	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
	ReAct (Ours)	**20.38**	**96.22**	**24.20**	**94.20**	**33.85**	**91.58**	**47.30**	**89.80**	**31.43**	**92.95**
MobileNet	MSP [15]	64.29	85.32	77.02	77.10	79.23	76.27	73.51	77.30	73.51	79.00
	ODIN [31]	55.39	87.62	54.07	85.88	57.36	84.71	49.96	85.03	54.20	85.81
	Mahalanobis [29]	62.11	81.00	47.82	86.33	52.09	83.63	92.38	33.06	63.60	71.01
	Energy [33]	59.50	88.91	62.65	84.50	69.37	81.19	58.05	85.03	62.39	84.91
	ReAct (Ours)	**42.40**	**91.53**	**47.69**	**88.16**	**51.56**	**86.64**	**38.42**	**91.53**	**45.02**	**89.47**

Table 1: **Main results.** Comparison with competitive *post hoc* out-of-distribution detection methods. All methods are based on a model trained on **ID data only** (ImageNet-1k), without using any auxiliary outlier data. ↑ indicates larger values are better and ↓ indicates smaller values are better. All values are percentages.

Our Dissection

Activation functions matter!

ReAct fails to improve metrics with tanh activation:

- Theoretical analysis and even the idea of upper clipping seem to rely on ReLU
- Effect of activation function not studied
- No comments about generalization to other activation functions and architectures

Activation functions matter!

With tanh activation

Without ReAct	FPR95	AUROC
	89.51	70.99

Became Worse!

With ReAct	FPR95	AUROC
	91.69	64.01

ReAct fails to improve metrics with tanh activation:

- Theoretical analysis and even the idea of upper clipping seem to rely on ReLU
- Effect of activation function not studied
- No comments about generalization to other activation functions and architectures

Activation functions matter!

With tanh activation

Without ReAct

FPR95	AUROC
89.51	70.99

Became Worse!

With ReAct

FPR95	AUROC
91.69	64.01

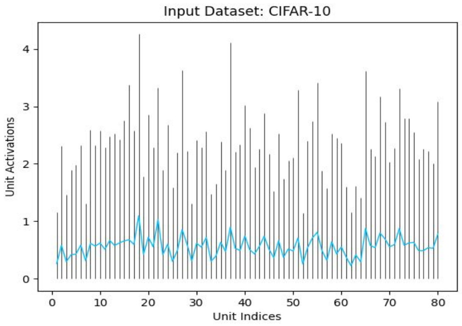
ReAct fails to improve metrics with tanh activation:

- Theoretical analysis and even the idea of upper clipping seem to rely on ReLU
- Effect of activation function not studied
- No comments about generalization to other activation functions and architectures

Going Ahead:

- Analyze ReAct for bounded activation functions
- Explore upper and lower clipping for symmetrical, zero-centered activations

Batchnorm isn't the only reason ReAct works; so what is?



Model With BatchNorm

AUROC 46.86 → 79.36

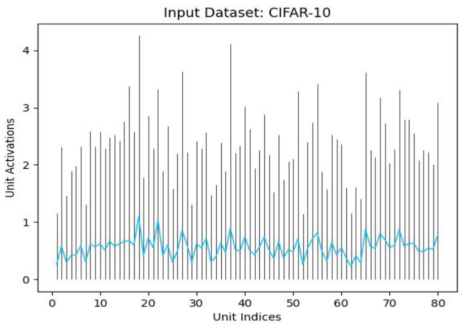
FPR95 95.34 → 78.10

ReAct improves metrics without BatchNorm for a smaller model:

- Unexplained by paper's BatchNorm reasoning

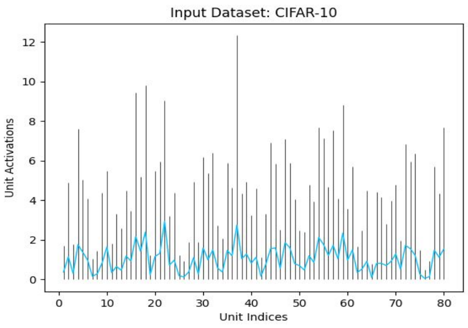
- The variation in ID activations' mean not constant without BatchNorm => Theoretical analysis breaks

Batchnorm isn't the only reason ReAct works; so what is?



Model With BatchNorm

AUROC	46.86	→	79.36
FPR95	95.34	→	78.10



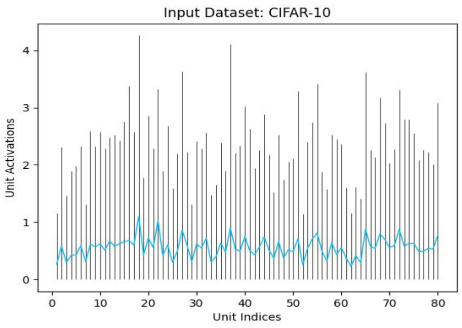
Model Without BatchNorm

AUROC	38.74	→	54.70
FPR95	96.50	→	93.27

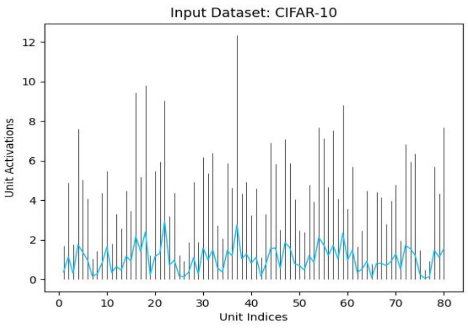
ReAct improves metrics without BatchNorm for a smaller model:

- Unexplained by paper's BatchNorm reasoning
- The variation in ID activations' mean not constant without BatchNorm => Theoretical analysis breaks

Batchnorm isn't the only reason ReAct works; so what is?



Model With BatchNorm



Model Without BatchNorm

	Model With BatchNorm		Model Without BatchNorm	
AUROC	46.86 →	79.36	38.74 →	54.70
FPR95	95.34 →	78.10	96.50 →	93.27

ReAct improves metrics without BatchNorm for a smaller model:

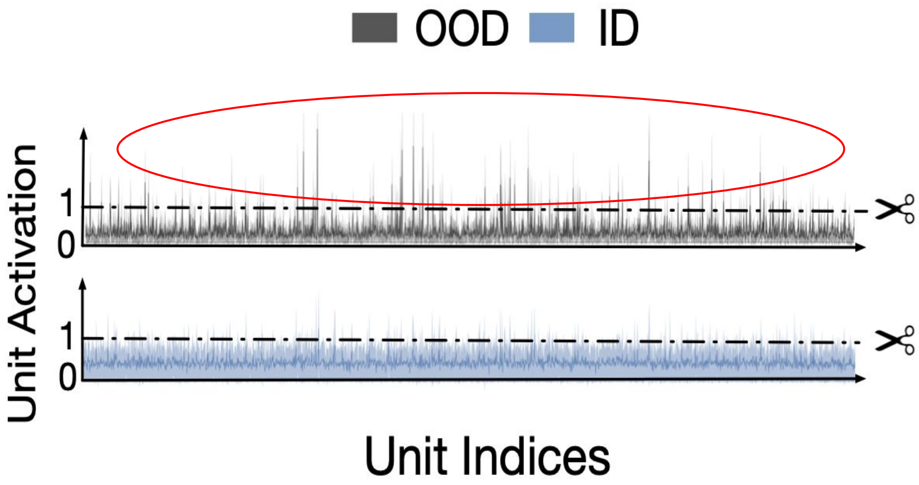
- Unexplained by paper's BatchNorm reasoning
- The variation in ID activations' mean not constant without BatchNorm => Theoretical analysis breaks

Going Ahead:

- Relax assumption about constant mean in theoretical analysis
- Run experiments on models without BatchNorm layers

If OOD activations look different, why not use that directly?

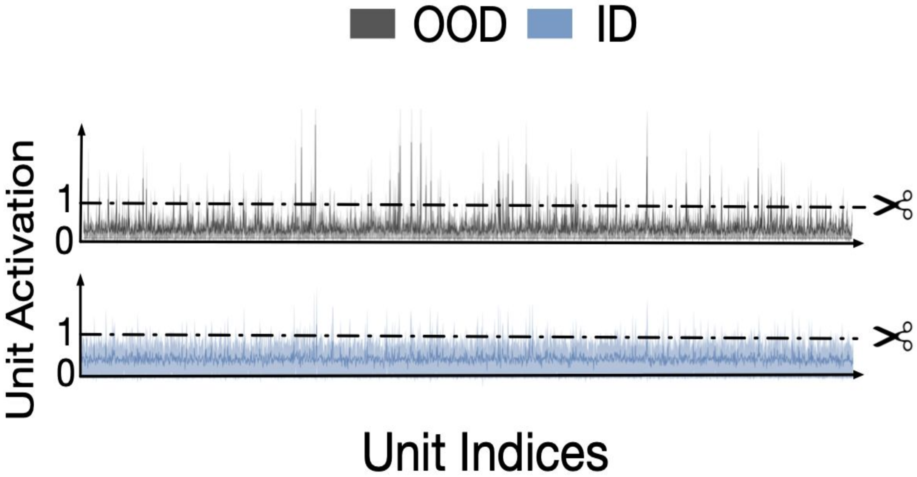
- ReAct relies on out-of-distribution activations being clipped more often



If OOD activations look different, why not use that directly?

- ReAct relies on out-of-distributions activations being clipped more often
- New proposed metric:
  Proportion of final layer activations below the clipping threshold

$$S(h(x); c) = \frac{(\#\text{ of activations below threshold c})}{(\#\text{ of total activations})}$$



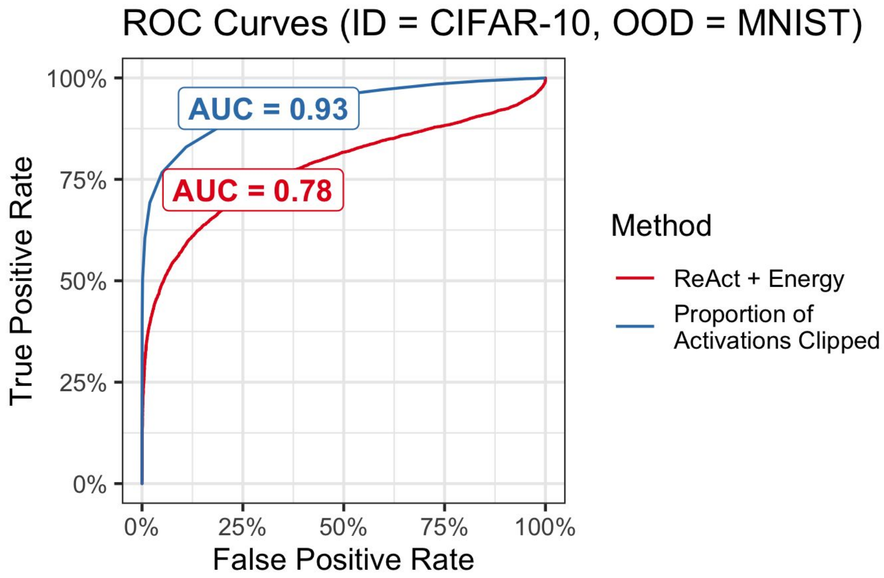
If OOD activations look different, why not use that directly?

- ReAct relies on out-of-distributions activations being clipped more often

- New proposed metric:
  Proportion of final layer activations below the clipping threshold

$$S(h(x); c) = \frac{(\text{\# of activations below threshold c})}{(\text{\# of total activations})}$$

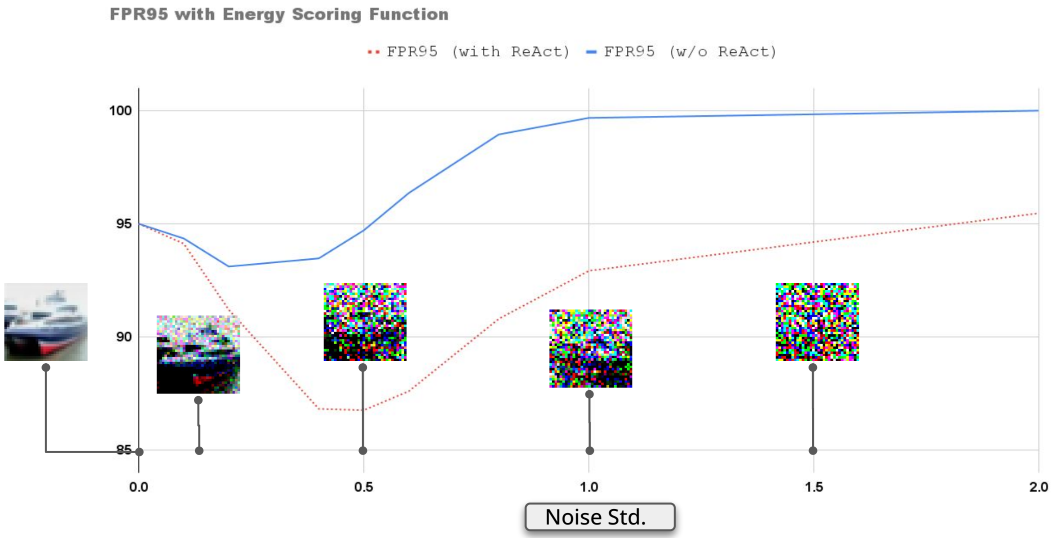
- Future research could use the activation patterns more thoughtfully



Metrics sensitive to OOD’s true mean, variance!

OOD sample’s true mean, variance is not known at test time

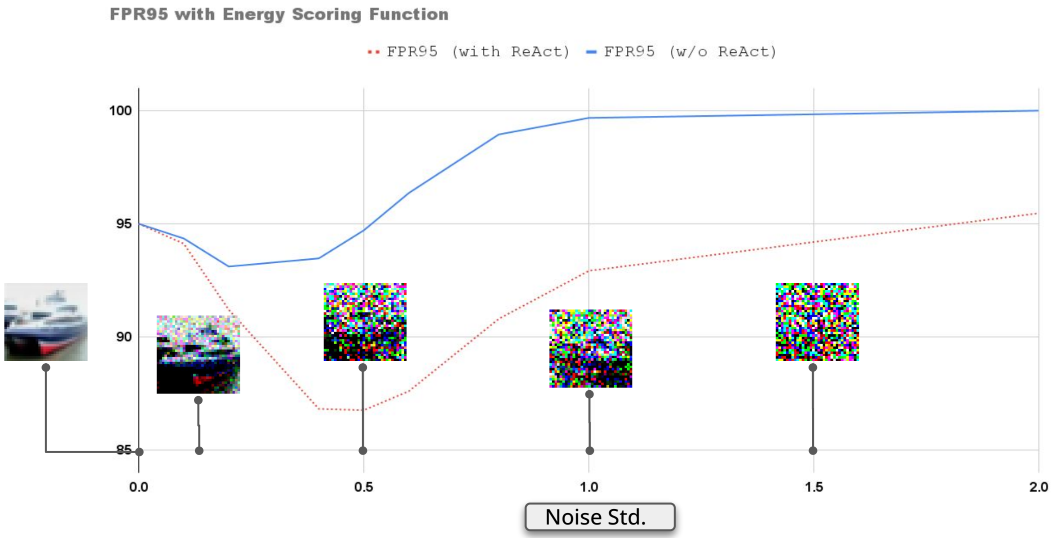
Metrics sensitive to OOD's true mean, variance!



OOD sample's true mean, variance is not known at test time:

- For eg., high noise values with ID's preprocessing lead to wrong predictions
- Sensitivity probably due to un-normalized Energy Score as the best metric for ReAct

Metrics sensitive to OOD's true mean, variance!



OOD sample's true mean, variance is not known at test time:

- For eg., high noise values with ID's preprocessing lead to wrong predictions
- Sensitivity probably due to un-normalized Energy Score as the best metric for ReAct

Going Ahead:

- Verify and handle effect of OOD statistics on the metrics
- Prioritize scoring functions that don't scale with logit values - like softmax