Proving Adversarial Robustness of Deep Neural Networks

Aiswarya G Raj
CET-IS
IIST INTERN

Guided by:-
Dr. Deepak Mishra
Dr. Mahesh Sreekumar Rajasree

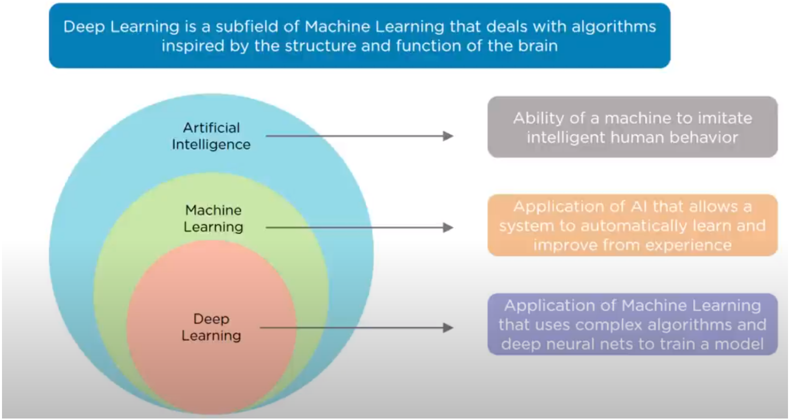
AVIONICS

IIST

Outline

1. Introduction
2. Comprehended Paper
3. Implementation and Output
4. Conclusion
5. Future Enhancement
6. Internship Summary

What is Deep Learning?



Deep Learning is a subfield of Machine Learning that deals with algorithms inspired by the structure and function of the brain

Ability of a machine to imitate intelligent human behavior

Application of AI that allows a system to automatically learn and improve from experience

Application of Machine Learning that uses complex algorithms and deep neural nets to train a model

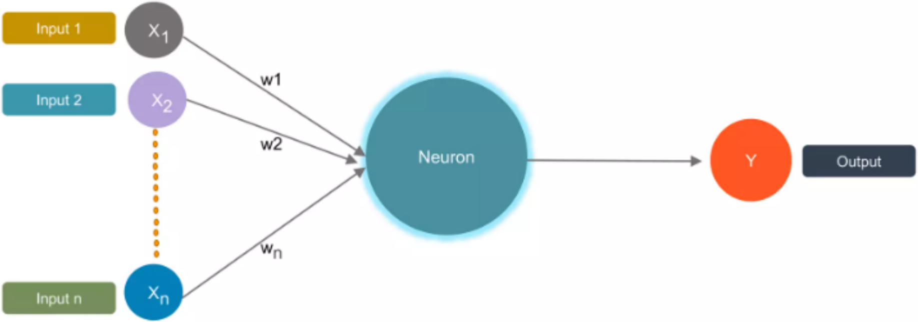
Why do we need Deep Learning?

- Process large amount of data
- Handle complex algorithms
- Achieve best performance with large amount of data
- Feature extraction

Applications of Deep Learning

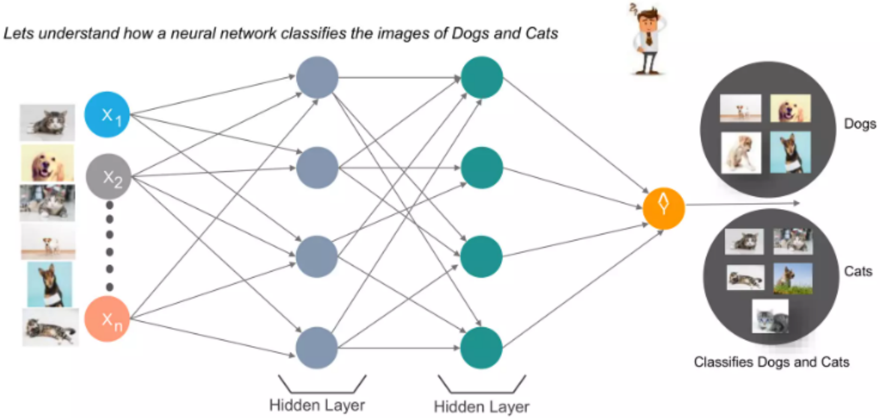
- Detection of Cancer
- Navigation of Robotics
- Machine translation
- Music composition
- Colorization of images
- Autonomous driving cars

What is a Neural Network?



- Deep learning works on neural network.
- Neural network is designed based on human brain.
- Smallest component of neural network is neuron.
- Neuron consists of small central unit which receives the input.
- Each input value is multiplied with a weight and is added with a bias value.
- Output from Step 1 (Transfer function) is taken as the input to the Step 2 (Activation function).
- Step 2 produces the final output which is a binary value.

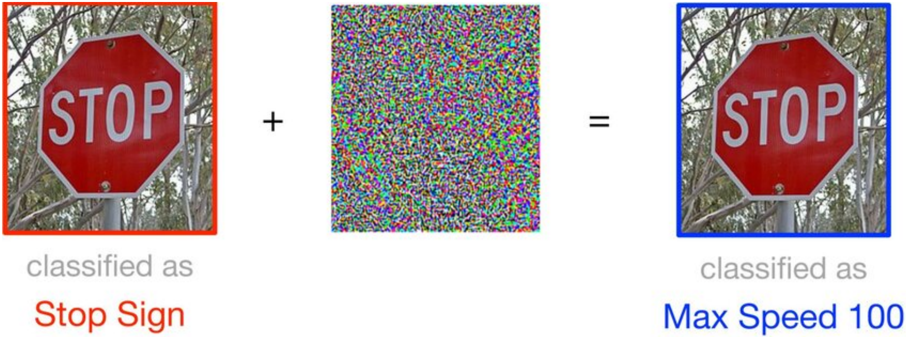
Working of a Neural Network


Lets understand how a neural network classifies the images of Dogs and Cats
X_1
X_2
X_n
Hidden Layer
Hidden Layer
Dogs
Cats
Classifies Dogs and Cats

- Provide labelled training data
- Training:
  - Iterative process
  - Input is given, weights and bias keep changing during each iteration
- Testing:
  - Fresh set of inputs are fed into the network
  - Checking whether accuracy is getting similar or not
  - Accuracy become maximum when the values for weight and bias become optimal

Adversarial Robustness

- Adversarial robustness refers to the ability of a machine learning model to maintain its performance and accuracy when exposed to adversarial examples.
- Adversarial examples are data points which are indistinguishable from our dataset but lead to wrong predictions.
- Adversarial inputs refers to inputs specifically designed to deceive a Deep Neural Network.
- These inputs are generated by making small changes to correctly classified inputs, causing the DNN to misclassify them.
- For example, an image of a stop sign.


classified as
Stop Sign
+
=
classified as
Max Speed 100

Reluplex

- Reluplex is an extension of the Simplex algorithm specifically designed to handle the piecewise linear nature of ReLU functions.
- It enables effective verification of ReLU-based neural networks.
- For each ReLU node, it introduces two linear constraints: $y \geq 0$, $y \geq x$.
- The algorithm determines whether the ReLU node is active (y=x) or inactive (y=0).
- It enhances the safety and reliability of AI systems.
- It can verify local robustness for networks with hundreds of nodes, but global robustness is limited to smaller networks with a few dozen nodes due to the increased complexity.

Comprehended Paper

“Towards proving the adversarial robustness of deep neural networks”

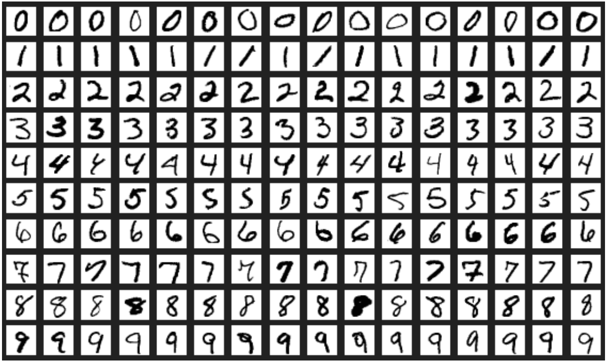
- It explores how neural networks can resist adversarial attacks by studying existing research on adversarial robustness and input perturbations designed to deceive the network.
- This paper focuses on understanding and proving the robustness properties of neural networks, ensuring they maintain performance and accuracy even when faced with adversarial inputs.
- The study investigates different types of robustness in neural networks, including resistance to adversarial attacks and overall stability across various conditions.

- The paper addresses practical aspects of verifying robustness in real-world networks, focusing on scalability and computational feasibility of the techniques.
- The authors present initial findings on establishing robustness properties specifically for ACAS Xu networks used in unmanned aircraft collision avoidance systems.
- The paper identifies unresolved research issues and outlines future plans involving new techniques, methodologies, or experiments to address these challenges.

Implementation and Output

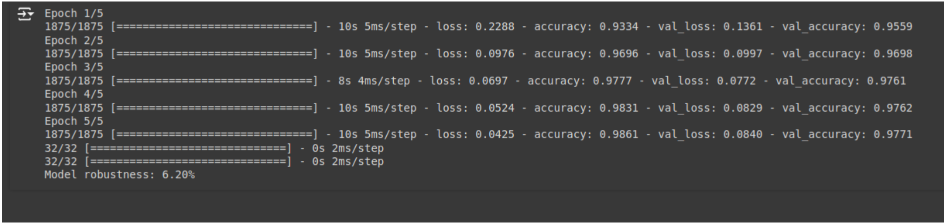
- The main objective of the code is the process of creating a neural network, generating adversarial examples to test the network and verifying the network's robustness against these adversarial examples.
- MNIST dataset is used to define, compile, and train a neural network model for a classification task.
- The dataset is relatively small and easy to handle, making it suitable for quick experiments and prototyping.
- The dataset is easily accessible through various machine learning libraries such as TensorFlow and Keras.

- Here's an example of how the images and labels look in the MNIST dataset:



- The Fast Gradient Sign Method (FGSM) is an algorithm used to generate adversarial examples.
- It is commonly used to evaluate the robustness of machine learning models, especially neural networks, against adversarial attacks.
- Robustness is evaluated by comparing the model's predictions on original inputs versus adversarially perturbed inputs.

Output



```
Epoch 1/5
1875/1875 [==============================] - 10s 5ms/step - loss: 0.2288 - accuracy: 0.9334 - val_loss: 0.1361 - val_accuracy: 0.9559
Epoch 2/5
1875/1875 [==============================] - 10s 5ms/step - loss: 0.0976 - accuracy: 0.9696 - val_loss: 0.0997 - val_accuracy: 0.9698
Epoch 3/5
1875/1875 [==============================] - 8s 4ms/step - loss: 0.0697 - accuracy: 0.9777 - val_loss: 0.0772 - val_accuracy: 0.9761
Epoch 4/5
1875/1875 [==============================] - 10s 5ms/step - loss: 0.0524 - accuracy: 0.9831 - val_loss: 0.0829 - val_accuracy: 0.9762
Epoch 5/5
1875/1875 [==============================] - 10s 5ms/step - loss: 0.0425 - accuracy: 0.9861 - val_loss: 0.0840 - val_accuracy: 0.9771
32/32 [==============================] - 0s 2ms/step
32/32 [==============================] - 0s 2ms/step
Model robustness: 6.20%
```

- The model trains for 5 epochs, improving in both training and validation accuracy while reducing loss.
- The final accuracy on the validation set is high, at around 97.71%.
- The model's robustness is relatively low at 6.20%, suggesting that it might still be susceptible to adversarial examples or perturbations.

Conclusion

- Deep Learning solves complex problems, achieves high accuracy in tasks like image and speech recognition.
- Neural Networks are computational models inspired by the human brain, consisting of interconnected layers of nodes.
- Adversarial Robustness is the ability of a neural network to resist adversarial attacks.
- Reluplex is a tool for verifying properties of neural networks, specifically those using ReLU activation functions.
- Successfully trained model with high validation accuracy (97.71%) and noted a robustness score of 6.20%, indicating areas for improvement in adversarial defense.

Future Enhancement

- Develop and integrate advanced defense mechanisms to enhance model robustness against adversarial attacks.
- Enhance the scalability of verification tools like Reluplex to handle larger neural networks efficiently.
- Train and test the model on more diverse and complex datasets to ensure robustness across various data distributions.
- Implement real-time detection systems to identify and mitigate adversarial attacks during model deployment.
- Explore the application of adversarial robustness techniques in other domains, such as finance, healthcare, and cybersecurity.

Internship Summary

- Gained in-depth knowledge about deep learning, neural networks, and their applications in various fields.
- Studied adversarial robustness and tools like Reluplex to verify the robustness of neural networks.
- Implemented a neural network using the MNIST dataset and generated adversarial examples to test robustness.
- Achieved high validation accuracy (97.71%) but identified low robustness (6.20%), indicating the need for further enhancements.
- Identified potential future enhancements, including better defense mechanisms, scalability improvements, and application across diverse datasets.

Thank you!