INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, NAGPUR
भारतीय सूचना प्रौद्योगिकी संस्थान, नागपुर

LEARNING TO PROTECT COMMUNICATIONS WITH ADVERSARIAL NEURAL CRYPTOGRAPHY

Final Year Project
Mentor: Dr. Nidhi Lal
Shreyash H. Turkar
BT17CSE026

PROBLEM STATEMENT

We here discuss whether we can employ neural nets to secure our communication channels. We take classic Alice, Bob and Eve example where Alice is trying to send a message to Bob and Eve is eavesdropping. Instead of using rigid symmetric encryption algorithms like AES, DES, we will create an adversarial neural network where Alice will be trained to encrypt using a shared key, Bob will be trained to decrypt using a shared key and Eve will be trained to reconstruct the message without a shared key.

Introduction

Today, there are several types of neural networks which are used for various purposes. As we discover more ways of exploiting their potential, we ask whether they can also be used to make our communication secure and replace the existing rigid cryptographic algorithms. For sake of this paper, we will focus on symmetric encryption algorithms.

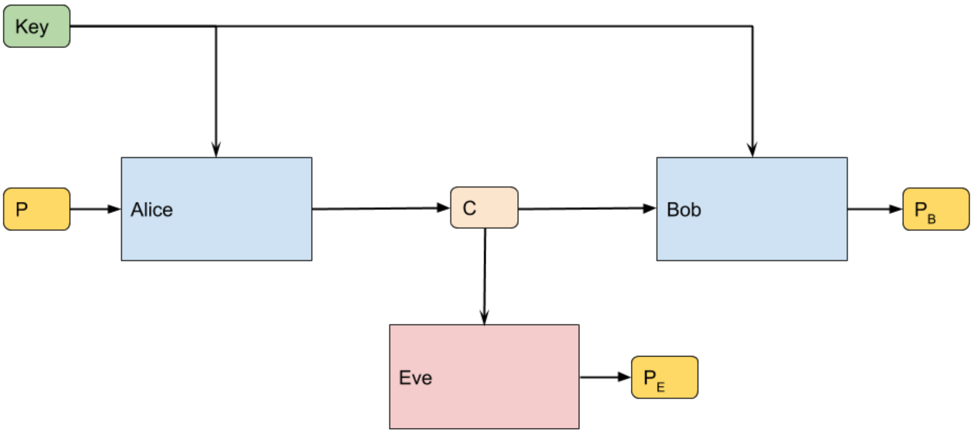


Fig. 1: Classic Alice, Bob and Eve Example

SYMMETRIC ENCRYPTION

Symmetric crypto-system can be easily explained by a classic example of three people named Alice, Bob and Eve. Both Alice and Bob hold shared a secret key. Alice wants to send a secret message to Bob which she encrypts using their shared secret key to create cipher-text, Alice sends this cipher-text to Bob through a channel, presumably insecure. Bob utilises the same shared secret key to decrypt. When cipher-text was intercepted by Eve but she can’t decrypt it without the shared secret key.

Classic Crypto-system

A classic scenario involving: Alice, Bob and Eve. Alice and Bob require to communicate securely, and Eve envies to snoop. We start with a particularly simple instance of this scenario, depicted in 1, in which Alice wishes to send a single confidential message P to Bob. The message P is input to Alice. Alice generates an output C. Bob and Eve take C to determine P. PBob and PEve, are what they compute respectively. Alice and Bob have a share a secret key K as a lead over Eve.

INTEGRATING NEURAL NETWORK INTO CRYPTO-SYSTEM

The Architecture of Model

We train three neural networks Alice, Bob and Eve whose jobs are as follows:

- Alice takes n-bit message + n-bit key and calculates n-bit cypher-text as an output
- Bob takes n-bit cypher-text created by Alice + n-bit key and computes original n-bit message as output.
- Eve takes n-bit cypher-text created by Alice and computes original n-bit message as output.

Rather than a rigid method, we chose the following architecture. It has a front fully-connected layer, where the number of outputs and inputs are equal. n-bits of plaintext and n-bits key are fed into this layer. (2n for Alice and Bob, n for Eve) This layer does not assure mixing between the key and the plaintext bits.

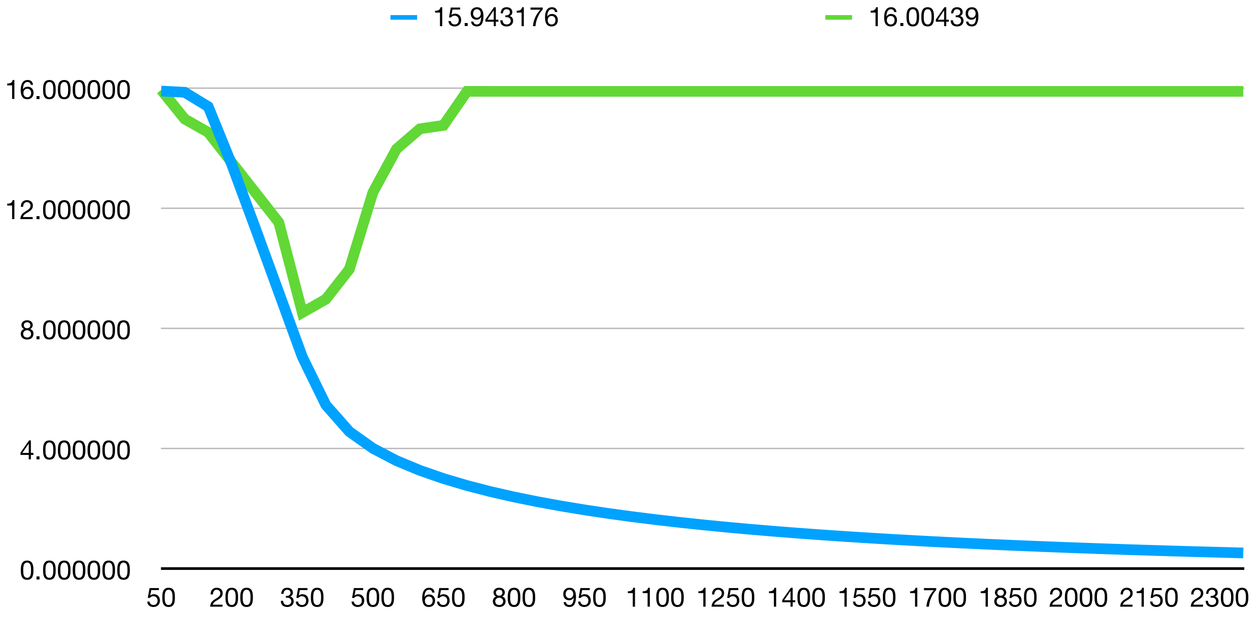
The 2n×2n fully-connected layer is followed by six of convolutional layers, the last layer produces an output of a size n. The convolutional layers take output by the previous layer and learn to apply some function to groups of the bits.

Input is processed by a dense layer using relu as an activation function. Following convolutional layers are described in terms of their window size, input depth, and output depth, stride—the amount by which the window is shifted and activation function as [4,1,1,1,leaky_relu], [2,1,2,2,leaky_relu], [1,2,1,1,leaky_relu], [1,1,4,1,sigmoid], [1,4,1,1,relu] and [1,1,1,1,tanh]. In final layer to bring the values back to range to map to binary we use tanh as activation function. Network of Alice and Bob are identical, whereas the network Eve network takes only the ciphertext as input explaining first N×2N FC layer. We train at learning rate of .0008.

Formulation of loss function

For Eve loss function is L1 distance between Eve's computation and the plaintext. The loss function for Alice and Bob has two components, first Bob's reconstruction error which is L1 distance between Bob's output and the input plaintext and Eve's success which is(N/2–Eve L1 error)^2/(N/2)^2. This takes care of Eve's random guessing. This is reduced when half of the message bits are wrong and half is right. A quadratic formula is prefered to emphasis on making Eve have a large error and to impose less of a penalty when Eve guesses a few bits correctly, as in occasional cases when Eve's guesses are random. This formulation allowed us to have a powerful loss function therefore improving the robustness of training.

Result



The figure shows Eve's, Bob's reconstruction error vs the number of training steps for 16-bits. The initial error rate of Eve and Bob are 16-bits and start reducing. Around 350 Eve's error rate is reduced to 8-bits and after 360 it starts increasing and around 800 it remains constant at 16-bits and the error rate of Bob is constant at 1 after 2150 and after 2300 is decreased to 0.

REFERENCES

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680, 2014a. http://papers.nips.cc/paper/5423-generative-adversarial-nets

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Info-GAN: Interpretable representation learning by information maximizing generative adversarial nets. CoRR, abs/1606.03657, 2016. URL https://arxiv.org/abs/1606.03657

Pengtao Xie, Misha Bilenko, Tom Finley, Ran Gilad-Bachrach, Kristin E. Lauter, and Michael Naehrig. Crypto-nets: Neural networks over encrypted data. CoRR, abs/1412.6181, 2014. URL http://arxiv.org/abs/1412.6181

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680, 2014a. http://papers.nips.cc/paper/5423-generative-adversarial-nets.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Info-GAN: Interpretable representation learning by information maximizing generative adversarial nets. CoRR, abs/1606.03657, 2016. URL https://arxiv.org/abs/1606.03657.

Pengtao Xie, Misha Bilenko, Tom Finley, Ran Gilad-Bachrach, Kristin E. Lauter, and Michael Naehrig. Crypto-nets: Neural networks over encrypted data. CoRR, abs/1412.6181, 2014. URL http://arxiv.org/abs/1412.6181.