Generating Steganographic Text with LSTMs

Tina Fang [a] · Martin Jaggi [b] · Katerina Argyraki [b]

[a] UNIVERSITY OF WATERLOO

[b] ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

overview

Challenge: hide secret information in natural-looking tweets and emails.

Contributions

✅ hides over 11 times more than existing steganographic systems[1]

✅ provides flexible trade-off between capacity and text quality

motivation

Problem

➡ private user data from free communication systems (e.g. emails, Twitter) are used for advertising or government investigations

➡ using encryption in messages interrupts free service

⇒ users cannot communicate privately

Solutions

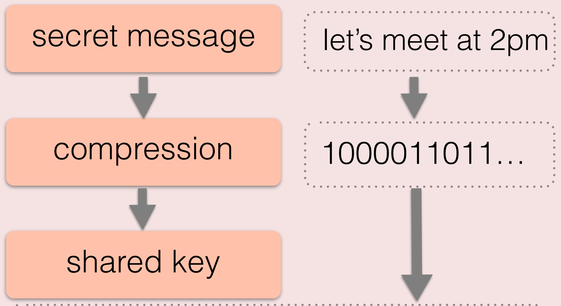
➡ existing steganographic techniques use hardcoded synonym and paraphrase substitution

➡ non-automated techniques use word lists to generate text manually or using n-grams

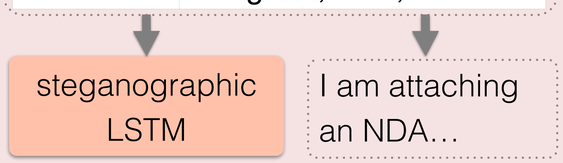
our solution: generate text with a neural network (LSTM)

steganographic system

Encoder



bit block	words
00	This, am, !, weather
01	was, attaching, today
10	I, better, an, Great
10	great, than, NDA



Decoder

recovers original text by mapping each word to its bit block

our modification

Regular LSTM:

- Given a word sequence $(x_1, x_2, \ldots, x_t)$, a regular LSTM selects the word with the highest probability $P[x_{t+1} \mid x_{<=t}]$ as its next word

Steganographic LSTM:

- restrict generation of the word to the bit block B given by the secret data
- $P[x = W_j] = 0$ for $j \notin W_B$

experiments

steganographic tweets

- I can't take the whole day to the car, the people I suddenly didnt understand.
- "where else were u making?... i feel fine? - e? lol" * does a voice for me & take it to walmart?
- Don't think I can trust anymore... but I know is my fault.
- I can't believe he is a freshman!

steganographic emails

- If you do like to comment on the above you will not contact me at the above time by 8:00 a.m. on Monday, March 13 and July 16 and Tuesday, May 13 and Tuesday, March 9 - Thursday, June 17, - 9:00 to 11:30 AM.
- At a moment when my group was working for a few weeks, we were able to get more flexibility through in order that we would not be willing.

discussion

Text Quality

- adding common tokens (e.g. ',', 'I', 'to', '…') improves grammar and context coherence (the 3rd and 4th tweet were generated with common tokens added)
- text quality increases when the number of bit blocks decrease (when less information is hidden)

Tweets vs Emails

- emails have a much longer range context dependency
- text quality of emails improves if the non-steganographic LSTM improves

future work

- evaluate against human judges
- evaluate against steganography classifiers
- generate tweets personalized to a user type or interest group
- open-source this system

reference

[1] Alex Wilson, Phil Blunsom, and Andrew D Ker. 2014. Linguistic steganography on twitter: hierarchical language modeling with manual interaction. In IS&T/SPIE Electronic Imaging . International Society for Optics and Photonics, pages 902803–902803.

Github: tbfang/steganography-lstm