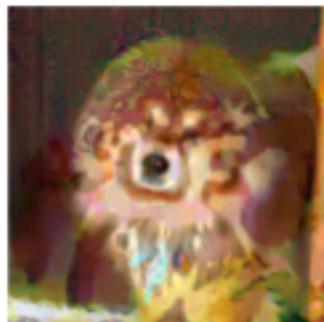


# Related: Adversarial Robustness

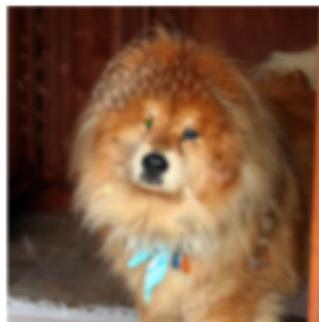
Previous Attacks



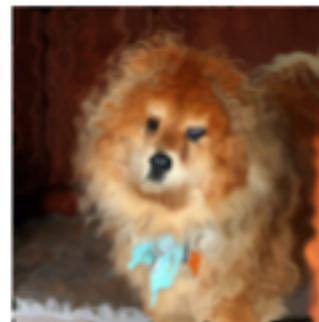
$L_\infty$



$L_2$



$L_1$



Elastic

Our New Attacks



JPEG



Fog



Snow



Gabor

# Related: Adversarial Robustness

Defense Robustness Under Different Attacks

Adversarially Trained Defense	Adversarial Attack							
	$L_\infty$	$L_2$	$L_1$	JPEG	Elastic	Fog	Snow	Gabor
None	7	17	22	0	31	16	10	5
$L_\infty$	88	42	15	14	49	20	37	55
$L_2$	80	88	79	67	48	18	38	53
$L_1$	62	71	89	56	43	18	31	47
JPEG	65	70	54	92	40	19	31	52
Elastic	23	25	11	1	91	25	40	41
Fog	1	3	8	0	28	91	43	54
Snow	13	15	9	1	39	37	93	60
Gabor	12	19	14	0	39	29	40	82

# Related: Adversarial Robustness

	Clean Accuracy	$L_\infty$	$L_2$	$L_1$	Elastic	JPEG	Fog	Snow	Gabor	mUAR
SqueezeNet	84.1	5.2	11.2	14.9	25.9	<b>1.9</b>	20.1	9.8	4.4	12.8
ResNeXt-101 (32×8d)	95.9	2.5	5.5	20.7	26.5	1.8	14.1	12.4	5.3	13.4
ResNeXt-101 (32×8d) + WSL	<b>97.1</b>	3.0	5.7	28.3	29.4	<b>1.9</b>	26.2	20.3	8.0	19.0
ResNet-18	91.6	2.7	8.2	13.5	22.6	1.8	20.3	9.5	4.2	12.0
ResNet-50	94.2	2.7	6.6	20.1	24.9	1.8	15.8	11.9	4.9	13.2
ResNet-50 + Stylized ImageNet	94.6	2.9	7.4	22.8	26.0	1.8	16.2	12.5	8.1	14.6
ResNet-50 + Patch Gaussian	93.6	4.5	10.9	27.4	28.2	1.8	23.9	10.5	5.2	16.2
ResNet-50 + AugMix	95.1	<b>6.1</b>	<b>13.4</b>	<b>34.3</b>	<b>38.8</b>	1.8	<b>28.6</b>	<b>24.7</b>	<b>11.1</b>	<b>23.2</b>

# Motivation

Folklore: ML does poorly OOD

Why and when? Can we predict it?

# Motivation

Folklore: ML does poorly OOD

Why and when? Can we predict it?



Model works



Model does poorly

Geirhos et al., 2018; Ford et al., 2019

# The Challenge

Test accuracy in-distribution well-defined (just measure it)

OOD not well-defined (many types); how to measure?

Danger: accidentally overfit to specific type

Geirhos et al., 2018; Ford et al., 2019

# The Solution

Will consider many types of shift at once



Original



ImageNet-C



ImageNet-A



ImageNet-v2



ImageNet-R

# Starting Point: ImageNet-C

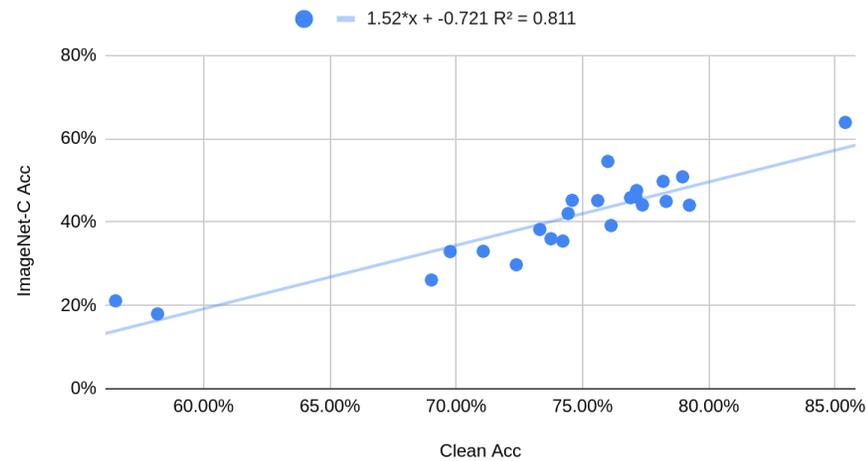
What helps with robustness?

# Starting Point: ImageNet-C

What helps with robustness?

In-distribution accuracy:

ImageNet-C Acc vs. Clean Acc



# Combined Results

	Resnet50	Resnet152	SE_Resnet152	SIN	WSL	AugMix
Orig.	76.1	78.3	78.7	74.6	85.4	77.6
IN-C	41.6	47.8	50.9	47.9	65.5	???
IN-v2	63.2	66.8	67.4	62.8	77.0	???
IN-R	37.1	42.9	40.9	44.0	81.6	42.8

## Conclusions:

- pre-training helps a *\*lot\**
- depth, data aug, accuracy also help
- SE accuracy gain was spurious