Intriguing Properties of Adversarial Examples

E.D. Cubuk, B. Zoph, S.S. Schoenholz, Q.V. Le

Google

arXiv:1711.02846

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Universality Observations

Mean-Field Theory

Network Architectures

Summary

Introduction

Basic Premise and Motivation

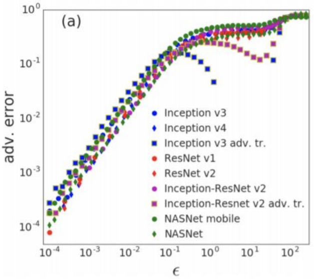
- Study properties of adversarial examples
- Calculate adversarial error, the difference between clean and adversarial accuracy with perturbation ϵ
  - At small ϵ, adversarial error has similar dependence on ϵ across all models and datasets; grows like $A\epsilon^B$
- Show that origin of adversarial examples is inherent uncertainty that neural networks have; dependent only on logit differences
- Using these results, propose new methods of combating adversarial examples

Universality Observations

- FGSM adversarial example calculation where x^{adv} is adversarial image, x is clean image, y is the correct label for x, ϵ is perturbation size, and L is the loss function

$$x^{adv} = x + \epsilon \operatorname{sign}(\nabla_x L(x, y))$$

- Plot adversarial error as function of ϵ on ImageNet for various models
  - For $\epsilon < 0.2$, obeys power law with exponent between 0.9 and 1.1

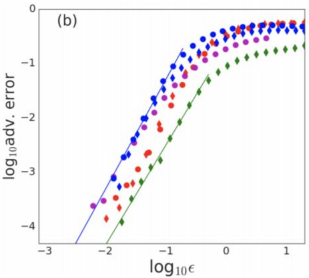


Universality Observations

- Try to get different form for adversarial error using different step l.l. attack, where y_{LL} is least likely predicted class

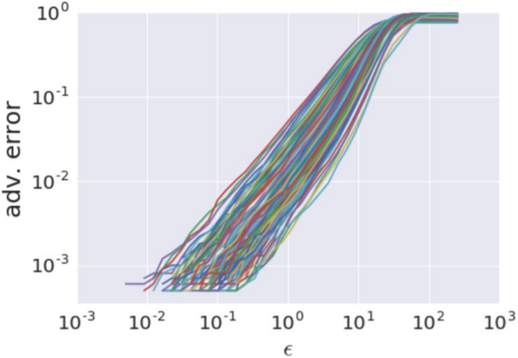
$$x^{adv} = x - \epsilon \ \mathrm{sign}(\nabla_x L(x, y_{LL}))$$

- Still obeys power law except with exponent between 1.8 and 2.2 for ImageNet and 1.8 and 2.5 for MNIST and CIFAR10; image is for ImageNet



Universality Observations

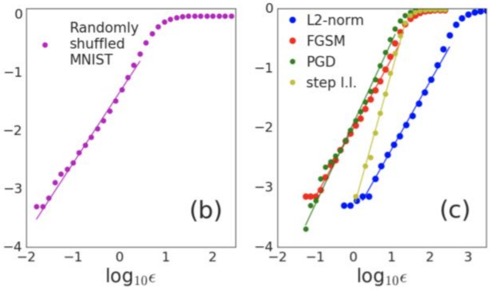
- Test more extensively by investigating effect of architecture by stochastically sampling thousands of neural networks and training on MNIST
    - Either fully-connected with 1-4 hidden layers and 30-2000 nodes/layer or simple CNN with dropout between 0-0.5
- Power law with exponent between 0.9 and 1.1



Universality Observations

- Probe relationship between generalization and adversarial robustness by training fully connected network on MNIST until perfect test set accuracy and conducting FGSM attack
  - Power law with exponent 1.2

  Investigate dependence of error on attack protocol using MNIST
  - Exponent values are L2-norm: 1.1, FGSM: 1.2, PGD: 1.3, step-l.l.: 2.3



Mean-Field Theory

Linear Response

- Focus on L2-variant of FGSM; form and exponent of power law insensitive to this choice as shown previously

$$x^{adv} = x + \epsilon \, \frac{\nabla_x L}{||\nabla_x L||_2}$$

- Find minimum ϵ such that class assigned to input x changes, $\hat{\epsilon}(x)$
- Assuming the network perfectly classifies clean images, adversarial error rate is equal to $P(\hat{\epsilon} < \epsilon)$
- Notation: $\hat{y}_i(x)$ is output, $h_i(x)$ is corresponding logits, choose ordering of logits such that $h_1(x) \geq h_2(x) \geq ... \geq h_N(x)$
- Logit differences: $\Delta_{ij}(x) = h_i(x) - h_j(x)$
- For a network to make an erroneous prediction, $h_1(x^{adv}) < h_j(x^{adv})$ for some j

Mean-Field Theory

Linear Response

- Through derivations (see paper appendix), arrive at equation where $J_{ij} = \frac{\partial h_j}{\partial x_i}$ is the input-logit Jacobian and $\delta_i = \frac{\partial L}{\partial h_i}$ is the error of network outputs; valid at small ϵ

$$h(x^{adv}) = h(x) + \epsilon \, \frac{J^T J\delta}{||J\delta||_2} + O(\epsilon^2)$$

- Define $\Gamma(x) = \frac{J^T J\delta}{||J\delta||_2}$. Then, $\Delta_{ij}(x^{adv}) = h_i(x^{adv}) - h_j(x^{adv}) \approx (h_i(x) - h_j(x)) + \epsilon(\Gamma_i(x) - \Gamma_j(x)) = \Delta_{ij}(x) + \epsilon(\Gamma_i(x) - \Gamma_j(x))$
- Misclassification occurs when $\Delta_{1j}(x^{adv}) = 0$ for some j. Therefore, for any given j,

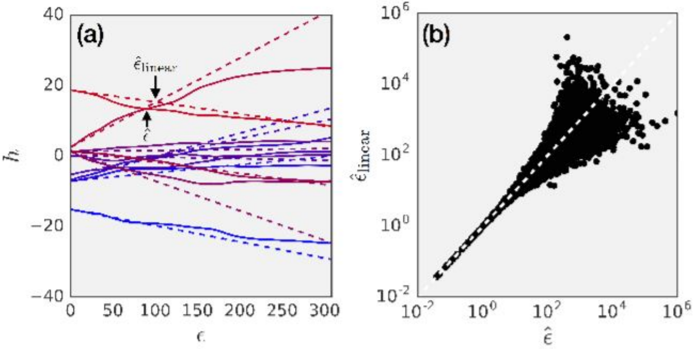
$$\hat{\epsilon}_j(x) = \frac{\Delta_{1j}(x)}{\Gamma_j(x) - \Gamma_1(x)}$$

which allows us to approximate $\hat{\epsilon}_{linear}(x) = min_j(\hat{\epsilon}_j(x))$

Mean-Field Theory

Linear Response

- Graph logit values as function of ϵ and observe crossover point; good approximation
- Graph $\hat{\epsilon}$ against ϵ; linear good fit for small ϵ

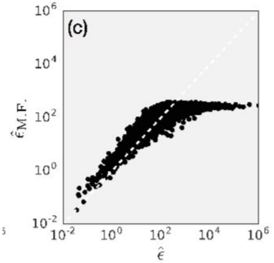


Mean-Field Theory

Linear Response

- However, $\Gamma_i(x)$ difficult to calculate because of Jacobian; introduce a mean-field approximation where we replace this with the average over entire dataset, $\langle \Gamma_i(x) \rangle$ (see paper 2 in references)
- Also, observe that vast majority of time second logit overtakes first logit most quickly, therefore assume $\Delta_{12}(x)$ results in minimum $\hat{\epsilon}$

$$\hat{\epsilon}_{M.F.} = \frac{\Delta_{12}(x)}{\langle \Gamma_2 \rangle - \langle \Gamma_1 \rangle}$$



Mean-Field Theory

Logit Differences

- Through derivations (see paper appendix), arrive at an approximation for $P(\Delta_{1j})$ at small Δ_{1j}

$$P(\Delta_{1j}) = C\Delta_{1j}^{j-2} + O(\Delta_{1j}^{j-1})$$

  where C is a network specific constant
- The earlier mean-field approximation for $\hat{\epsilon}$ implies that

$$P(\hat{\epsilon} \leq \epsilon) \approx P(\Delta_{12} \leq \epsilon(\langle \Gamma_2 \rangle - \langle \Gamma_1 \rangle)) = P(\Delta_{12} \leq \tilde{\epsilon})$$

  where $\tilde{\epsilon} = \epsilon(\langle \Gamma_2 \rangle - \langle \Gamma_1 \rangle)$
- Combining the two results,

$$P(\hat{\epsilon} < \epsilon) \approx P(\Delta_{12} < \tilde{\epsilon}) \approx C\tilde{\epsilon} + O(\tilde{\epsilon}^2)$$

Mean-Field Theory

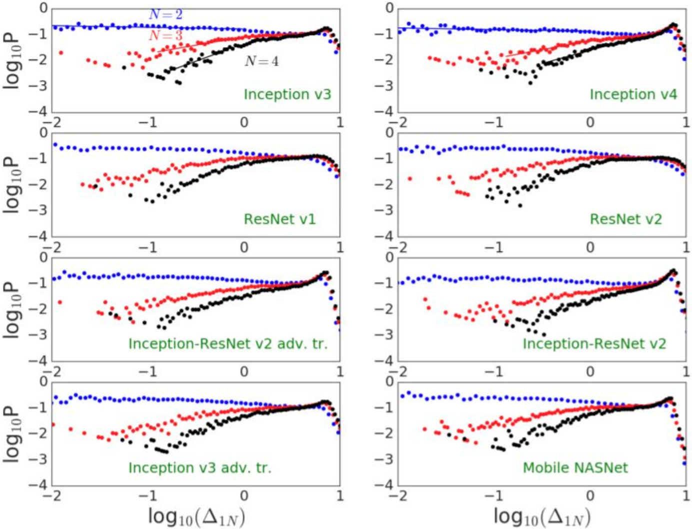
Logit Differences

- In order to further test the mean-field approximation, Δ_{1j} was evaluated for various neural network architectures and datasets
  - Δ_{12} does seem to have exponent of 0, for higher j, also power law, but typically not integral
  - Compared to randomly sampled i.i.d. logits
- Shows that adversarial error has power law form for various datasets and models; implies that commonality of adversarial examples is not due to depth of model or high dimensionality of data, but rather because difference between logit 1 and 2 is frequently small

Mean-Field Theory

Logit Differences

- Graphs of logit differences for models trained on ImageNet



Mean-Field Theory

Logit Differences

- Given the large density of small Δ_{12} values, propose new loss function to increase confidence of network at each sample to increase logit differences

$$\text{loss} = \text{old loss} - \Sigma_{i=1}^{n} p_i \log p_i$$

- Logit difference distribution was verified to have overall larger values

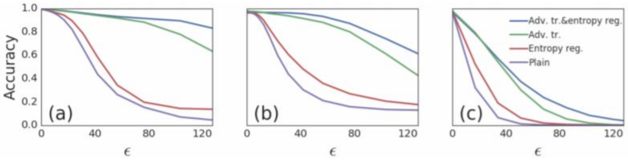


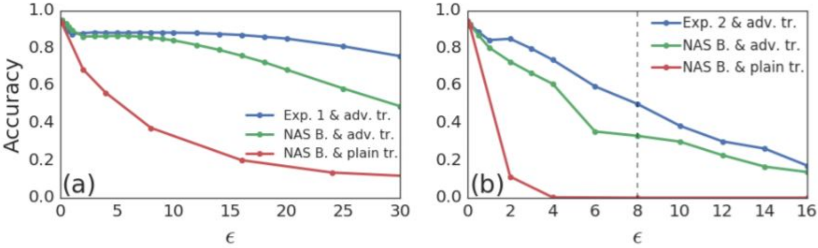
Figure 5: Step l.l. attack adversarial accuracy as a function of ϵ for CNN (a) and permutation invariant (b) MNIST, with regular training (purple), with entropy regularization (red), adversarial training (green), and adversarial training with entropy regularization (blue). Adversarial training was done using the step l.l. method. In (c), we show the PGD attack adversarial accuracy on permutation invariant MNIST trained with and without step l.l. adversarial training.

Network Architectures

- Recent papers suggest that larger networks are more resistant against adversarial examples regardless of adversarial training
- To test effects of architecture, conduct experiments
  - Neural architecture search (NAS) where child models are trained with clean and either step l.l. or PGD adversarial examples and reward is computed on validation set with FGSM adversarial accuracy
- Step l.l training
  - Child models trained for 10 epochs on training batches where half the samples are adversarially perturbed
  - Pick model with highest adversarial accuracy and enlarge by scaling up number of filters; train for 100 epochs on full training set for 12 different hyperparameter sets
  - Pick hyperparameter set with highest adversarial accuracy

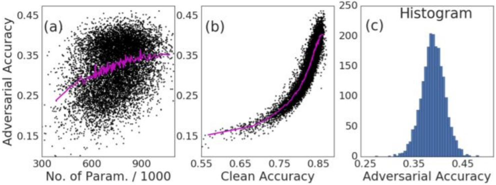
Network Architectures

- PGD Training
  - Follow procedures in Madry et al. (2017)
- Compared two experiments with baseline NAS which rewarded clean accuracy; NAS baseline had 4.9 million trainable parameters vs. 2.3 million (Exp. 1) and 3.5 million (Exp. 2)
- Graphs: left is step l.l adversarial accuracy, right is PGD



Network Architectures

- Next, examine performance statistics of 9,360 child models in Experiment 1 (only 10 epochs of training)
- Very little correlation between number of trainable parameters and adversarial accuracy
- High correlation between clean accuracy and adversarial accuracy
  - Explains why larger models are more robust against adversarial examples
- High clean accuracy insufficient; high variance of adversarial accuracy among high clean accuracy models



Summary

- The functional form of adversarial error and logit differences are universal at small ϵ
- Entropy regularization and better network architectures can help protect against adversarial examples
- Model architecture affects adversarial accuracy due to its effects on clean accuracy

References

- https://arxiv.org/pdf/1711.02846.pdf
- https://arxiv.org/pdf/1611.01232.pdf