Towards Near-imperceptible Steganographic Text

Falcon Z. Dai [1] dai@ttic.edu
Zheng Cai [2] jon.z.cai@colorado.edu

[1]Toyota Technological Institute at Chicago
[2]University of Colorado at Boulder

Motivations

Computationally resourceful entities could monitor *ostensibly* private communications at scale. The mere presence of encryption may raise suspicion in the eavesdropper.

- Can we hide secrets in natural text? **Yes**, via linguistic steganography. A language model (LM) lets us sample fluent text.
- How hard is it to tell steganographic text from fluent text? Quantified by total variation distance (TVD), the existing methods rely on **unrealistic** assumptions.
- Can we do better? **Yes**, we propose a method with a stronger guarantee.

Highlights

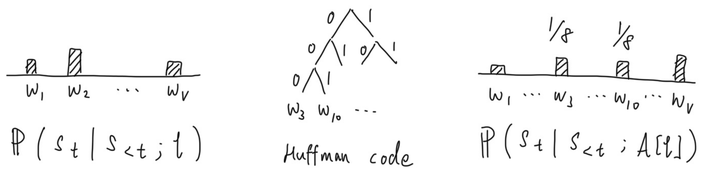
- We quantify statistical imperceptibility with total variation distance (TVD) between language models. We study the TVD of several encoding algorithms [FJA17, YGC+18] and point out the implicit assumption for them to be near-imperceptible.
- We use a state-of-the-art transformer-based, subword-level LM, `GPT-2-117M` [RWC+19], to empirically evaluate the plausibility of these assumptions.
- We propose an encoding algorithm `patient-Huffman` with strong relative statistical imperceptibility.

Intuition

Consider plausible continuations of the following two prefixes.

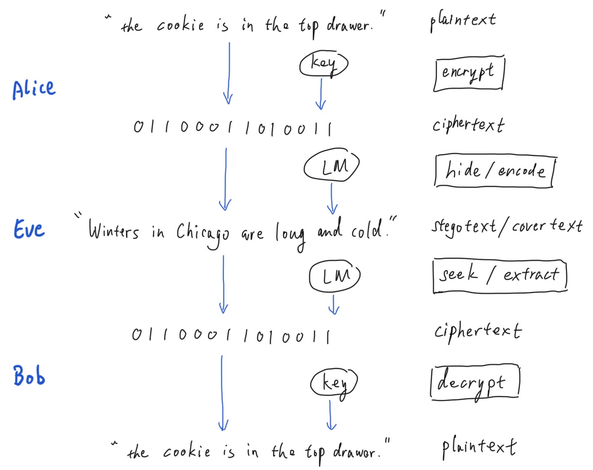
- "I like your" ⟶ {"work", "style", "idea", "game", "book", ... }
- "It is on top" ⟶ {"of", ",", "and", ".", ... }

Non-standard sampling at the latter can expose the stegosystem.



Communication protocol

Alice wants to send a secret message to Bob via a channel monitored by Eve who expects to see fluent text.



Algorithm

Be patient and skip encoding steps that can expose the stegosystem. Choose $\delta_t \in o(1/t)$ for each step, so the TVD is bounded.

Algorithm 1 `patient-Huffman` (one encoding step)

Input: a language model ℓ, prefix $h \in \Sigma^*$, an imperceptibility threshold δ, a ciphertext b.
Output: a stegotext from Σ^*.
Compute the distribution of the next token $p \leftarrow \mathbb{P}[\cdot|h; \ell]$.
Construct a Huffman tree c for p.
Compute the TVD (or the KL divergence) between p and the Huffman measure m_c corresponding to c.
if TVD (or KL divergence) $< \delta$ **then**
    Decode a token w by consuming the ciphertext b and following its bits starting at the root of Huffman tree c.
else
    Sample a token w according to p.
end if
Append the token to prefix $h \leftarrow h; w$
return h

Formalism

Total variation distance (TVD)

$d(p, q) \coloneqq \sup_{E \in \mathcal{F}} |p(E) - q(E)| = \frac{1}{2} \sum_{x \in X} |p(x) - q(x)|$

It takes at least $\Omega(1/d(p,q)^2)$ samples to distinguish two distributions p and q.

Decomposition of TVD

Suppose the true LM of the monitored channel is ℓ^*, and we have access to some base LM ℓ, then running encoding algorithm $\mathfrak{A}_\ell$ induces an effective LM $\mathfrak{A}[\ell] \coloneqq \mathbb{E}_b[\mathfrak{A}_\ell(b)]$. The TVD between the effective LM and the true LM

$$d(\ell^*, \mathfrak{A}[\ell]) \leq d(\ell^*, \ell) + d(\ell, \mathfrak{A}[\ell]).$$

By Pinsker's inequality, a bound via the KL divergence (in bits) on each step

$$d(\ell, \mathfrak{A}[\ell]) \leq \sqrt{\frac{\ln 2}{2} \sum_{t=1}^{\infty} D_{KL}\left(\mathbb{P}[\cdot|s_{<t}; \ell] \,||\, \mathbb{P}\left[\,\cdot\,|s_{<t}; \mathfrak{A}[\ell]\right]\right)}.$$

Open problems

- Can the eavesdropping adversary achieve $O(1/d^2)$? That is, is there a detection algorithm matching the lower bound? This seems to require some extra assumptions on fluent text.
- The entropy of fluent text is not uniform over steps and it reflects a kind of structure.

References

[Cac04] Christian Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, 2004.

[FJA17] Tina Fang, Martin Jaggi, and Katerina Argyraki. Generating steganographic text with lstms. In *Proceedings of ACL 2017, Student Research Workshop*, pages 100–106, 2017.

[Huf52] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[HvAL08] Nicholas Hopper, Luis von Ahn, and John Langford. Provably secure steganography. *IEEE Transactions on Computers*, 58(5):662–676, 2008.

[MHC+08] Peng Meng, Liusheng Huang, Zhili Chen, Wei Yang, and Dong Li. Linguistic steganography detection based on perplexity. In *2008 International Conference on MultiMedia and Information Technology*, pages 217–220. IEEE, 2008.

[RWC+19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. (Accessed on 2019-4-23).

[YGC+18] Zhongliang Yang, Xiaoqing Guo, Ziming Chen, Yongfeng Huang, and Yu-Jin Zhang. Rnn-stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security*, 2018.