Generative Models for Information Security: Literature Review

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

September 19, 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. By abstracting the problem in this way, we can take the cutting-edge research that powers Google Translate and apply it to our problem in a way that has not been previously attempted.

Choosing a Dataset

we basically just need a lot of english text, no need for labels etc lots of choice everybody uses twitter for good reason pretty standard form compared to prose character limit ¡- good since RNNs tend to struggle on longer sequences lots of data available can compare against existing solutions misspellings + slang -¿ more synonyms has an immediate use case

Choice of Model

gan vs autoencoder charnn vs seq2seq vs paragraph vectors

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* .

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