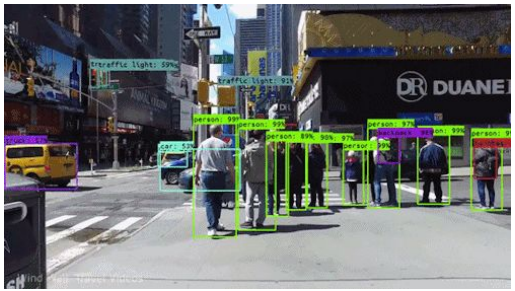
Perplexing Image Recognition Systems Using Mutation Generative Adversarial Networks

ROHAN ACHARYA

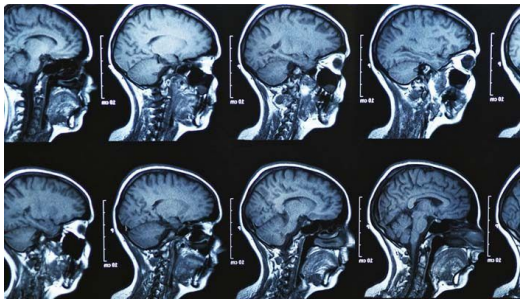
Background

- AI is becoming a pervasive and disruptive force
- Safety-critical applications rely on deep learning for image classification



Self-driving Cars

Facial-recognition ATMs



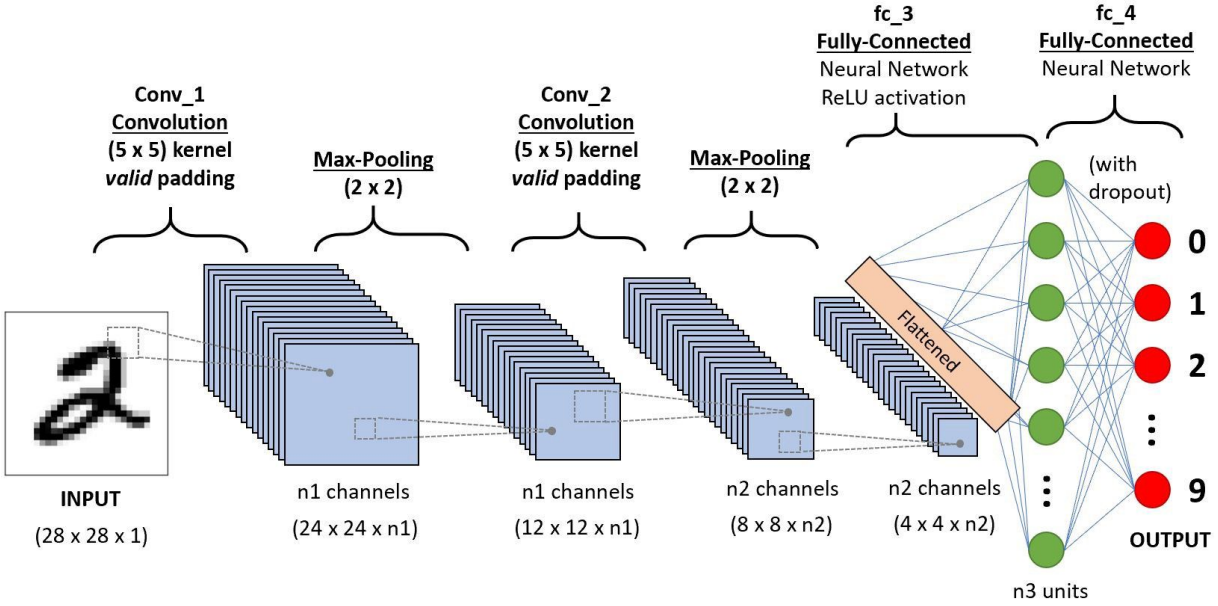
Medical Imaging Systems

What if an AI is hacked? What if an AI goes rogue?

April 2019: Keen Security Lab researchers trick a Tesla Model S in Autopilot mode to swerve towards oncoming traffic

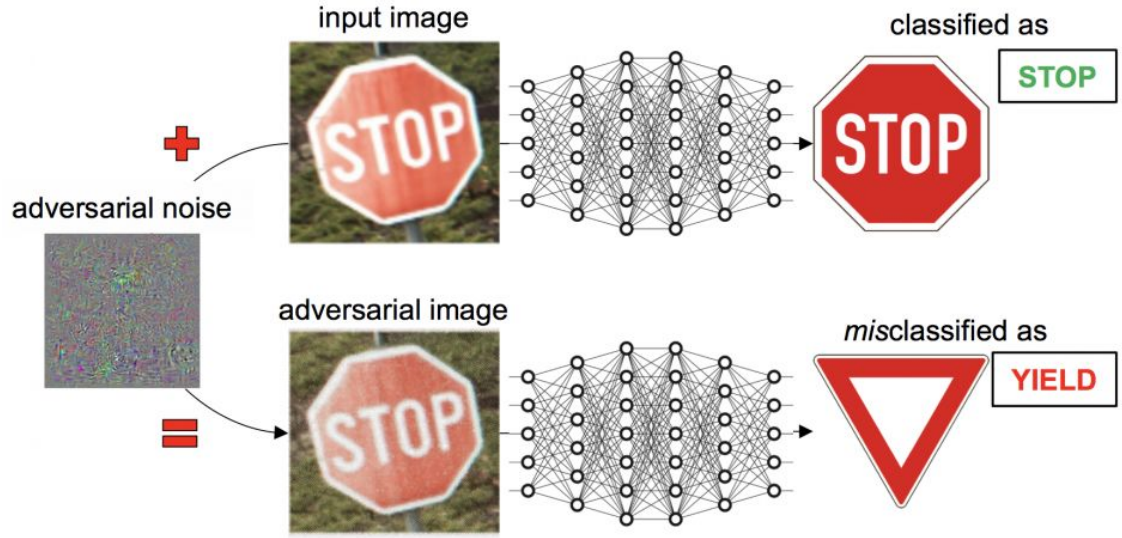
Image Recognition Systems

Convolutional Neural Networks (CNNs)[*LeCun, 1998*]



Adversarial Machine Learning[*Szegedy, 2014*]

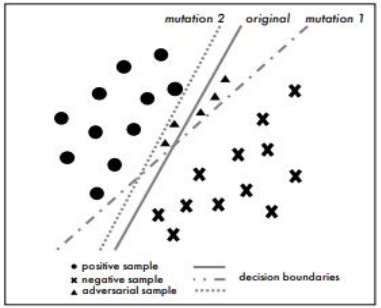
Fooling machine learning models using “adversarial samples” or intentionally perturbed inputs



Model Mutation Testing[*Wang, 2018*]

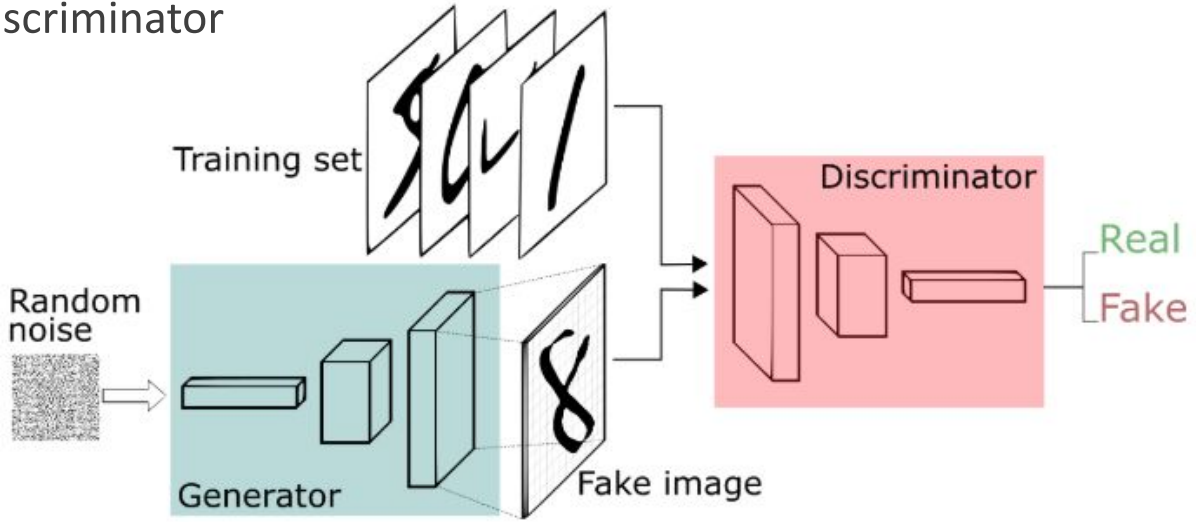
- Detect adversarial samples through the generation of mutant CNNs obtained by applying mutation operators[*Ma, 2018*]
- Label Change Rate (LCR)
  Measures how sensitive an input sample is to mutations imposed on the target model (should be larger for adversarial samples)

Mutation Operator	Level	Description
Gaussian Fuzzing (GF)	Weight	Fuzz weight by Gaussian Distribution
Weight Shuffling (WS)	Neuron	Shuffle selected weights
Neuron Effect Block. (NEB)	Neuron	Block a neuron effect on following layers
Neuron Activation Inverse (NAI)	Neuron	Invert the activation status of a neuron
Neuron Switch (NS)	Neuron	Switch two neurons of the same layer
Layer Deactivation (LD)	Layer	Deactivate the effects of a layer
Layer Addition (LA_m)	Layer	Add a layer in neuron network
Act. Fun. Remov. (AFR_m)	Layer	Remove activation functions



Generative Adversarial Networks (GANs)[*Goodfellow, 2014*]

- Machine learning model used for image generation
- Two-party architecture
  - Generator
  - Discriminator



- State-of-the-Art Defense Mechanisms
  - Detect adversarial samples by taking advantage of their sensitivity to random mutations imposed on the target model
- **Gap:** Current adversarial sample generation methods yield samples that are easily detected, giving a false sense of security!
  - Mutation Generative Adversarial Network (MGAN)
  - Possible to generate more robust adversarial samples i.e. malicious inputs will be misclassified as safe
- **Statement of Purpose:** Improve the security of safety-critical image recognition systems by exposing this vulnerability

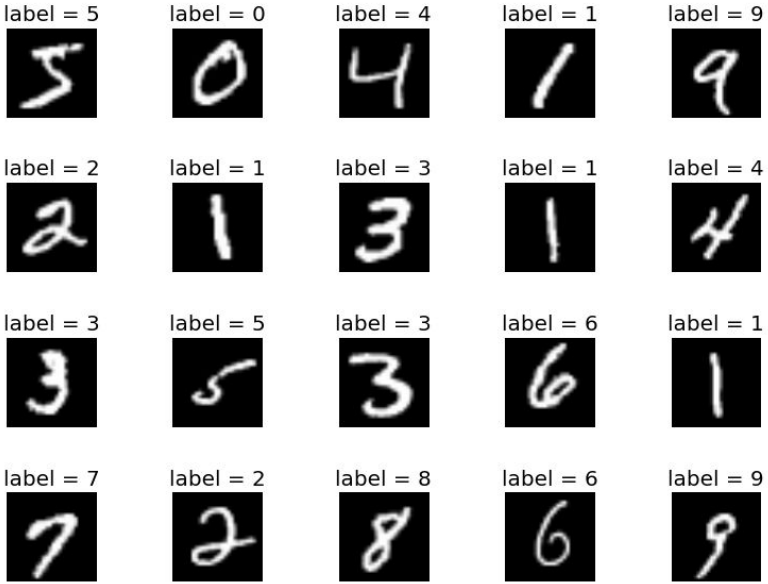
Hypothesis

By wiring mutants as a third component in the original two-party architecture of GANs, we can guide the generation of mutation-consistent yet effective adversarial samples

Data

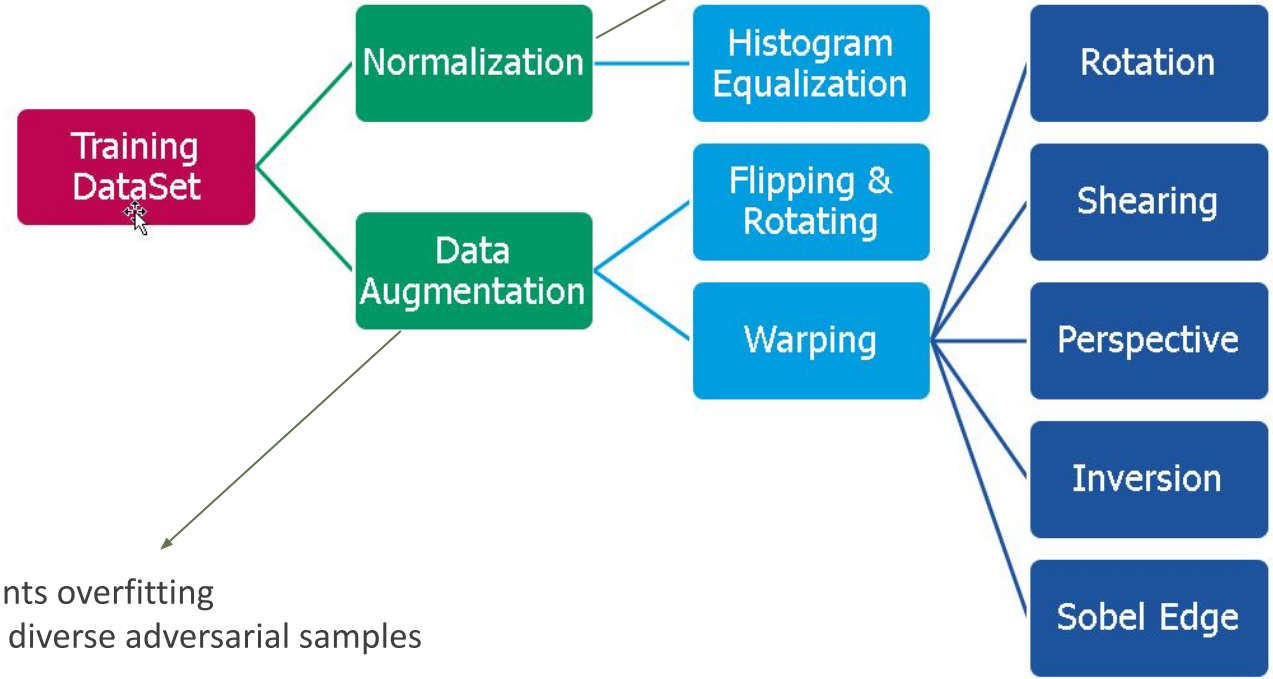
MNIST Dataset[*LeCun*]

- Handwritten digits from 0-9
- Training set of 60,000 samples
- Test set of 10,000 samples



Data Preprocessing

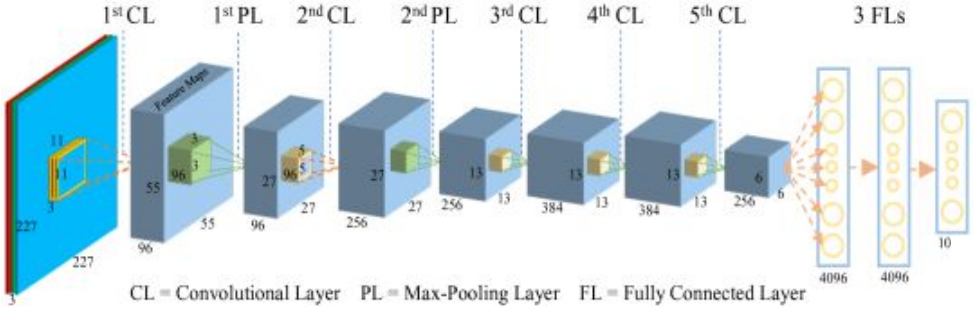
Insensitive to magnitude
Faster convergence



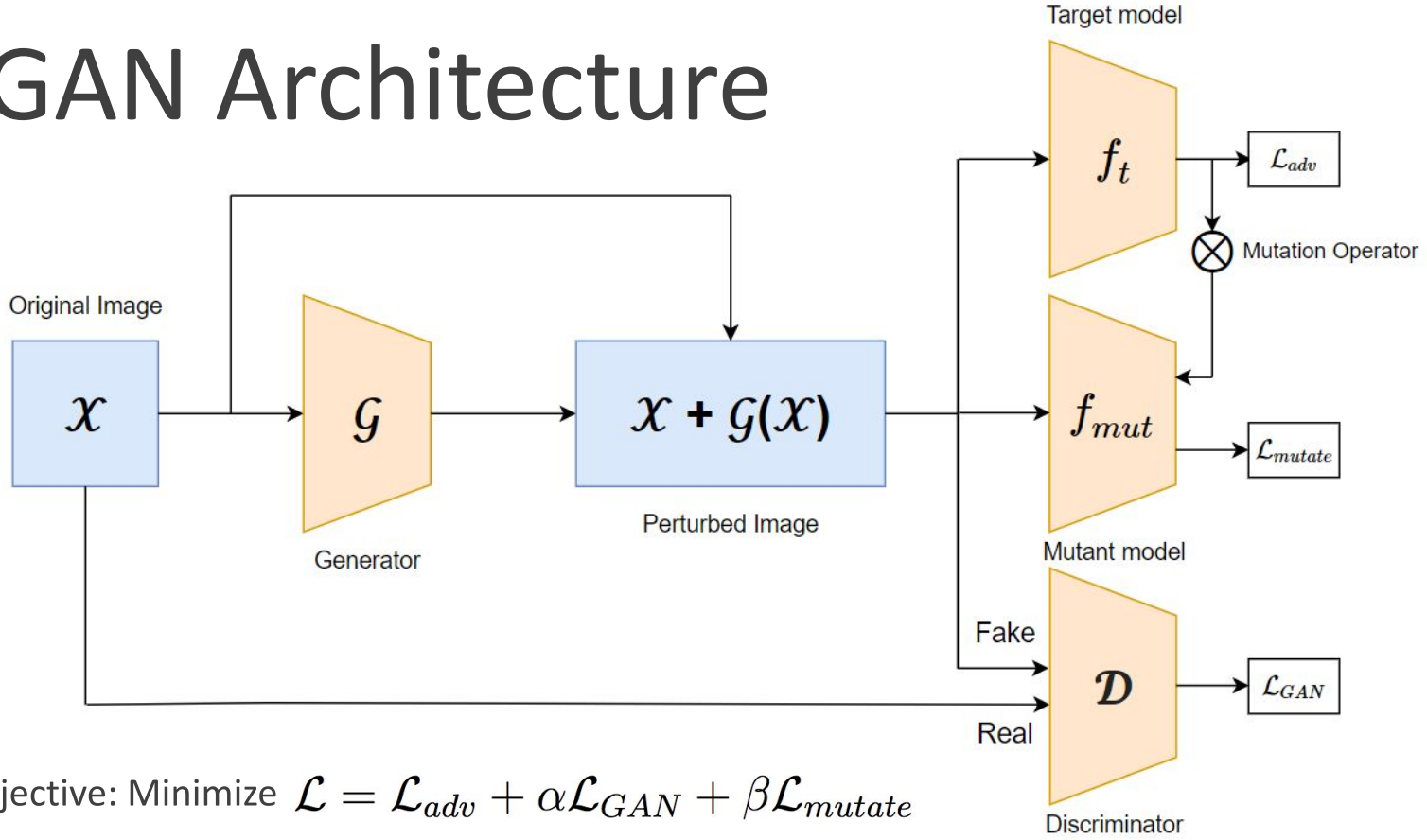
Prevents overfitting
More diverse adversarial samples

Training the Target Model

- Implemented using PyTorch, a Python-based machine learning library
- Trained on PC with RTX 2080-Ti and Intel i-9 9900k
- 99.1% accuracy on testing set



MGAN Architecture



Final objective: Minimize $\mathcal{L} = \mathcal{L}_{adv} + \alpha\mathcal{L}_{GAN} + \beta\mathcal{L}_{mutate}$

Evaluation Criteria

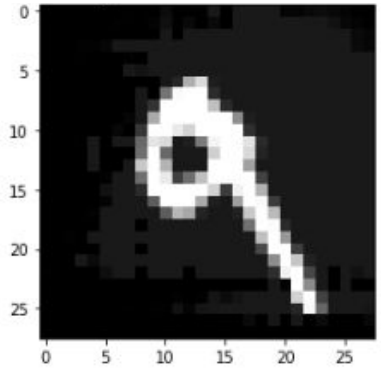
- Compared the results of MGAN and AdvGAN (state-of-the-art)[*Xiao, 2018*]
- Measured:
  - **Accuracy**: the misclassification rate of the adversarial samples
  - **Label Change Rate**: percentage of adversarial samples that are classified differently by the target model after a mutation is applied

Comparable

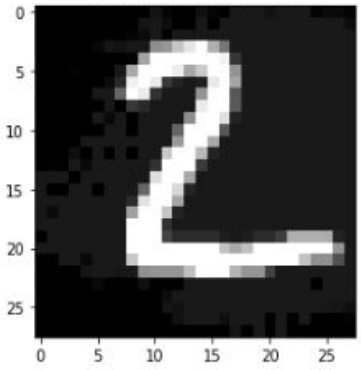
Indistinguishable
Much better than AdvGAN

Mutation Operator	MGAN Accuracy	AdvGAN Accuracy	MGAN LCR	AdvGAN LCR	Normal LCR
WS	96.83%	98.96%	12.21%	24.56%	11.52%
NAI	93.41%		44.83%	77.70%	41.50%
NS	98.21%		7.38%	18.39%	6.57%
GF	97.23%		16.21%	30.28%	15.67%

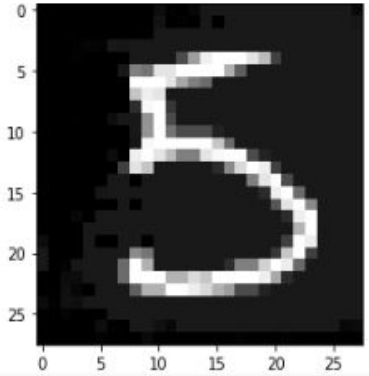
classified as a 3



classified as a 0



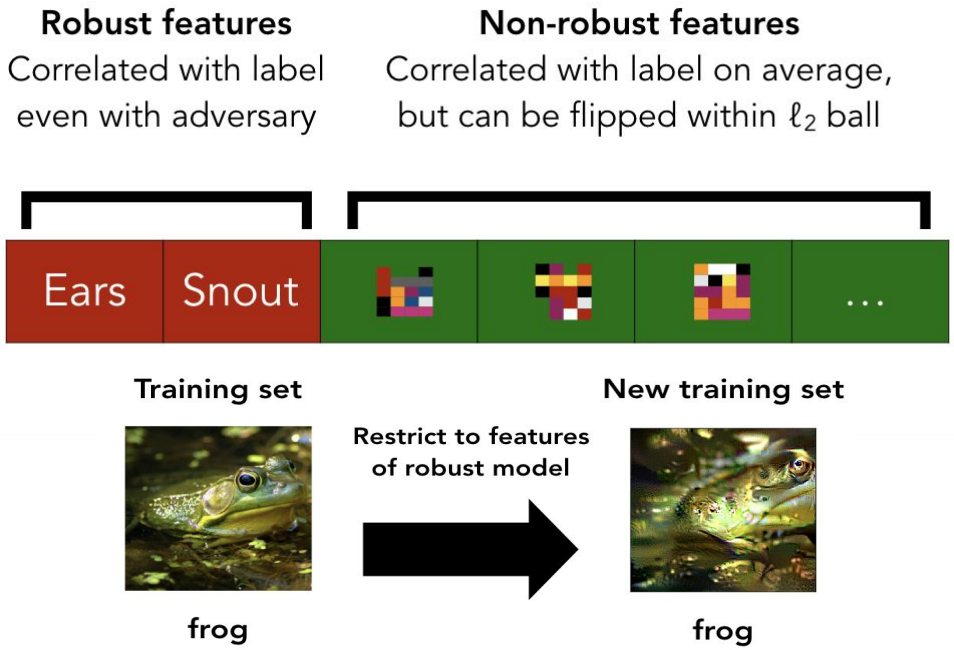
classified as a 6



Conclusion

- Hypothesis supported!
  - MGAN was able to produce effective and mutation-consistent adversarial samples that were undetectable by model mutation testing
- Adversarial samples pose a real threat to AI security
- Does this mean that we should stop deploying AI?
  - No!

Future Steps



Acknowledgements

I would like to thank my mentor for introducing me to the field of adversarial machine learning, guiding me through the initial literature search, and providing me with crucial assistance in the brainstorming phase.

I would also like to thank my science research teachers for providing me with unconditional support throughout the research process.

Thank you!

Questions?

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[2] George Saon, Hong-Kwang J. Kuo, Steven Rennie, and Michael Picheny. The IBM 2015 English Conversational Telephone Speech Recognition System. *arXiv e-prints*, page arXiv:1505.05899, May 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv e-prints*, page arXiv:1506.01497, Jun 2015.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv e-prints*, page arXiv:1312.6199, Dec 2013.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, page arXiv:1412.6572, Dec 2014.

[6] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. *arXiv e-prints*, page arXiv:1608.04644, Aug 2016.

[7] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating Adversarial Examples with Adversarial Networks. *arXiv e-prints*, page arXiv:1801.02610, Jan 2018.

[8] Jingyi Wang, Guoliang Dong, Jun Sun, Xinyu Wang, and Peixin Zhang. Adversarial Sample Detection for Deep Neural Network through Model Mutation Testing. *arXiv e-prints*, page arXiv:1812.05793, Dec 2018.

[9] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. *arXiv e-prints*, page arXiv:1811.03378, Nov 2018.

[10] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. *arXiv e-prints*, page arXiv:1511.04599, Nov 2015.

[11] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. *arXiv e-prints*, page arXiv:1905.02175, May 2019.

Perplexing Image Recognition Systems Using Mutation Generative Adversarial Networks

Background

Review of Literature

- Image Recognition Systems
- Adversarial Machine Learning
- Model Mutation Testing
- Generative Adversarial Networks

Problem & Hypothesis

- AI Security Risk
- Hypothesis

Methods & Materials

- Data
- Data Preprocessing
- Training the Target Model
- MGAN Architecture

Conclusion

Future Steps