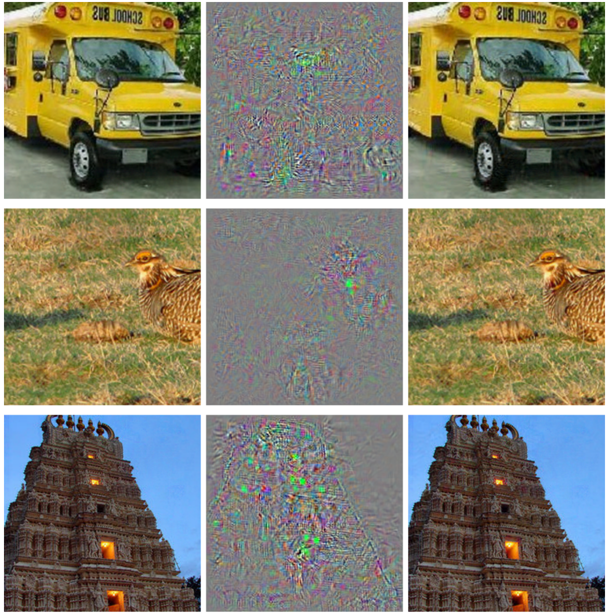
# Blind Spots in Neural Networks

by **Wojciech Zaremba**
with Christian Szegedy, Ilya Sutskever, Joan Bruna,
Dumitru Erhan, Ian Goodfellow, and Rob Fergus

# Blind Spots in Neural Networks

Correctly predicted object



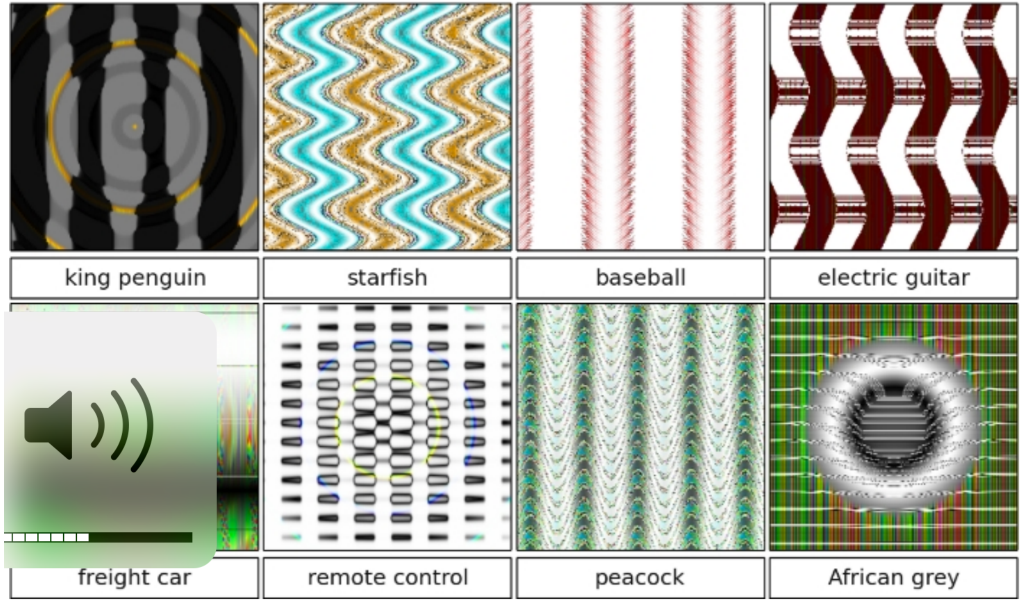
Predicts ostrich, Struthio camelus

# Blind Spots in Neural Networks

- Negative examples generated with Backpropagation

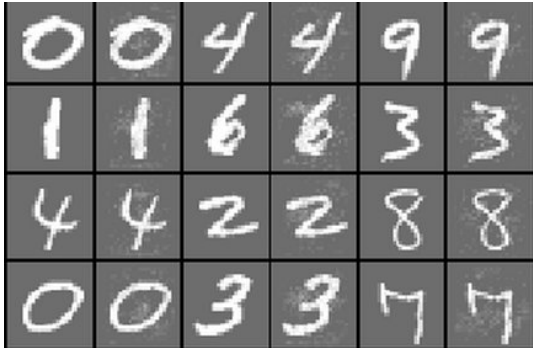
- Constrained to be in feasible set (proper color range)

# Other examples



ref.: Nguyen et al. “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images“

# Cross model transfer



| | Training Error |
|---|---|
| **Model A** | 0% |
| **Model B** | 0% |

| | **Negative examples for Model A** | **Negative examples for Model B** | **Gaussian noise std = 0.1** |
|---|---|---|---|
| **Model A** | 100% | 6.6% | 0% |
| **Model B** | 20.3% | 100% | 0% |

Different fully connected networks trained on MNIST dataset. Average distortions by ~6%.

# Cross training data transfer

| | **Training P1** | **Training P2** |
|---|---|---|
| **Model A** | 0% | 2.4% |
| **Model B** | 2.5% | 0% |

| | **Test distortion for A** | **Test distortion for B** |
|---|---|---|
| **Model A** | 100% | 6.25% |
| **Model B** | 26.2% | 100% |

Different fully connected networks trained on MNIST dataset. Distortions by ~6%.

# Possible explanations

- High-dimensional dot-product is unstable under small perturbations in every dimension.*

- Our linear operators are dominated by few high eigenvalues.

* ref.: Goodfellow et al. “Explaining and Harnessing Adversarial Examples”

# Conclusions

- Different networks share properties, which are dependent on statistics of training sets (not only particular samples).

- Can be negative examples used to improve generalization ?

# Q & A

- Adversarial example generation
- Cross model transfer
- Cross different training data transfer