Defendr: A Robust Model for Image Classification

Apoorva Dornadula, Ashwini Pokle, Akhila Yerukola

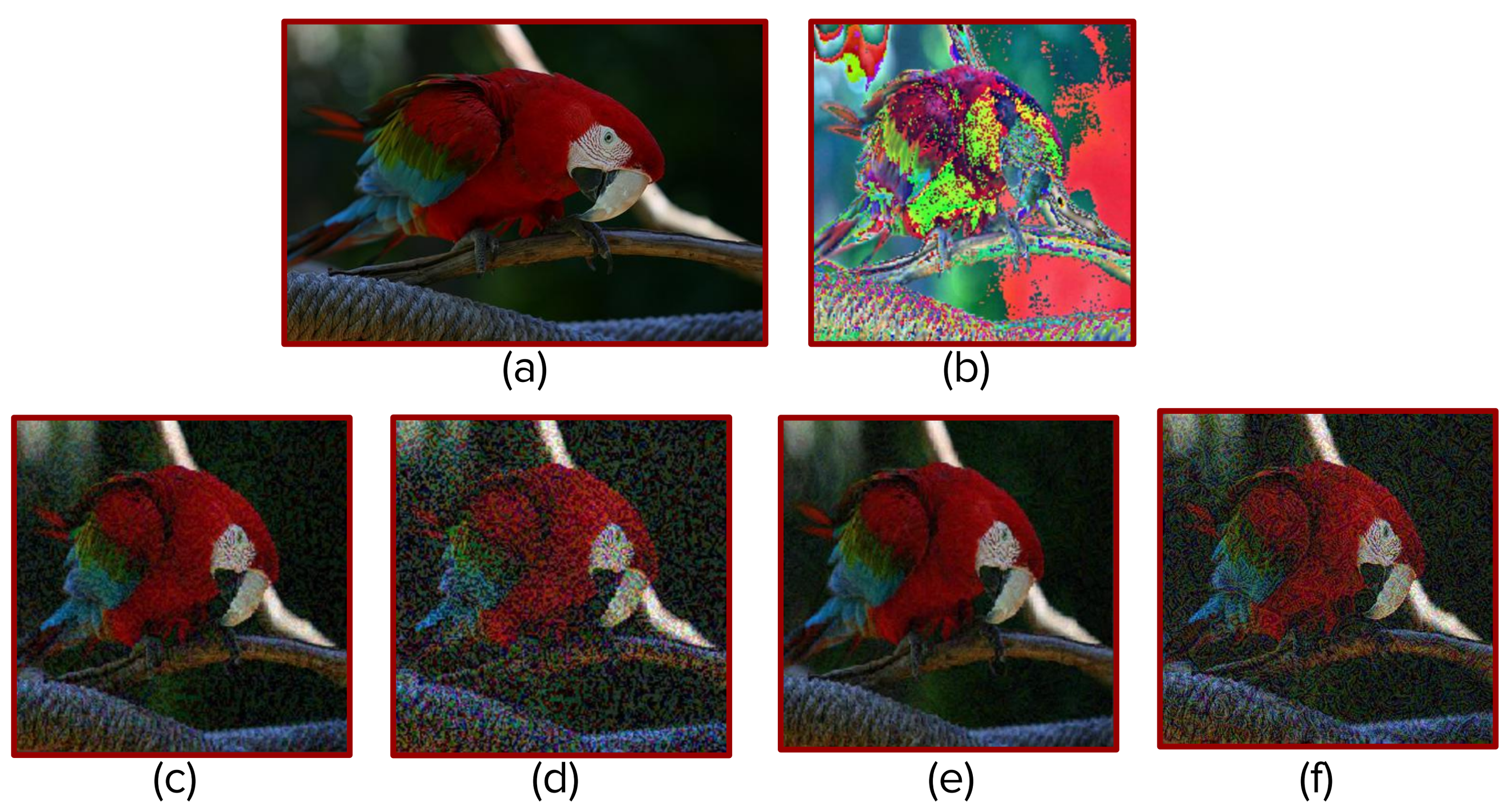
Stanford University

Motivation/Problem Statement

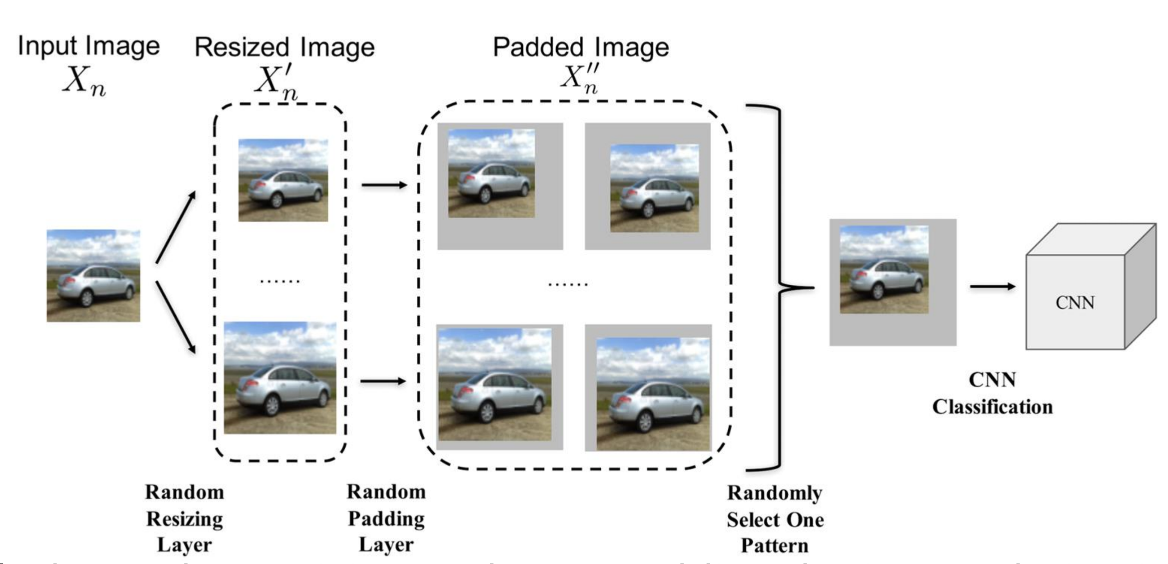
Motivation: Numerous computer vision applications, such as self driving cars, are susceptible to being fooled by adversarial images.
Problem Statement: create a robust model that is able to correctly classify adversarial examples without compromising performance on normal images.

Dataset/Preprocessing

- We used a subset of ImageNet with 201 classes
- 10854 images in the training set, 1206 in validation set and 2010 in test set
- All images were reshaped to 299x299, with the mean subtracted from all three channels.
- The red, green, and blue channels are normalized to have mean 0.485, 0.456, 0.406 and standard deviation 0.229, 0.224, 0.225, respectively
- Data set was augmented with adversarial images generated by using FGSM and PGD



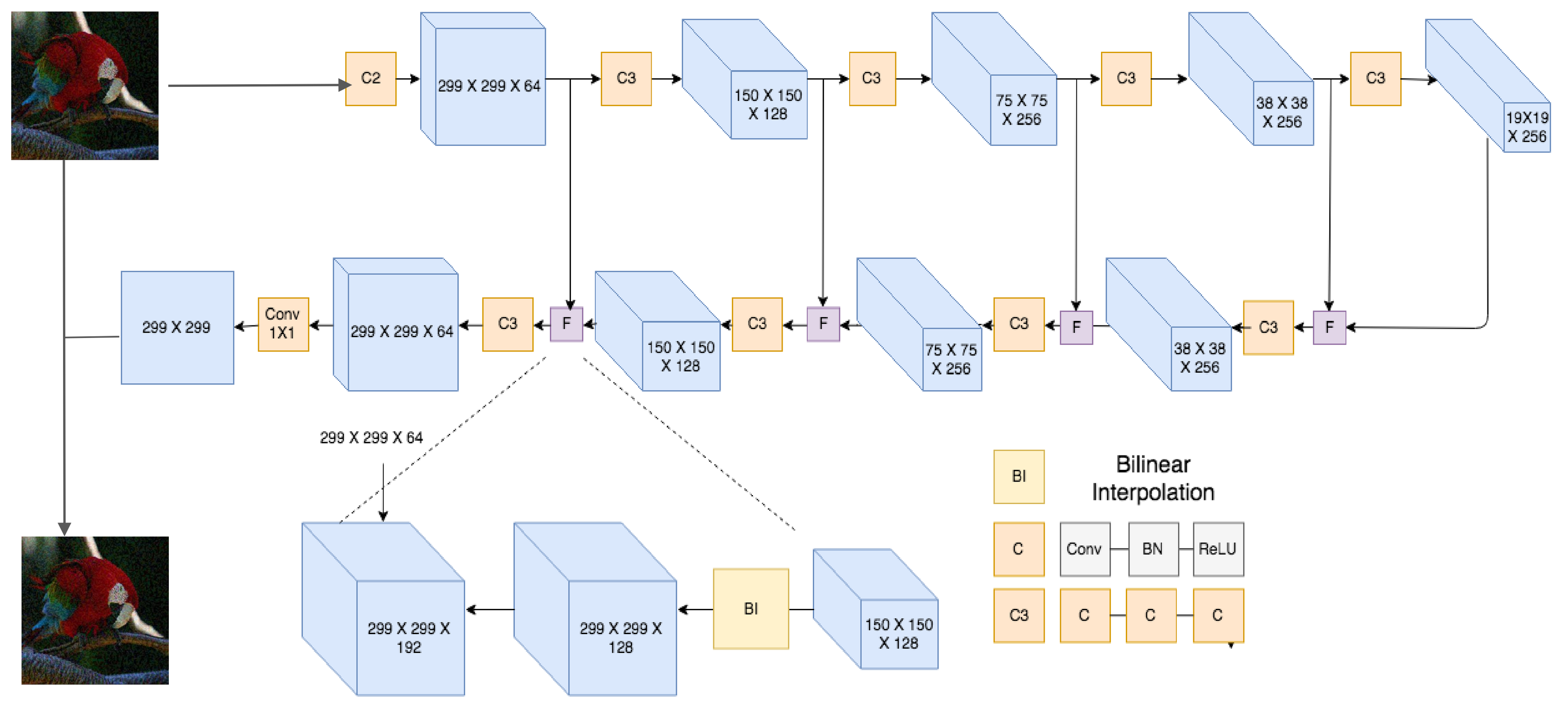
(a) Original Image, (b) Original Image after Normalization, (c) FGSM w/ ϵ = 0.12, (d) FGSM w/ ϵ = 0.3, (e) PGD w/ ϵ = 0.12, (f) PGD w/ ϵ = 0.3



Xie et. al. shows that resizing and zero-padding the image during inference increases model robustness to adversarial examples

Methods

Denoising U-Net (DUNET)



- Uses L1 loss between the original image and denoised image
- Denoised image is input to a ResNet-101

Adversarial Logit Pairing (ALP) & Clean Logit Pairing (CLP)

$$J(M,\theta) + \lambda \frac{1}{m}\sum_{i=1}^{m} L(f(x(i);\theta) f(\tilde{x}^{(i)},\theta))$$

$$J(M,\theta) + \lambda \frac{2}{m}\sum_{i=1}^{m/2} L(f(x(i);\theta) f(x(i+\frac{m}{2}),\theta))$$

- *Top*: Loss function for ALP - encourages logits between normal and corresponding adv. Image to be similar
- *Bottom*: Loss function for CLP - encourages logits of normal images in a batch to be similar

Experiments/Results

Single Architecture Performance (Adv. Ex: FGSM with ϵ = 0.12)				
		No Adv. Ex.	50% Adv. Ex.	100% Adv. Ex.
BASELINE	No Adv. Train	90.70	78.46	66.22
	Adv. Train	90.40	80.70	70.00
CLP	No Adv. Train	90.85	78.66	66.47
	Adv. Train	-	-	-
ALP	No Adv. Train	-	-	-
	Adv. Train	90.75	80.65	70.55
DUNET	No Adv. Train	-	-	-
	Adv. Train	92.4	-	86.6
RANDOMIZATION	No Adv. Train	85.92	75.55	65.82
	Adv. Train	85.47	76.69	68.01

Table 1: Effects of Adversarial Training on Single Model Architectures

Attack Effectiveness (Tested on 100% Adv. Ex.)				
	FGSM(ϵ = 0.12)	FGSM(ϵ = 0.3)	PGD(ϵ = 0.12)	PGD(ϵ = 0.3)
BASELINE	70.00	47.81	80.00	63.88
ALP	70.55	48.91	80.95	62.64
DUNET	86.6	72.5	85.8	79.0

Table 2: Attack Type Effectiveness

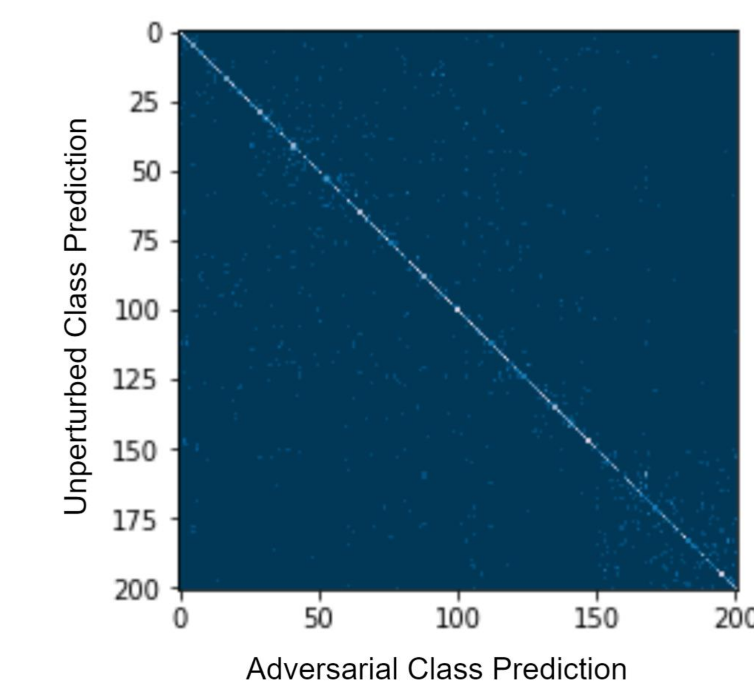
Experiments/Results (cont'd)

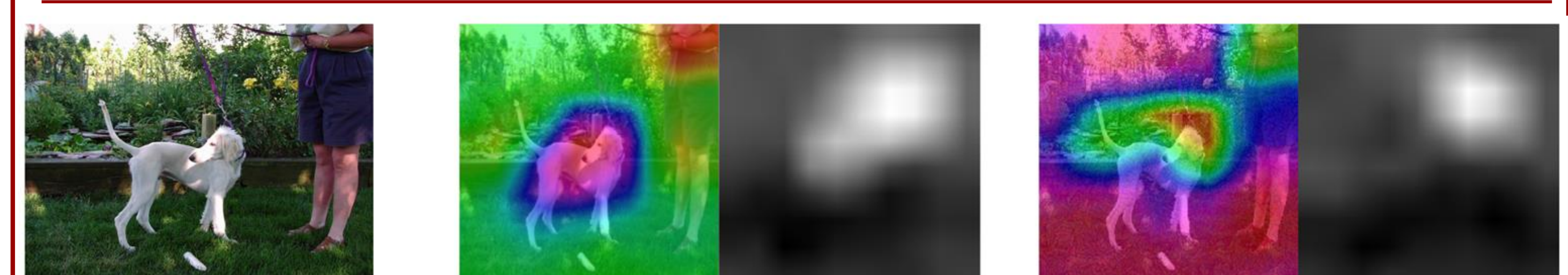
- Models trained on FGSM are more robust to other attacks compared to those trained on PGD.
- Ensemble of DUNET, ALP and CLP performs the best among ensembles and achieves accuracy of 84.86%.
- DUNET outperforms all other models.

Model Robustness		
Tested 100% Adversarial Examples		
	FGSM	PGD
Baseline FGSM (0.12)	71	80
Baseline PGD (0.12)	69.35	81
ALP FGSM (0.12)	69.35	79.40
DUNET FGSM (0.12)	**86.6**	**82.9**
DUNET FGSM (0.3)	**72.5**	68.0
DUNET PGD (0.12)	70.9	**85.8**
DUNET PGD (0.3)	49.7	79.0

Model Ensemble Accuracy	
	Accuracy
ALP + CLP	66.22
DUNET + CLP	77.71
DUNET + ALP	77.71
DUNET + ALP + CLP	84.86

- For the most part, class predictions for normal images match to those of the adversarial class
- The bottom right square of noise is present because those labels correspond to dog types which can be easily confused.





Left: dog of breed 1 image, *Middle*: class activations on original dog breed 1 image, *Right*: class activations on perturbed dog image; prediction: dog of breed 2

Future Work

- Ensemble adversarial training to decouple generation of adversarial examples from the model being trained. This will further improve the robustness of model against black-box attacks.
- Include techniques to make the model robust against white-attacks.

References

[1] Liao, Fangzhou, et al. "Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser." *arXiv preprint arXiv:1712.02976* (2017).
[2] Kannan, Harini, Alexey Kurakin, and Ian Goodfellow. "Adversarial Logit Pairing." *arXiv preprint arXiv:1803.06373*(2018).
[3] Xie, Cihang, et al. "Mitigating adversarial effects through randomization." *arXiv preprint arXiv:1711.01991* (2017).