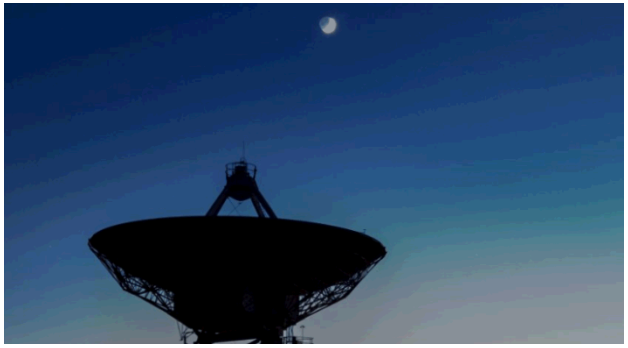




# Exploring Mitigations to Universal Adversarial Manipulation of Deep Neural Networks


**Lior Attias** and Tim Draelos

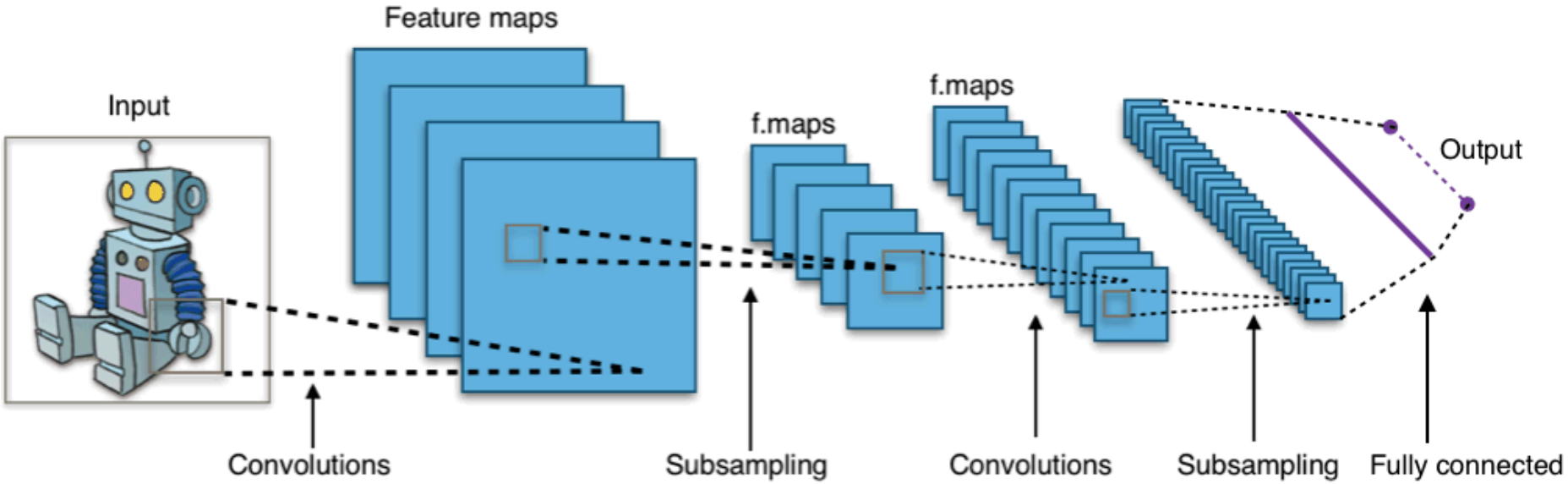
Funded by center 6300 CASA (Center for Analysis Systems and Applications) summer intern program



SAND2017-8755C

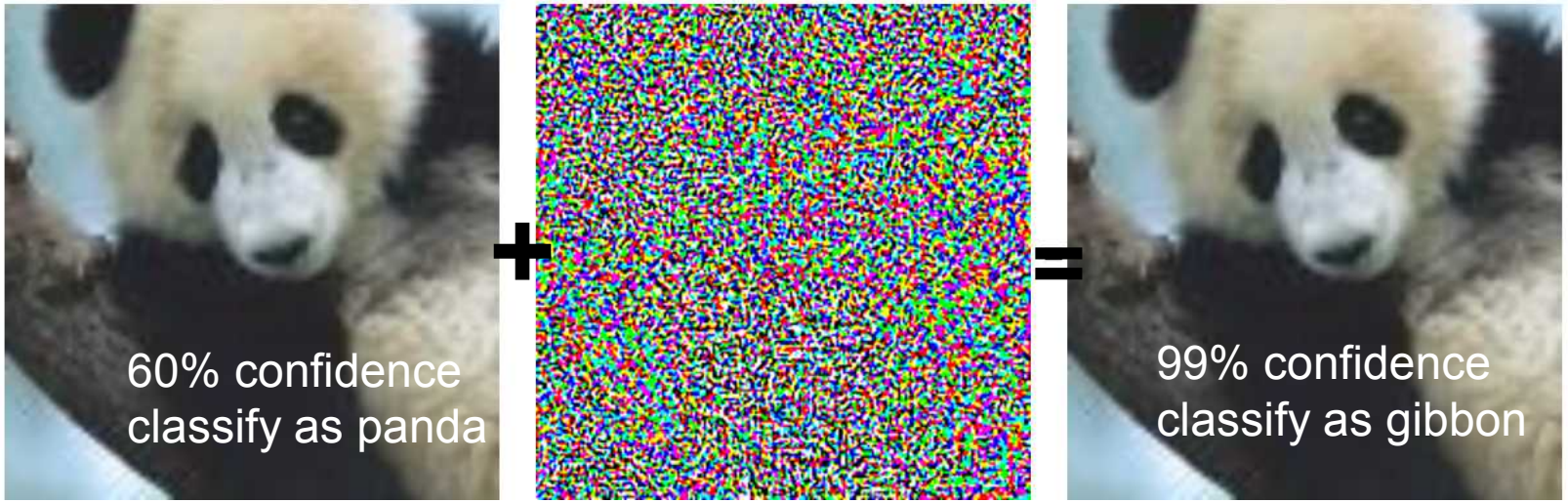
# Classifiers

- Many neural nets function as classifiers (GoogLeNet, ALEXnet, etc.)
- A central goal of classifiers is to “learn” to correctly classify novel data (images)
- Applications to image-based big-data: satellite information, video footage, etc.



# Perturbations

- Images are represented as complex matrixes where each number represents a RBG pixel value
- Because of this representation, you can add a vector to each image mathematically to change the computer's representation of an image, but retain human perception of the image
- This vector is known as a "perturbation"



# “Universal” perturbation*

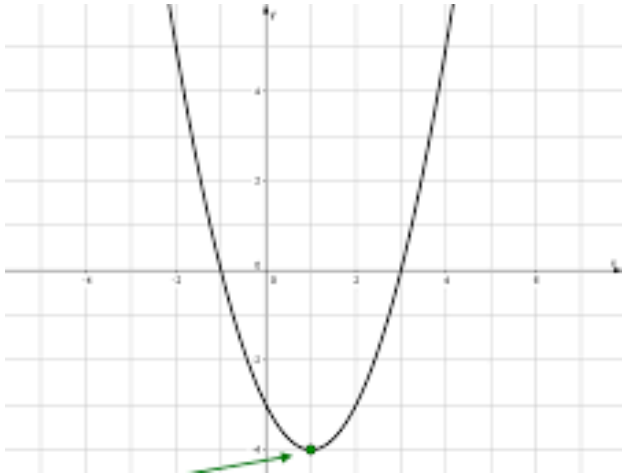

- If calculated correctly, the perturbation calculated on ONE set of images using ONE neural net will fool many images on many different neural nets.
  - A perturbation calculated on GoogLeNet will fool CaffeNet, VGG, etc.
  - A perturbation calculated on a subset of images will fool images outside that subset.

| | VGG-F | CaffeNet | GoogLeNet | VGG-16 | VGG-19 | ResNet-152 |
|---|---|---|---|---|---|---|
| VGG-F | **93.7%** | 71.8% | 48.4% | 42.1% | 42.1% | 47.4 % |
| CaffeNet | 74.0% | **93.3%** | 47.7% | 39.9% | 39.9% | 48.0% |
| GoogLeNet | 46.2% | 43.8% | **78.9%** | 39.2% | 39.8% | 45.5% |
| VGG-16 | 63.4% | 55.8% | 56.5% | **78.3%** | 73.1% | 63.4% |
| VGG-19 | 64.0% | 57.2% | 53.6% | 73.5% | **77.8%** | 58.0% |
| ResNet-152 | 46.3% | 46.3% | 50.5% | 47.0% | 45.5% | **84.0%** |

*S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in IEEE CVPR, 2017.

# Calculating the universal perturbation

- As you iterate through the list:
- Let v be the perturbation vector. Find $v_i$ by solving the optimization problem (cross the decision boundary
- Aggregate $v = v + v_i$



**Algorithm 1** Computation of universal perturbations.

**input:** Data points $X$, classifier $\hat{k}$, desired $\ell_p$ norm of the perturbation $\xi$, desired accuracy on perturbed samples $\delta$.
**output:** Universal perturbation vector $v$.
Initialize $v \leftarrow 0$.
**while** $\mathrm{Err}(X_v) \leq 1 - \delta$ **do**
  **for** each datapoint $x_i \in X$ **do**
    **if** $\hat{k}(x_i + v) = \hat{k}(x_i)$ **then**
      Compute the *minimal* perturbation that sends $x_i + v$ to the decision boundary:

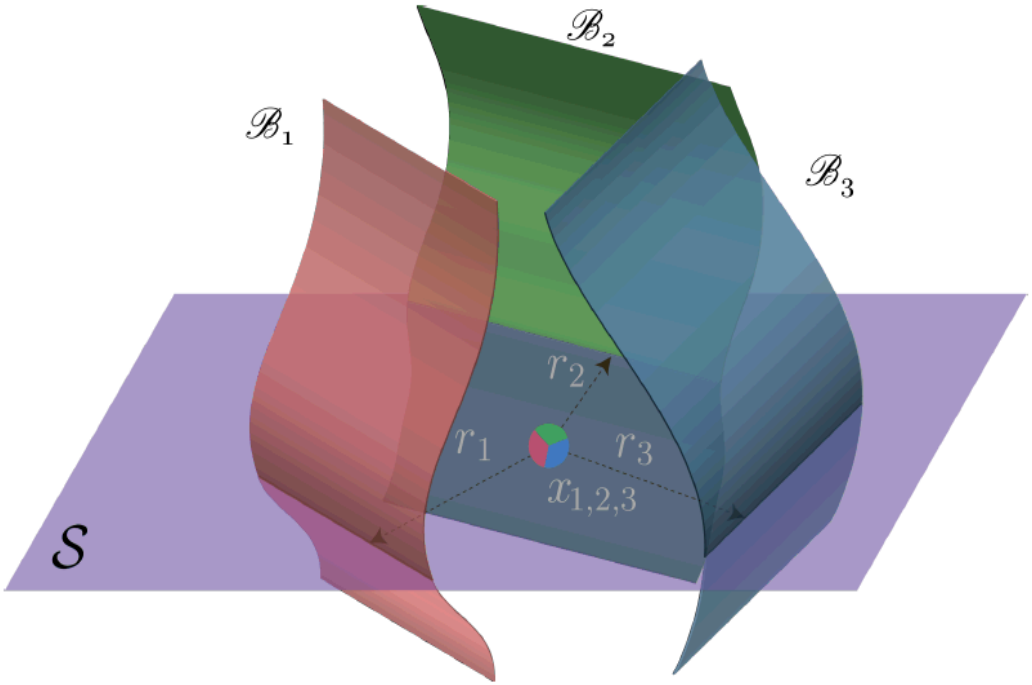
$$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$

      Update the perturbation:

$$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$

    **end if**
  **end for**
**end while**

# Visualizing the perturbation




S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in IEEE CVPR, 2017

# Our question


- Given a set X, apply the universal perturbation vector on each image in set X to create new set Xv
- How does the behavior of a **compressed** neural net compare to that of an **uncompressed** net when trying to classify the perturbed set Xv?
  - **Will the compressed net be able to overcome the perturbation and classify the images in Xv correctly?**
- How can we increase the **resiliency** of the net to the perturbation?

# Methods


- Select a known subset of images from ImageNet
- Perturb this subset using a pre-calculated universal perturbation to create a set of perturbed images.
  - Universal perturbation vector was calculated on GoogLeNet
- Pass the perturbed subset and the unperturbed subset to an uncompressed and a compressed neural net
  - Alexnet
- Compare percentage of mislabeled images for both the compressed and uncompressed images

# Results (compare to True)


- Percentage of correct classification as compared to **true** labels

| AlexNet | Images | Correct Classification | Incorrect Classification |
|---|---|---|---|
| Uncompressed | Normal | **68.30%** | **31.79%** |
| Uncompressed | Perturbed | **61.54%** | **38.46%** |
| Compressed | Normal | **69.58%** | **30.42%** |
| Compressed | Perturbed | **61.92%** | **38.08%** |

- **Conclusion**: Compressed network does not appear to improve (decrease) the vulnerability to universal perturbation
- Compressed network classifies slightly better than uncompressed, but no improved resiliency to perturbation
- Uncompressed: Pert. causes a 6.67% increase in misclassification
- Compressed: Pert. causes a 7.66% increase in misclassification

# Results (take AlexNet as “True”)


- Percentage of correct classification as compared to **AlexNet,** where **classification by AlexNet is taken to be the “true”** classification
- AlexNet uncompressed on normal images is taken to be 100% accurate for comparison

| AlexNet | Images | Correct Classification | Incorrect Classification |
|---|---|---|---|
| Uncompressed | Normal | **100%** | **0%** |
| Uncompressed | Perturbed | **77.75%** | **22.25%** |
| Compressed | Normal | **83.41%** | **16.59%** |
| Compressed | Perturbed | **71.66%** | **28.34%** |

# Discussion (AlexNet as "True")

| AlexNet | Images | Correct Classification (take AlexNet as True) | Incorrect Classification (take AlexNet as True) |
|---|---|---|---|
| Uncompressed | Normal | **100%** | **0%** |
| Uncompressed | Perturbed | **77.75%** | **22.25%** |
| Compressed | Normal | **83.41%** | **16.59%** |
| Compressed | Perturbed | **71.66%** | **28.34%** |

- Consider how the perturbation affects a network's classification
- Compressed: Pert. causes an 6.09% difference in classification
- Uncompressed: Pert. causes a 22.25% difference in classification

- The perturbation changes the classification decision boundary *within a specific network* in a more pronounced manner

# Discussion (Comparison)

| AlexNet | Images | Correct compared to True | Fail compared to True | Correct compared to AlexNet | Fail compared to AlexNet |
|---|---|---|---|---|---|
| Uncompressed | Normal | **68.30%** | **31.79%** | | |
| Uncompressed | Perturbed | **61.54%** | **38.46%** | **77.75%** | **22.25%** |
| Compressed | Normal | **69.58%** | **30.42%** | **83.41%** | **16.59%** |
| Compressed | Perturbed | **61.92%** | **38.08%** | **71.66%** | **28.34%** |

- When compared to True:
  - Perturbation only causes slight increase in misclassification
  - Compression only yields slight improvement in classification and slight increased vulnerability to perturbation
- Isolating the effects on the net:
  - Uncompressed net seems to be more vulnerable to perturbation (by 6.09%)
  - Compression seems to worsen classification slightly (16.59%)