XAE^2: Double Autoencoders for Data Security

Marco Ottavio Podesta Vezzali, Felíx, F. de J., Hérnandez G., A., Nieto, N., Morales, B., Gómez B., J. A., L. A. Munoz

Emails: marcoopodesta@gmail.com, ffelix3ro@gmail.com, alexhergomz@gmail.com, nezihng@gmail.com, a01661598@tec.mx, a01736171@tec.mx, amunoz@tec.mx

Introduction

Overview: XAE^2 introduces a double autoencoder architecture for text obfuscation and deobfuscation. It transforms plaintext into discrete latent representations and reconstructs it with up to **98% accuracy**. Key features include:

- Intermediate Teacher Forcing (ITF) mechanism.
- Robust training strategies ensuring inference accuracy.

Hypothesis

The double autoencoder architecture of XAE^2 can effectively perform text obfuscation and deobfuscation. It achieves this by:

- Transforming plaintext into discrete latent representations through the **Obfuscator**.
- Reconstructing the original text accurately via the **Deobfuscator**.

Definitions

Autoencoders: Autoencoders are neural networks designed for data compression and feature extraction. They consist of an **encoder** $E : X \to Z$, which maps input data X to a latent representation Z, and a **decoder** $D : Z \to \hat{X}$, which reconstructs an approximation of the original input [1].

$$X \xrightarrow{E} Z \xrightarrow{D} \hat{X}$$

Double Autoencoders: Double autoencoders extend traditional autoencoders with a sequence of two architectures. The first autoencoder (Obfuscator) generates an obfuscated representation Y, while the second autoencoder (Deobfuscator) reconstructs the original input X.

$$X \xrightarrow{E_1} Z_1 \xrightarrow{D_1} Y \xrightarrow{E_2} Z_2 \xrightarrow{D_2} \hat{X}$$

Obfuscator: Encodes plaintext into obfuscated text.
Deobfuscator: Recovers plaintext from obfuscated text.

Proposed Architecture

Overview of XAE^2: The architecture consists of two autoencoders:

- **Obfuscator:** Encodes plaintext into discrete latent representations.
- **Deobfuscator:** Reconstructs the original text from obfuscated representations.

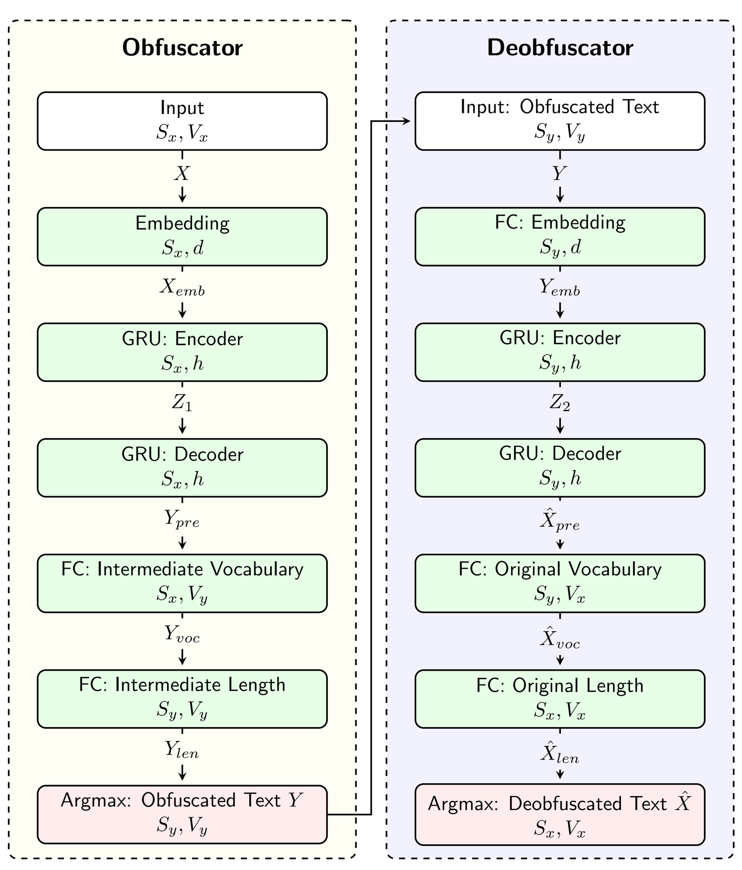

Figure 1: Proposed architecture.

Intermediate Teacher Forcing (ITF)

A training strategy that alternates between discrete and continuous intermediate representations to bridge the gap between training and inference, ensuring robustness under real-world conditions.

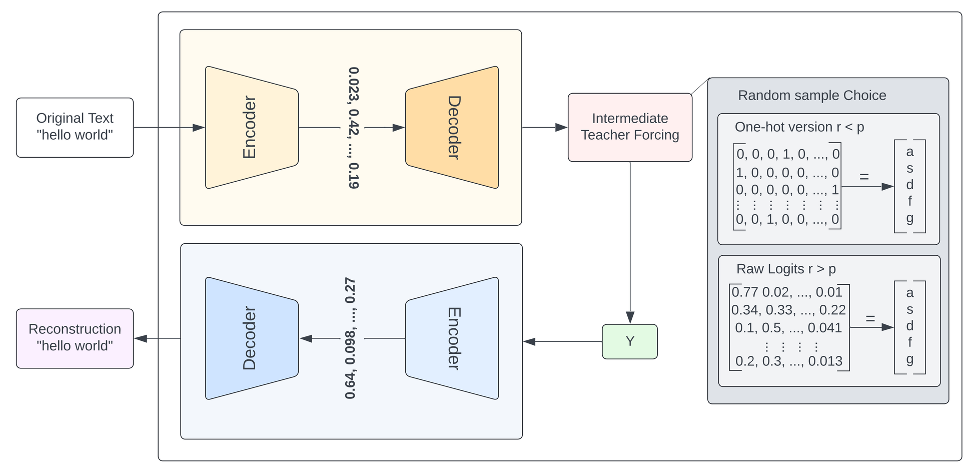

Figure 2: Intermediate Teacher Forcing Mechanism.

Training and Evaluation

Training Details:

- Dataset: 30,000 sequences of text for training.
- Optimizer: Adam.
- Batch Size: 32.
- Sequence Lengths: $S_X = 16$, $S_Y = 128$.
- Embedding Dimension: $d = 64$, Hidden Dimension: $h = 64$.

Evaluation Metrics: Accuracy was measured for both continuous and discrete inputs, demonstrating high performance with and without ITF during inference.

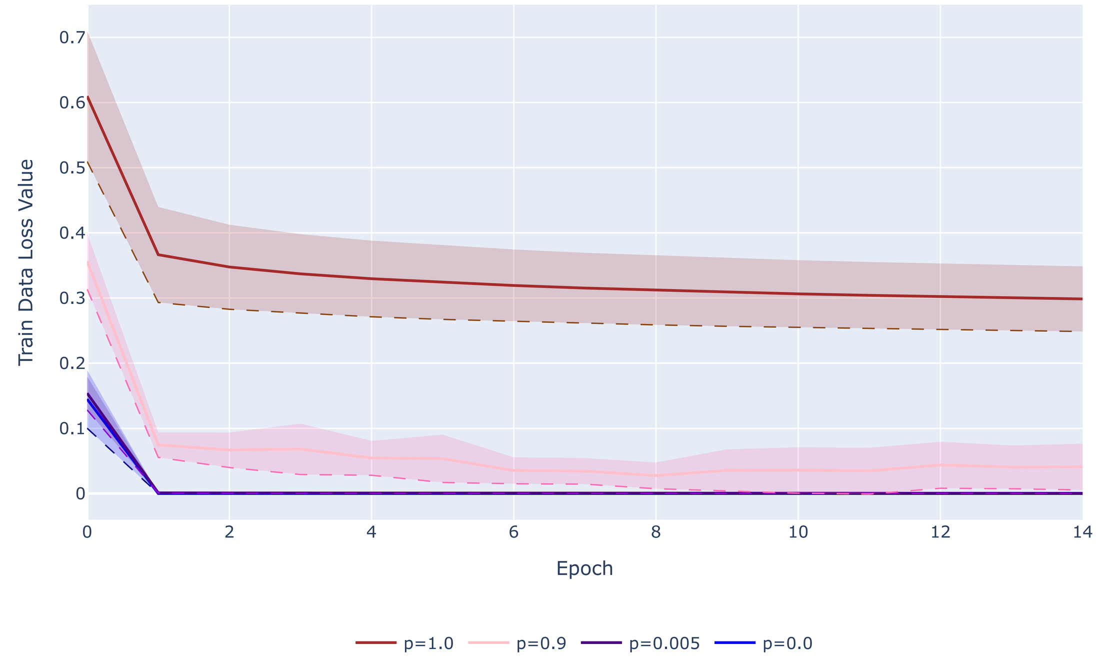

Figure 3: Training loss trends across epochs. $\mathbf{p = 0.0}$ and $p = 0.005$ are at the bottom line (overfitting around a loss of 0), followed by $p = 0.9$ and $p = 1.0$

Results

Key Observations:

- Achieved **98% accuracy** in deobfuscation with ITF-enabled training.
- Without ITF, accuracy dropped significantly when discrete inputs were used.
- Loss curves indicate strong learning convergence across configurations.

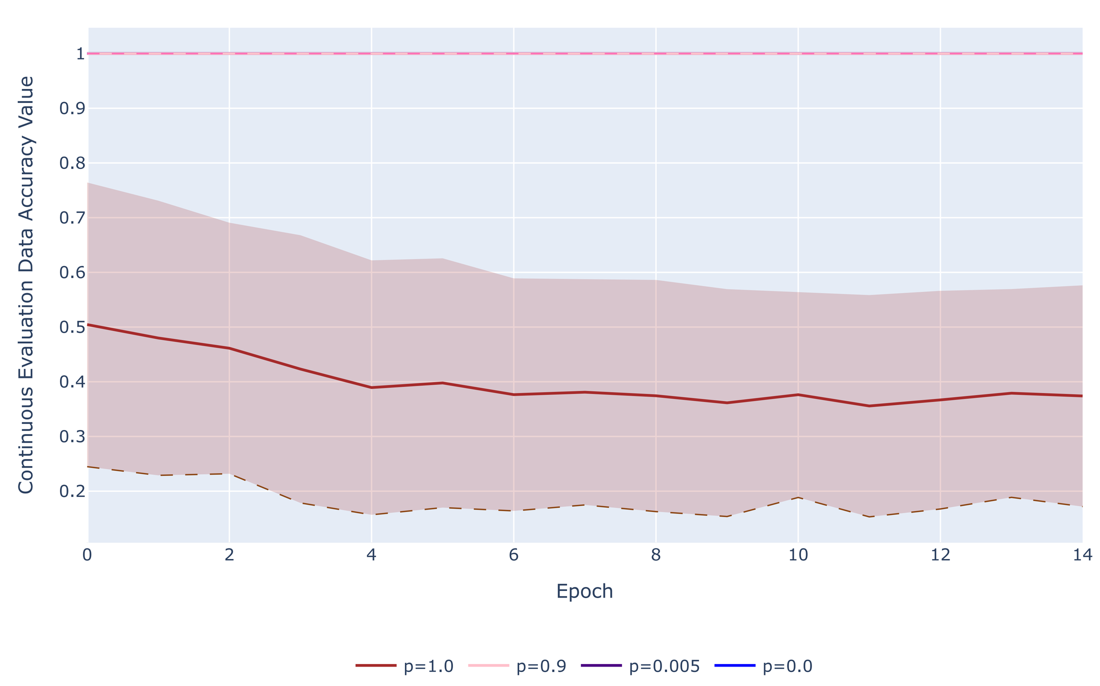

Figure 4: Accuracy trends with continuous inputs to the **Deobfuscator** during inference. $\mathbf{p = 1.0}$ at the bottom line (not being able to adapt to data it has not seen yet)

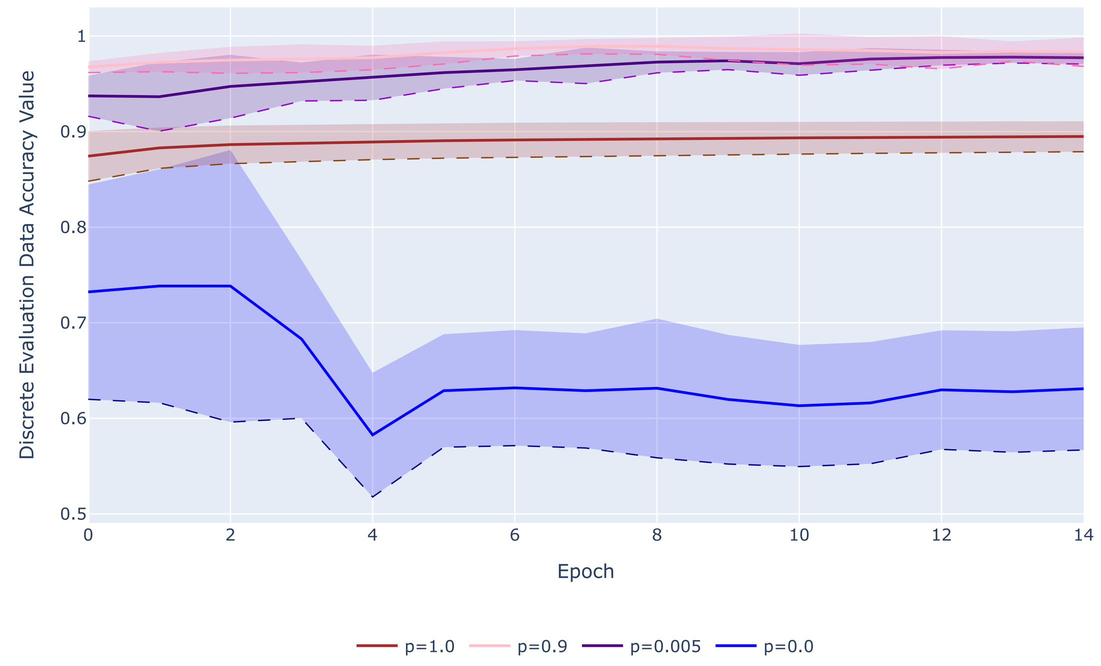

Figure 5: Accuracy trends with continuous inputs to the **Deobfuscator** during inference. $\mathbf{p = 0.0}$ at the bottom line (not being able to adapt to data it has not seen yet).

Applications and Extensions

Applications:

- Secure communication (chat encryption, email obfuscation).
- Privacy-preserving data transmission in IoT and healthcare.
- Neural cryptographic methods for real-time data security.

Future Directions:

- Extending XAE^2 to multimodal data obfuscation (image, audio, video).
- Incorporating user-specific private keys for customizable security.
- Advancing the Intermediate Teacher Forcing mechanism.

Conclusion

XAE^2 demonstrates high accuracy in text obfuscation and deobfuscation, leveraging a novel ITF mechanism. Its modular design and robustness make it suitable for various privacy-preserving applications, paving the way for innovative extensions and security protocols. Summary:

- XAE^2 achieves high accuracy in text obfuscation and deobfuscation.
- ITF bridges the training-inference gap effectively.
- Modular design enables broad applicability in secure and privacy-preserving domains.

References

[1] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *CoRR*, vol. abs/2003.05991, 2020. [Online]. Available: https://arxiv.org/abs/2003.05991

Downloads

PAPER POSTER