Generative Models for Information Security: Literature Review

Student: John Jennings **Supervisor:** Chris Willcocks

September 26, 2018

Problem Definition

Generative models are important tools in machine learning, seeing recent success in a wide variety of applications, from synthesizing realistic images of human faces and house interiors, to transforming photos from summer to winter and night to day. Within the information security sector, they have been used for learning novel encryption schemes, modelling password distributions, and steganography techniques for hiding information within an image. This project will expand upon current research, and investigate the applications of generative models to lexical steganography.

Definition of Terms

Steganography - The art of concealed communication, differing from cryptography in that the very existence of the message is secret. This is achieved by embedding the message into a seemingly innocuous **cover text** that can be sent to the intended recipient without arousing suspicion.

Stegotext - The result of steganographically embedding a message into a particular cover text.

Lexical Steganography - A form of steganography concerned with embedding a message within written text through the use of redundancies in natural language e.g. word ordering, inclusion of adjectives or choice of synonyms.

Generative Model - A model that learns the underlying distribution of the training data in order to generate new samples from the same distribution.

Identified Issues and Themes

Correspondence to Neural Machine Translation

While the concept of lexical steganography has existed for over a decade (Winstein 1999), there has been little research conducted in the area, with almost all existing solutions employing simple techniques based on synonym substitution (Alabish et al. 2013). A thorough search of the existing literature yielded only one paper that approached the problem from a machine learning perspective, Fang et al. (2017), who proposed a solution for the weaker problem of *cover generation* in which the user does not have control over the content of the cover text.

As a result, we will instead draw inspiration from the more mature field of Neural Machine Translation. Intuitively, lexical steganography can be thought of as a translation task, in that we are transforming the input text from one form to another while preserving the meaning. Instead of translating English text into the equivalent German text, we are translating English text into an equivalent English text that embeds the secret message.

Modelling the Problem

From this perspective, we can view lexical steganography as a problem of learning a generative model over the distribution of cover texts, such that given a particular cover text, we can generate new texts that are semantically equivalent. Embedding the hidden message can then be achieved through a learned algorithm that maps a particular codeword to a particular equivalent text. Decoding the hidden message from the stegotext is then simply a case of undoing this mapping.

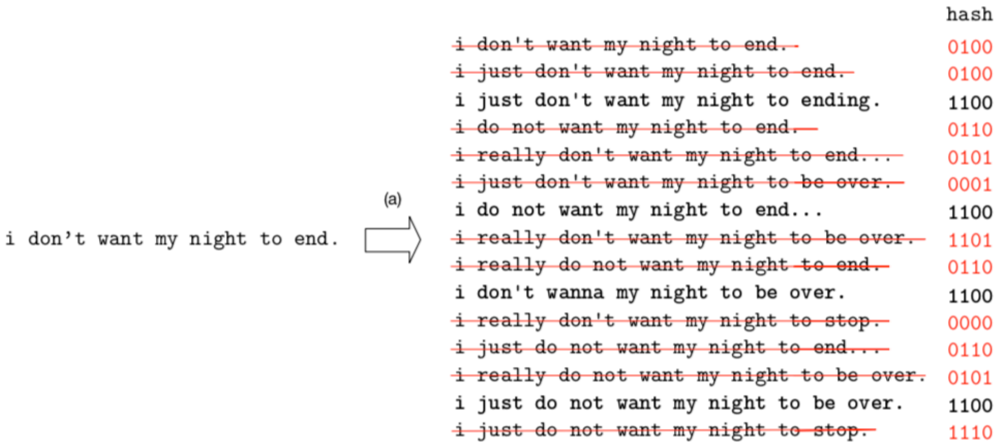
hash
i don't want my night to end.
(a)
~~i don't want my night to end.~~ 0100
~~i just don't want my night to end.~~ 0100
i just don't want my night to ending. 1100
~~i do not want my night to end.~~ 0110
~~i really don't want my night to end...~~ 0101
~~i just don't want my night to be over.~~ 0001
i do not want my night to end... 1100
~~i really don't want my night to be over.~~ 1101
~~i really do not want my night to end.~~ 0110
i don't wanna my night to be over. 1100
~~i really don't want my night to stop.~~ 0000
~~i just do not want my night to end...~~ 0110
~~i really do not want my night to be over.~~ 0101
i just do not want my night to be over. 1100
~~i just do not want my night to stop.~~ 1110

Figure 1: An example of mapping semantically equivalent texts to codewords (Wilson et al. 2014). In this case, the codeword is determined by hashing the generated text.

We will now outline some examples of cutting-edge generative models and show how they can be modified for the purpose of lexical steganography.

Autoencoder

In its simplest form, an autoencoder is a feed-forward neural network consisting of an input layer, a single hidden layer, and an output layer of the same size as the input. The model is then trained to output a reconstruction of the original input, thereby learning a latent representation of the input within the hidden layer. Modified versions of this relatively simple model have been successfully used for a number of applications, such as variable-rate image compression (Toderici et al. 2017) and pretraining deep neural networks to have meaningful starting weights (Poultney et al. 2007). Compared to more traditional methods of unsupervised representation learning, such as principal component analysis, the autoencoder is shown to have a better expressive power, due to its ability to learn non-linear encodings (Goodfellow et al. 2016).

One type of autoencoder that is of particular use for lexical steganography is the variational autoencoder (VAE), which learns to encode a training sample as a probability distribution over the latent space, instead of a single point. This is effective in generalizing the autoencoder to act as a generative model over the training distribution. In the case of lexical steganography, a VAE could be trained on a set of cover texts such that at inference time, sampling from the distribution of a particular cover text will yield equivalent texts that can then each be mapped to a codeword.

seq2seq seqgan

char/word/tweet level

Dataset

For training

Additional tuning

word embeddings skip connections dropout

Loss Function

nll vs bleu

Proposed Direction of the Project

References

Alabish, A., Goweder, A. & Enakoa, A. (2013), ‘A universal lexical steganography technique’, *International Journal of Computer and Communication Engineering* **2**(2), 153.

Fang, T., Jaggi, M. & Argyraki, K. (2017), ‘Generating steganographic text with lstms’, *arXiv preprint arXiv:1705.10742* .

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016), *Deep learning*, Vol. 1, MIT press Cambridge.

Poultney, C., Chopra, S., Cun, Y. L. et al. (2007), Efficient learning of sparse representations with an energy-based model, *in* ‘Advances in neural information processing systems’, pp. 1137–1144.

Toderici, G., Vincent, D., Johnston, N., Hwang, S. J., Minnen, D., Shor, J. & Covell, M. (2017), Full resolution image compression with recurrent neural networks., *in* ‘CVPR’, pp. 5435–5443.

Wilson, A., Blunsom, P. & Ker, A. D. (2014), Linguistic steganography on twitter: hierarchical language modeling with manual interaction, *in* ‘Media Watermarking, Security, and Forensics 2014’, Vol. 9028, International Society for Optics and Photonics, p. 902803.

Winstein, K. (1999), Lexical steganography through adaptive modulation of the word choice hash, *in* ‘Secondary education at the Illinois Mathematics and Science Academy’.