Inatel
Instituto Nacional de Telecomunicações

IoT Research Group
Inatel - Brazil

Physical Adversarial Attacks Against End-to-End Autoencoder Communication Systems

Mauro A. A. Da Cruz[1] and Sheila C. S. Janota[1]
maurocruzter@gmail.com and sheilajanota@hotmail.com

[1] Instituto Nacional de Telecomunicações (INATEL), Santa Rita do Sapucaí-MG, Brazil

Original Authors: Meysam Sadeghi and Erik G. Larsson

In IEEE Communications Letters, vol. 23, No. 5, May 2019
doi: 10.1109/LCOMM.2019.2901469.

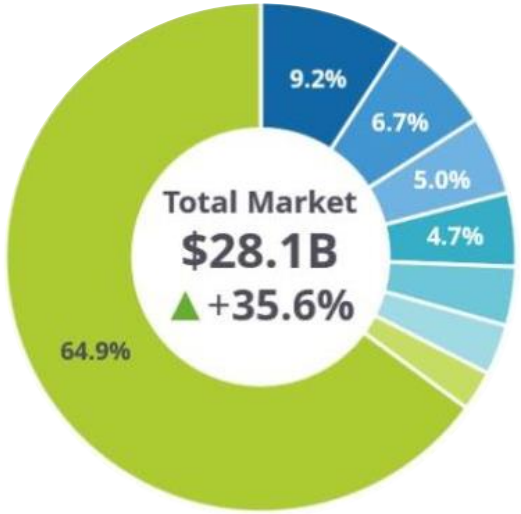
Presentation for TP 555 – Inteligência Artificial e Machine Learning

Santa Rita do Sapucaí, 23 de Junho de 2020

Outline

- Introduction
  - Artificial Intelligence
  - Problem definition
  - Proposal overview
- Deep Neural Networks (DNNs)
- End-to-end learning of communication systems using autoencoders
- The adversarial attack vulnerability
- Experimentation scenario
- Conclusions and future work

Introduction



28.1 Billion USD marketshare in 2018

- Nowadays it is hard to talk about technology without AI in the conversation
- Various business are trying to integrate it into their core
- The reason is because the end product results are more tailored to the user

Problem definition

- An increasing popular way of doing AI is through Neural Networks (NNs)
- They are popular because their applications are very flexible
- One of these applications is on Wireless Communications such as end-to-end learning communication systems using autoencoders
- Using autoencoders brings many benefits
- However, one of the main vulnerabilities in NN are adversarial attacks
- Since autoencoders use NN, they inherit this vulnerability

Proposal overview

- Physical adversarial attacks against end-to-end autoencoder communication systems
- White-box adversarial attacks
- Black-Box adversarial attacks
- Jamming attacks
- Analyze black-box and Jamming in traditional and autoencoder systems

Deep Neural Networks (DNNs)

- DNNs are basically Neural Networks with more layers
- DNNs are very flexible and can be used in various industries
- We will focus on End-to-end learning of communication systems using autoencoders

End-to-end learning of communication systems using autoencoders

- The goal is to learn full transmitter and receiver implementations which are optimized for a specific performance metric and channel model
- This can be achieved by representing transmitter and receiver as NNs and by interpreting the whole system as an autoencoder
- Autoencoder is an unsupervised artificial neural network that learns how to efficiently compress and encode data then learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible
- The advantage is that no math model of the cannel is required and therefore can be applied to any type of channel without prior analysis

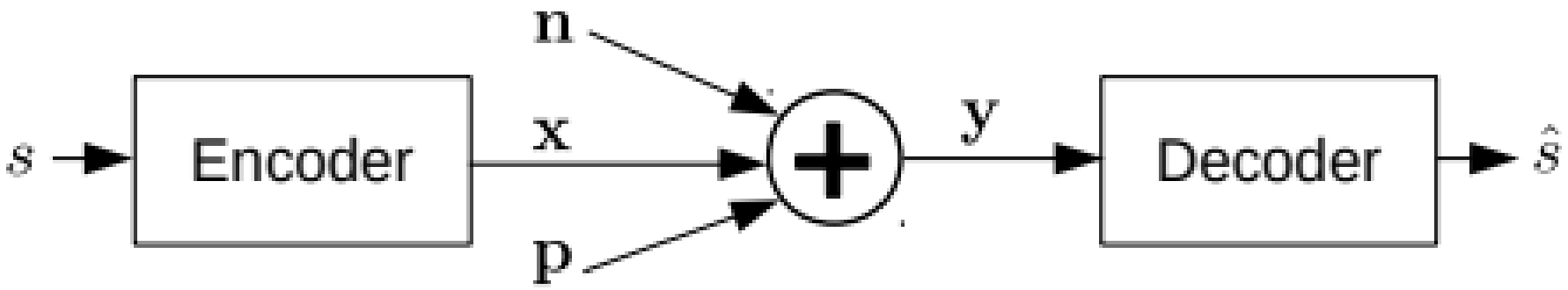
The adversarial attack vulnerability

- Inputs to ML models designed to cause the model to make a mistake
- Similar to optical illusions for machines
- Are one of the main vulnerabilities of Deep Neural Networks

The adversarial attack vulnerability

- DNNs are always vulnerable to these attacks
- Mitigating them is an important research topic
- Since the autoencoders are based on DNNs they are also vulnerable
- Adversarial attacks can be classified into digital and physical attacks
- The focus will be on Physical Attacks

Adversarial attack against autoencoder system



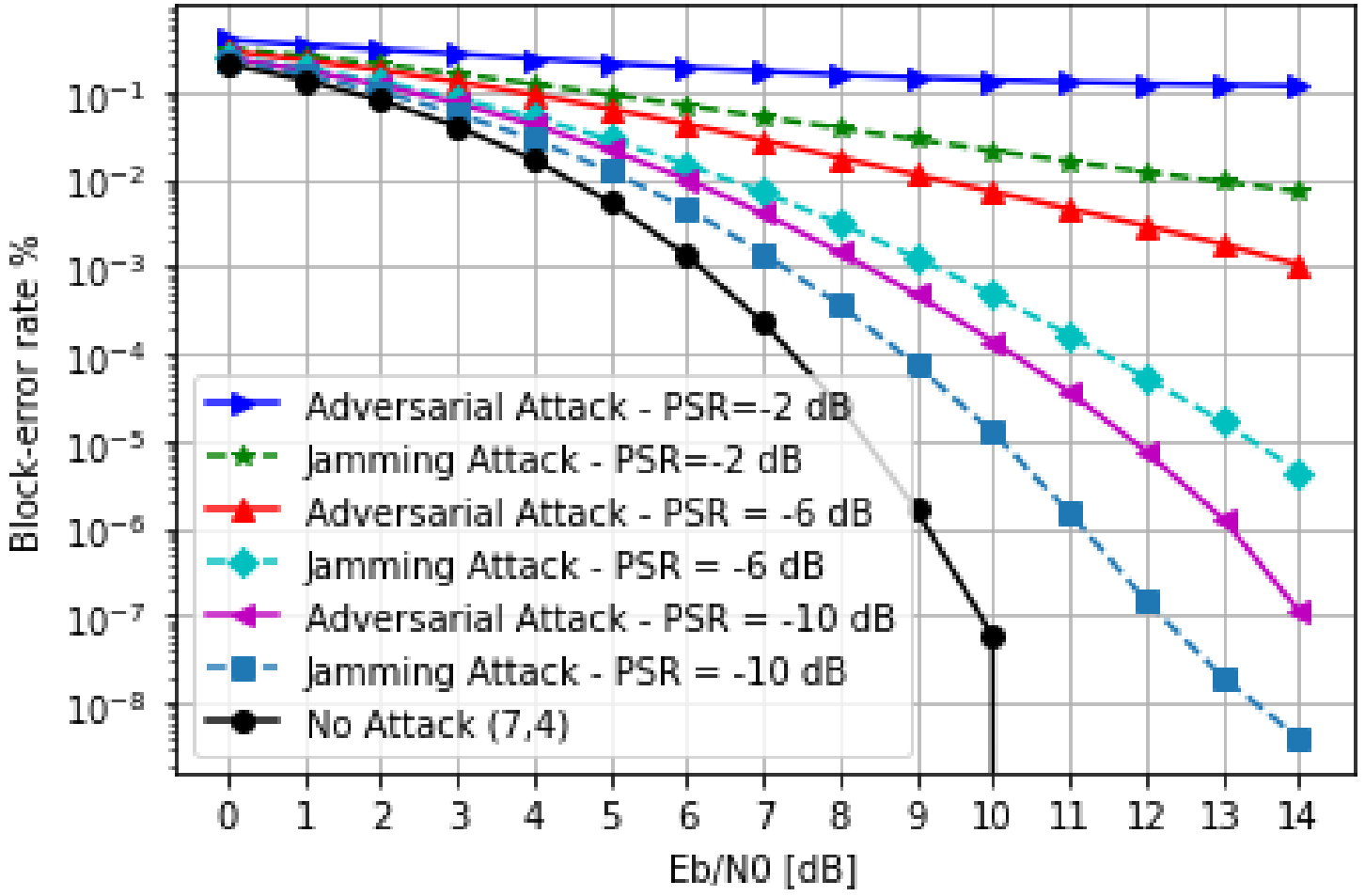
- s – Input signal
- x – Output of the encoder
- p – Atacker perturbation
- n – Noise

Experimentation scenario

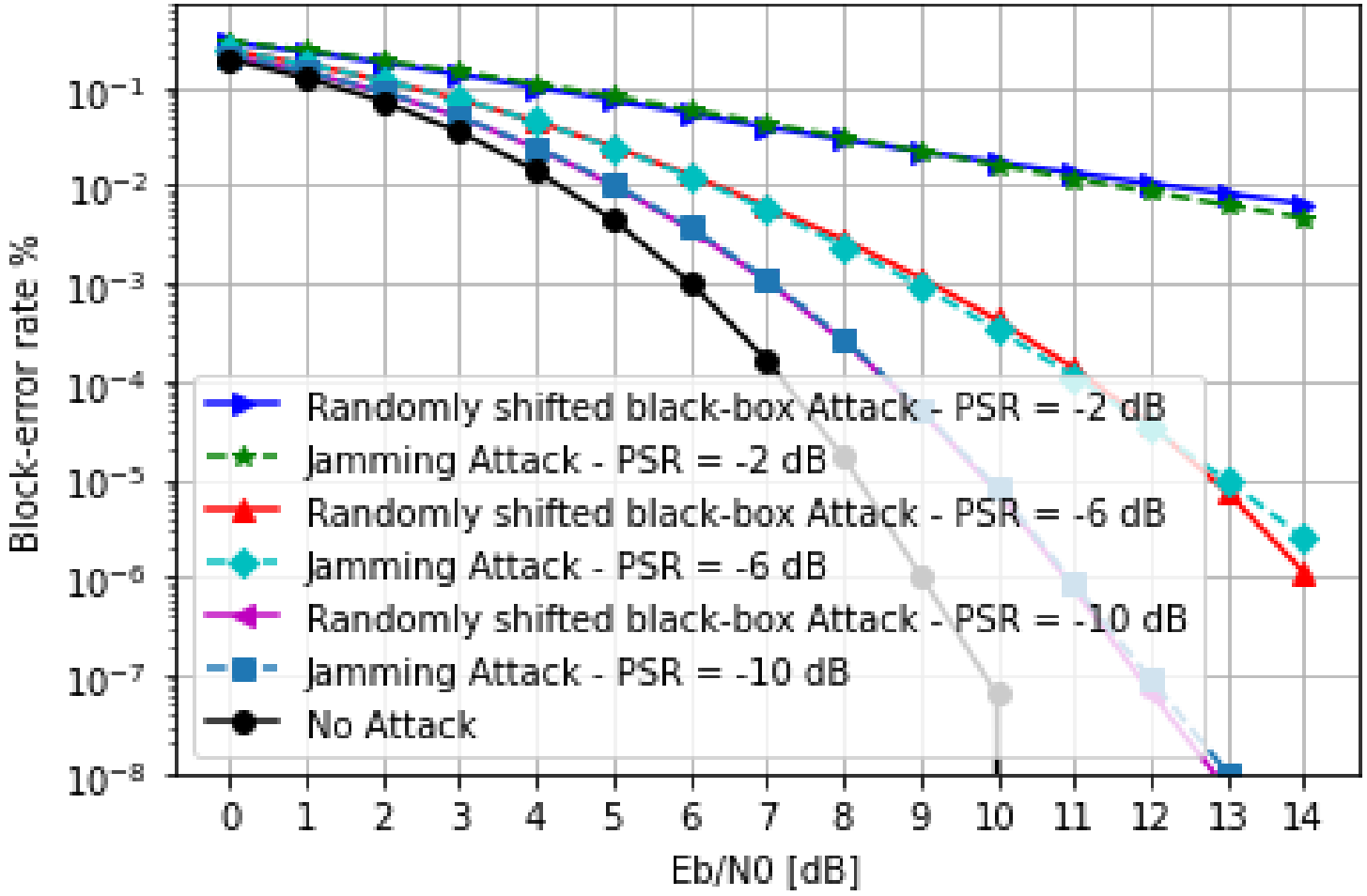
- The classic Hamming(7,4)
- AWGN (Additive White Gaussian Noise) Channel considered
- PSR (Perturbation-to-Signal Ratio): -2, -6, -10 dB

- Three scenarios
  - White-box attack (Full knowledge of the system)
  - Black-box attack (limited or no knowledge of the system)
  - Autoencoders VS Classical Approaches (PSR = -6dB)

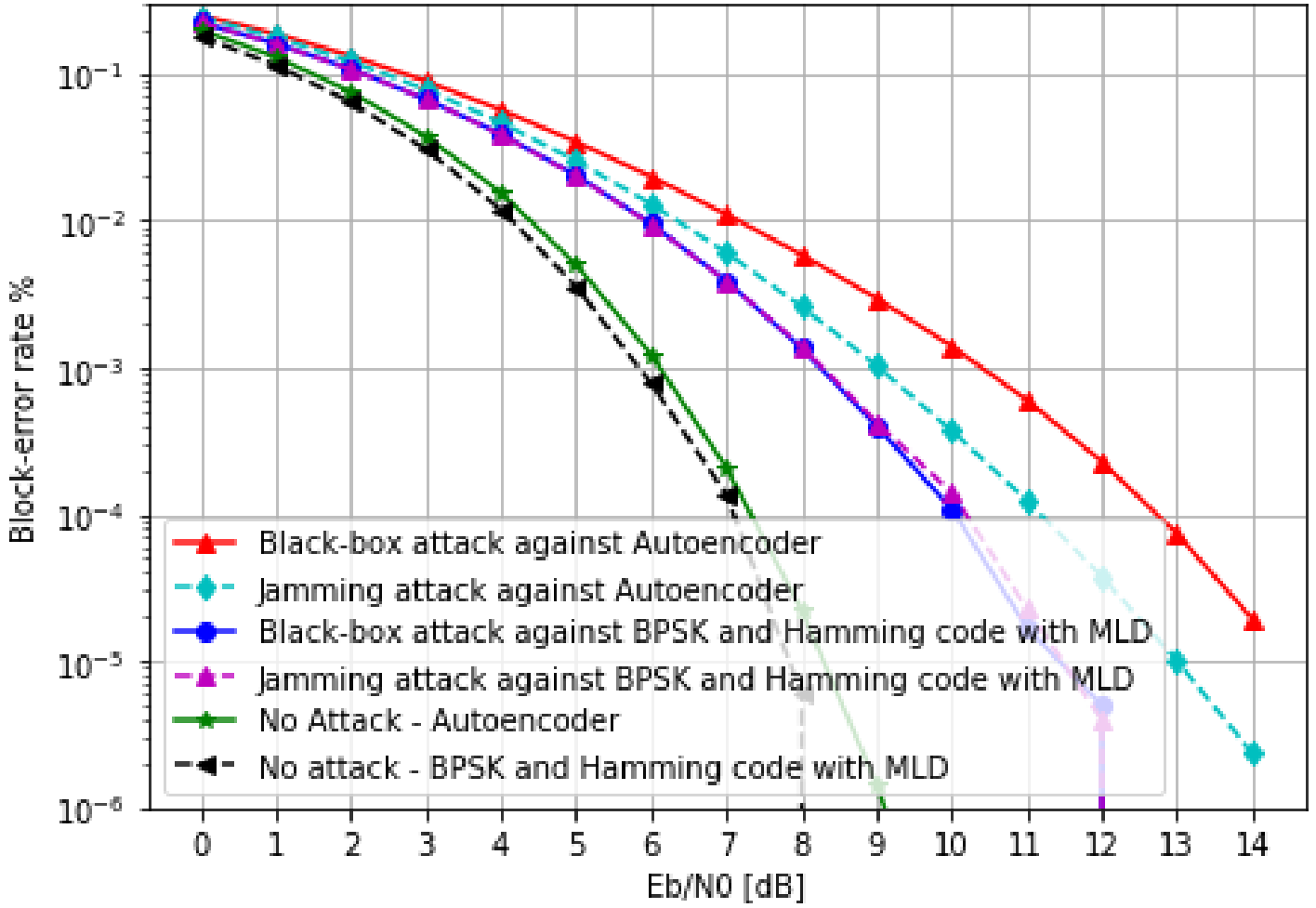
Scenario 1 – White-box



Scenario 2 – Black-box



Scenario 3 – Autoencoders vs Classical Approaches (PSR -6dB)



Conclusions

- Adversary transmitters can increase the BLER of a communication system by orders of magnitude by transmitting a well-designed perturbation signal
- The adversarial attacks are more destructive than the jamming attacks in Autoencoders
- Classical coding schemes are more robust than the autoencoders against both adversarial and jamming attacks

Future Work

- Mitigate the effects of Adversarial Attacks
- Investigate other channel models such as Rayleigh
- Use other hyperparameters other than the classic Hamming(7,4)
- Use more advanced Jamming strategies

Inatel
Instituto Nacional de Telecomunicações

IoT Research Group
Inatel - Brazil

Physical Adversarial Attacks Against End-to-End Autoencoder Communication Systems

Mauro A. A. Da Cruz[1] and Sheila C. S. Janota[1]
maurocruzter@gmail.com and sheilajanota@hotmail.com

[1] Instituto Nacional de Telecomunicações (INATEL), Santa Rita do Sapucaí-MG, Brazil

Original Authors: Meysam Sadeghi and Erik G. Larsson

In IEEE Communications Letters, Vol. 23, No. 5, May 2019
doi: 10.1109/LCOMM.2019.2901469.

Presentation for TP 555 – Inteligência Artificial e Machine Learning

Santa Rita do Sapucaí, 23 de Junho de 2020