International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Theoretical Analysis and Review of Adversarial Robustness in Deep Learning

Karthick Kumaran Ayyalluseshagiri Viswanathan

Member, IEEE

Abstract: ***Deep learning and neural networks are widely used in many recognition tasks including safety critical applications like self-driving cars, medical image analysis, robotics, etc., and have shown significant potential in various computer vision applications. The performance and accuracy of the deep learning models is highly important in safety critical systems. Recently some researchers have disclosed that deep neural networks are vulnerable to adversarial attacks. This paper talks about the adversarial examples, analyzes how adversarial noise can affect the performance and accuracy of deep learning models, potential mitigation strategies and the uncertainties in the deep learning models.***

Keywords: Deep learning, Adversarial Example, Adversarial Attack, Uncertainty, Bayesian Inference

1. Introduction

Deep neural networks have shown wide adoption in various computer vision techniques including safety critical applications like autonomous vehicles, medical image segmentation and surgery assistance etc., The performance of the deep learning models is highly important not just for the safety critical tasks but also for the general computer vision applications. Various articles and news articles quote that the deep neural networks beat human accuracy but the reality is that the deep learning models can be fooled easily to make wrong outputs or misclassify by making small changes in the input data that neither human eyes nor monitoring systems can notice easily.

2. Review on Adversarial Examples

This section reviews the paper by Goodfellow et al on adversarial examples and summarizes the key findings.

This paper discovers that the machine learning models including neural networks are vulnerable to adversarial examples. Adversarial examples make the ML models to misclassify examples that are only slightly different from correctly classified examples. Linear behavior in high dimensional spaces is sufficient to cause adversarial examples and that helped to design the fast method to generate adversarial examples which helps the adversarial training. It also shows that the adversarial training can provide an additional regularization benefit beyond dropout.

The precision of the input feature is limited and we can make many infinitesimal changes to the input that add up to one large change to the output. This shows that the linear model can have adversarial examples if its input has sufficient dimensionality.

The linear view of adversarial examples provides a fast way to generate them. Many networks like LSTM, ReLUs and maxout networks are designed in linear ways so that they are easier to optimize. The “fast gradient sign method” by linearizing the cost function is used to generate adversarial examples. This method causes multiple models to misclassify the input and the paper shows the demonstration on ImageNet with an error rate of 99.9%. They also proposed another simple method of rotating input by a small angle in the direction of the gradient produces adversarial examples. These simple algorithms can generate misclassified examples in the linear way and speeds up the adversarial training.

Goodfellow's paper considers both logistic and multiclass softmax regression for adversarial training with the fast gradient sign method for analyzing how adversarial training impacts weight decay. It has been found that the adversarial training will worsen underfitting. Thus, weight decay is viewed as being worst case in the underfitting scenarios.

A neural network can be regularized to some extent with the mixture of adversarial and clean examples. Training on adversarial examples is different from data augmentation techniques like transformations in the test set. Adversarial examples are inputs that are unlikely to occur naturally. They found that the training with the adversarial function based on the FGSM was an effective regularizer. This approach helps to reduce the error rate from 0.94% without adversarial training to 0.84% with adversarial training. They also observed that the error rate was not reaching zero on adversarial examples on the training dataset and fixed them by making the model larger and increasing the number of epochs. After fixing the model became resistant to adversarial examples and the error rate fell to 17.9% from 89.4%.

Here is the summary of the key findings from the Goodfellow et al paper:

- The adversarial training procedure can be seen as minimizing the worst-case error when the data is perturbed by an adversary.
- Linear models are easy to optimize and easy to perturb.
- Adversarial training can result in regularization.
- Adversarial perturbations yield the best regularization when applied to the hidden layers.
- Adversarial training is useful only if the model has the capacity to learn to resist adversarial examples.

3. Analysis of Adversarial Noise vs Performance

Adversarial noise can significantly affect the performance of deep learning models in various safety critical applications like self-driving cars, medical image analysis and health care. Several practical deep learning models have been trained with specific training data collected from a particular type of sensor. If either the sensor is changed or the input is perturbed the model will misclassify or produce wrong outputs which can cause chaos to human fatalities. An example of chaos is that the Waymo cars honk each other in the morning in the San Francisco parking lot disturbing neighbors. There were few robotaxi incidents not recognizing traffic signs and humans. These are a few examples of the failure of deep learning algorithms. Adversarial noise can misclassify signs board with something else, not stopping at red signals and stop signs, wrong diagnosis of the disease.

4. Mitigation Strategies

Adversarial attacks are minor perturbations in an image that can easily confuse the classifier. These perturbations are very small changes in the input image that the human eyes cannot notice. These attacks are in general white box attacks meaning that they use the model parameters to perturb the images. There are few mitigation strategies that work on imperceptible adversarial attacks.

- Integrating image acquisition and recognition processing stages in the deep learning network. This is the stage where any malicious software can embed perturbations into the image to attack the DNN.
- Another mitigation strategy proposed by Zhu and Ziang is to carry out multi-path synchronous prediction. If the results of the multi-path prediction are different, the original input might have been perturbed by adversarial attacks.
- Another recent development in the mobile and safety critical systems is the implementation of the secure camera. With a secure camera end-to-end pipeline, the image buffer cannot be altered by any non-secure entity like the malicious software. The secure camera buffer can only be written by the camera sensor and can only be read by the DNN model.
- Encrypting the input image will also help in mitigating the adversarial attacks significantly as modifications to the encrypted images will make the model identify the perturbations easily.

5. Challenges in Training on One Data Set and Testing on Another Data Set

Traditional machine learning paradigms are based on the assumption that both training and test data follow the same statistical pattern. This is called in-distribution. This scenario occurs when we have a lot of training data for a set of scenarios and try to generalize the model for all kinds of similar scenarios. An example of this is training a satellite remote sensing model with training data from North America and testing the data from Africa. During training the model is trained only with limited Africa data. When the model is provided with unseen data from Africa it performs poorly. This data from Africa is called the out-of-distribution. One of the biggest challenges in designing the deep learning model is to think about how to generalize the model so that it can provide accurate predictions against the OOD data.

6. Transfer Learning to Improve Performance

Transfer learning is a machine learning method where a pre-trained model can be used as a starting point for a new task. This model can achieve higher performance with transfer learning than training with only a small amount of data. In medical imaging, transfer learning can be used for tumor detection, disease diagnosis by fine tuning the pre-trained models on large image datasets. Pre-trained models like ResNet or Inception can be fine-tuned on new image dataset for image segmentation, object detection etc.

7. Uncertainty in Deep Learning Models

There are two main types of uncertainty in deep learning models: aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty arises from the randomness or noise inherent in the data which could be sensor noise, pixel noise. This means that our training data is not perfect in representing the true relationship between the input and the target. Epistemic uncertainty arises from the lack of knowledge about the model or its parameters. Epistemic uncertainty is associated with the model structure.

8. Why Aleatoric Uncertainty Alone is Insufficient

Aleatoric uncertainty arises from the noise inherent in the data. This is something that cannot be reduced even with more training data. The uncertainty in the image acquisition process can be due to various factors like instrument errors, thermal impacts, transmission errors etc., So we also use epistemic uncertainty to make the model more robust. Epistemic uncertainty can be reduced with more training data.

9. Bayesian Neural Networks to Improve Deep Learning Model Robustness

A Bayesian Neural Network is a stochastic artificial neural network trained using Bayesian inference models. The main goal of bayesian neural networks is to obtain a better idea of the epistemic uncertainty associated with the underlying model. This is usually accomplished by comparing the predictions of multiple sampled model parameterizations. If the multiple models agree then the uncertainty is low and if they disagree then the uncertainty is high.

10. Conclusion

This paper started with a review of the adversarial examples and explained how adversarial noise can affect the performance of the deep learning models. It also discussed the potential mitigation strategies for the adversarial attacks. Challenges in training the deep learning model with limited data and how transfer learning potentially improves the performance of the models have been explained. Finally, it

walks through the uncertainties in the deep learning models and how Bayesian neural networks can be utilized to estimate uncertainties for making the models more robust.

References

[1] Goodfellow et al. (2015) Goodfellow, I.J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR) 2015.
[2] Zhu and Jiang (2021) Zhu, Y., & Jiang, Y. (2021). Imperceptible adversarial attacks against traffic scene recognition, Soft Computing, 25(19), 13069-13077
[3] Zhang et al (2021) Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z. (2021). Deep Stable Learning for Out-Of-Distribution Generalization
[4] Christopher M. Bishop, Pattern Recognition and Machine Learning
[5] https://abc7news.com/post/waymo-cars-honk-each-other-night-disturbing-san-francisco-neighbors/15179709/
[6] https://www.reuters.com/business/autos-transportation/how-gms-cruise-robotaxi-tech-failures-led-it-drag-pedestrian-20-feet-2024-01-26/