

Learning cryptography using adversarial neural networks

メラウーシュ, イザーク

<https://hdl.handle.net/2324/7157357>

出版情報 : Kyushu University, 2023, 博士 (工学), 課程博士
バージョン :
権利関係 :



氏 名 : MERAUCHE ISHAK (メラウーシュ イザーク)

論 文 名 : Learning cryptography using adversarial neural networks
(敵対的ニューラル ネットワークを利用した暗号学習)

区 分 : 甲

論 文 内 容 の 要 旨

The field of Artificial Intelligence (AI) has seen an outstanding growth in the recent years in response to the increase of the needs of society in different areas, such as healthcare, finance, information processing, and cybersecurity. Still, on the subject of cybersecurity, one area that remained unexplored for decades is cryptography. In the early days of AI, neural networks struggled in learning the simplest forms of cryptography, such as the exclusive or operation, which discouraged researchers to further explore this area.

This situation changed with the seminal work by Abadi and Andersen, who took initial steps in 2016 to build neural network models that can learn secure forms of Cryptography using adversarial training. They propose a Generative Adversarial Networks model that can learn encryption. In their proposal, two neural networks train to protect their communication from a third neural network which is the attacker playing the role of an eavesdropper. The results were successful, and the two parties have been able to protect their communication from the attacker. The two parties are not trained to learn any specific encryption algorithm but rather train to achieve a specific goal. Their assigned goal is to generate ciphertexts that can be decrypted by Bob using a symmetric key but cannot be decrypted by the attacker without the key. The attacker is given the goal of decrypting the ciphertexts without the key. By continuously training against each other, the two parties end up coming with a way to protect their communication from this attacker.

This seminal work has also been used as a foundation for other cryptology solutions. For example, Hayes and Danezis show that it is possible to use the same model setup with a few changes to the neural network structure to make the neural networks learn Steganography. In their model, the sender learns to hide messages inside an image and the receiver learns to extract these messages properly. The attacker is a classifier that receives the original image and the steganographic image as input and tries to tell which one contains a hidden message.

Despite the positive feedback and follow up contributions that the Abadi and Andersen model has seen, it has been shown to have multiple flaws that need to be addressed, some of which were yet to be resolved in the state of the art. For example, Lu Zhou et al show that the

ciphertexts generated failed the National Institute of Standards and Technology (NIST) statistical test and are weak against multiple statistical attacks like the χ^2 attack. This means that the ciphertexts might contain sensitive information about the key and/or the plaintext.

Another important challenge was uncovered by Jamie Hayes and George Danezis, who show that even if two separate neural networks train on the same data, they are not guaranteed to learn the same encryption/decryption technique. This is known as the problem of non-convexity where deep learning models that train with the same data are not guaranteed to reach the same results. The reason is the big number of possible outcomes makes it almost impossible for the models to converge to the same outcome. As two machines that train on the same dataset of plaintexts and keys are not guaranteed to learn the same encryption algorithm, communicating among more than two parties becomes a difficult problem that needs to be addressed.

Finally, Abadi and Andersen themselves also point out that the neural networks cannot learn asymmetric encryption. In all their experiments, they are only able to protect their communication using symmetric keys. Giving a pair of public and private keys to the two parties makes the receiver not able to use the private key to decrypt the messages that have been encrypted with the public key.

This thesis explores the flaws and challenging issues stated above – weak ciphertexts, problem of non-convexity, inability to learn asymmetric encryption – and proposes solutions to address these issues, advancing the state of the art in this area.

To address the problem of non-convexity and weak ciphertexts, we propose a neural networks model that allows multiple parties to synchronize together and learn the same encryption/decryption technique. We also add more attackers that apply different cryptanalysis attacks which pushes the sender to generate stronger ciphertexts. The ciphertexts generated become resistant to the cryptanalysis attacks employed by the attackers and their encryption is equivalent to the One Time Pad encryption. Our model is a combination of different contributions that improve the security of the ciphertexts. We also show how this model can be used to obtain secret sharing schemes realizing any general access structure.

Jamie Hayes and George Danezis stress out that their model is also subject to the problem of non-convexity and that it is not guaranteed that two parties training on different machines can learn the same steganography algorithm. We address this issue by showing how more than two parties can learn the same steganography technique with the focus on a 3-party case.

Lastly, we address the issue of the model not being able to learn asymmetric encryption. We propose a model, the first of its kind, which learns to encrypt and decrypt data using asymmetric information. This removes the overhead of using a key exchange scheme to agree on a common state beforehand. Our model can encrypt data using a public key that can be known to anyone and decrypt data using a private key that is only known to the receiver.