GENERATIVE MODELS FOR INFORMATION SECURITY

Degree: MEng. Computer Science

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

Description:

Generative models, particularly Generative Adversarial Networks (GANs), 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, GANs 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 machine learning to lexical steganography - the practice of hiding a hidden message within a 'cover text'.

Preliminary Preparation:

- Collection of an appropriate dataset of cover texts e.g. BBC News articles, tweets, Wikipedia articles
- Survey of existing methods for lexical steganography with focus on machine learning applications, as well as methods for attacking these stegosystems.

Minimum Objectives:

- Simple autoencoder-inspired network takes cover text and message as input. Encoder aims to embed message within cover text, decoder aims to extract the embedded message. Loss function is a combination of reconstruction error of decoder with respect to original message and reconstruction error of encoder with respect to the cover text, where 'reconstruction error' takes into account semantic coherence through the use of pre-trained word embeddings.
- Comparison of this network against existing methods e.g. TLex, LUNABEL.
- Users will be able to interact with the system through a web interface, embedding their own message into a cover text of their choice.

Intermediate Objectives:

- Encoder output will be robust to random deletion of words/sentences i.e. hidden message can still be decoded from partial output.
- Encoder will be robust to an adversarial network attempting to identify if the output text contains a hidden message.
- Encoder will additionally take a random key as input. Will be robust to an adversarial network attempting to decode the hidden message without knowledge of the key.

Advanced Objectives:

- Cryptanalysis of system to give some theoretical bounds on the level of security that is provided e.g. Does partial knowledge of the key yield a partially decoded message? How much information can be successfully hidden within a typical sentence?

References

Chand, V. & Orgun, C. O. (2006), Exploiting linguistic features in lexical steganography: design and proof-of-concept implementation, *in* 'System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on', Vol. 6, IEEE, pp. 126b–126b.

Chang, C.-Y. & Clark, S. (2012), 'The secrets in the word order: Text-to-text generation for linguistic steganography', *Proceedings of COLING 2012* pp. 511–528.

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

Taskiran, C. M., Topkara, U., Topkara, M. & Delp, E. J. (2006), Attacks on lexical natural language steganography systems, *in* 'Security, Steganography, and Watermarking of Multimedia Contents VIII', Vol. 6072, International Society for Optics and Photonics, p. 607209.

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