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:

- Implement a sequence-to-sequence model in which a slight change to the latent representation of an input sentence will result in a semantically similar sentence as output.

Intermediate Objectives:

- Tailor the encoder-decoder architecture to improve the quality of the decoded output, such as with an adversarial component, residual-connections, regularisation, and tuning of the number of layers/parameters
- Investigate a method of modifying the latent representation to store a hidden message.
- Comparison of this network against existing methods, e.g. TLex, LUNA...
- Investigate/collect results on robustness to random deletion, adversarial attacks etc.

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?
- Deploy the model and create a front-end, making it easy to use in inference (hide/reveal messages on new texts).

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'.