CERIAS

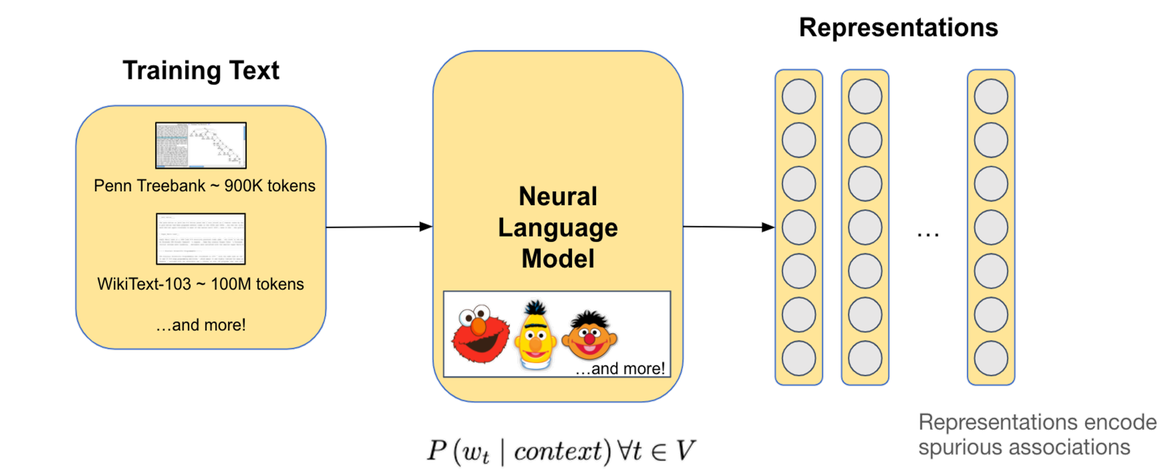
The Center for Education and Research in Information Assurance and Security

Interpretable Privacy Preservation of Text Representations Using Vector Steganography

Geetanjali Bihani, Julia Taylor Rayz
{gbihani, jtaylor1}@purdue.edu

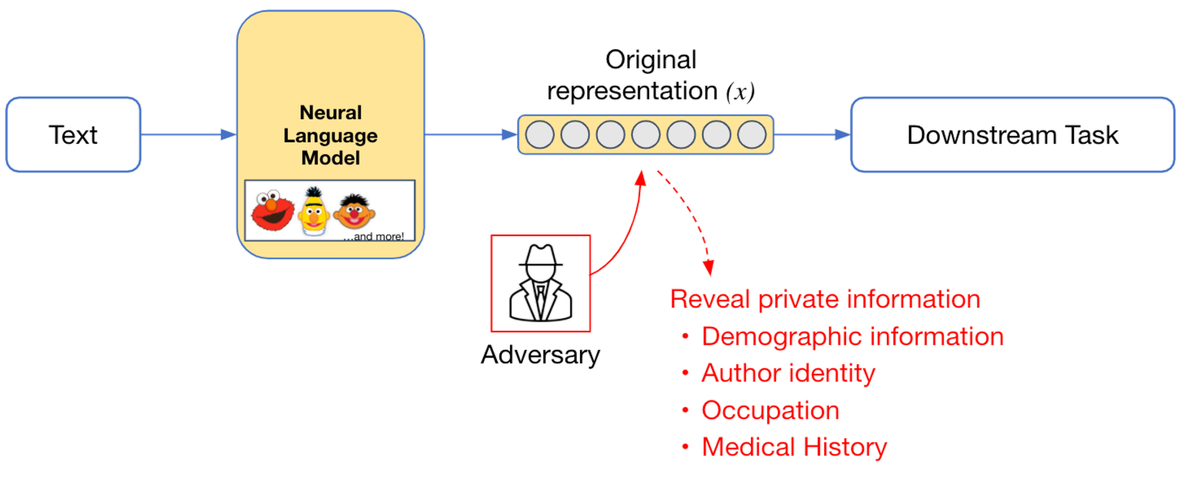
Privacy Risks of Neural Language Models

- Deep neural language model-based word representations encode spurious associations learnt from natural language text
- Adversaries can reverse-engineer private information and exploit neural LMs



Neural language models trained on large text corpora learn probability distribution over word sequences

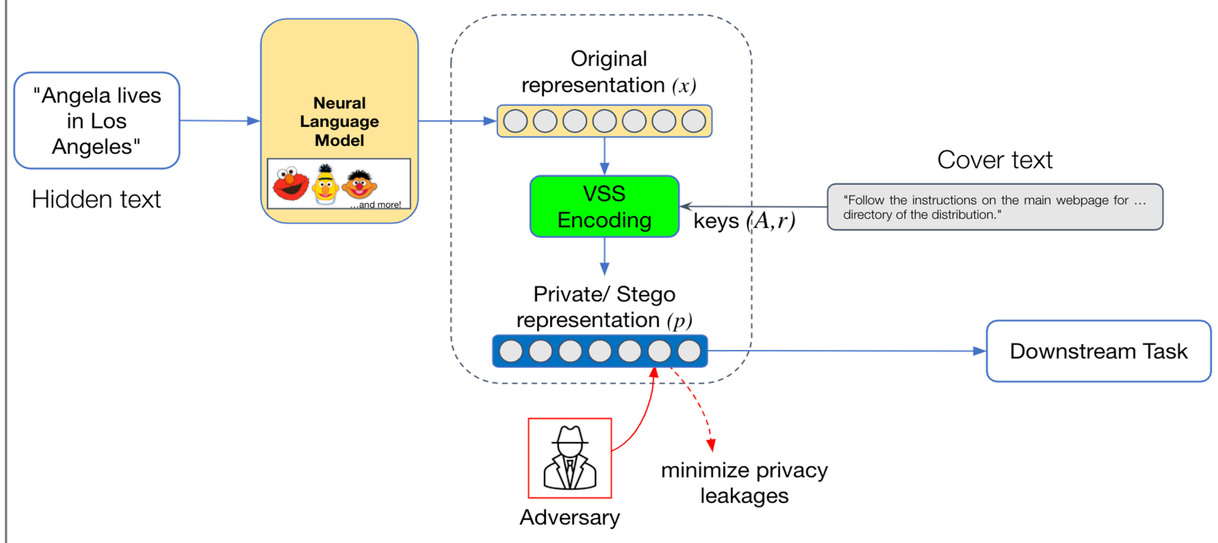
- Current privacy preservation methods cause privacy leakages
- Modify training objectives to achieve privacy



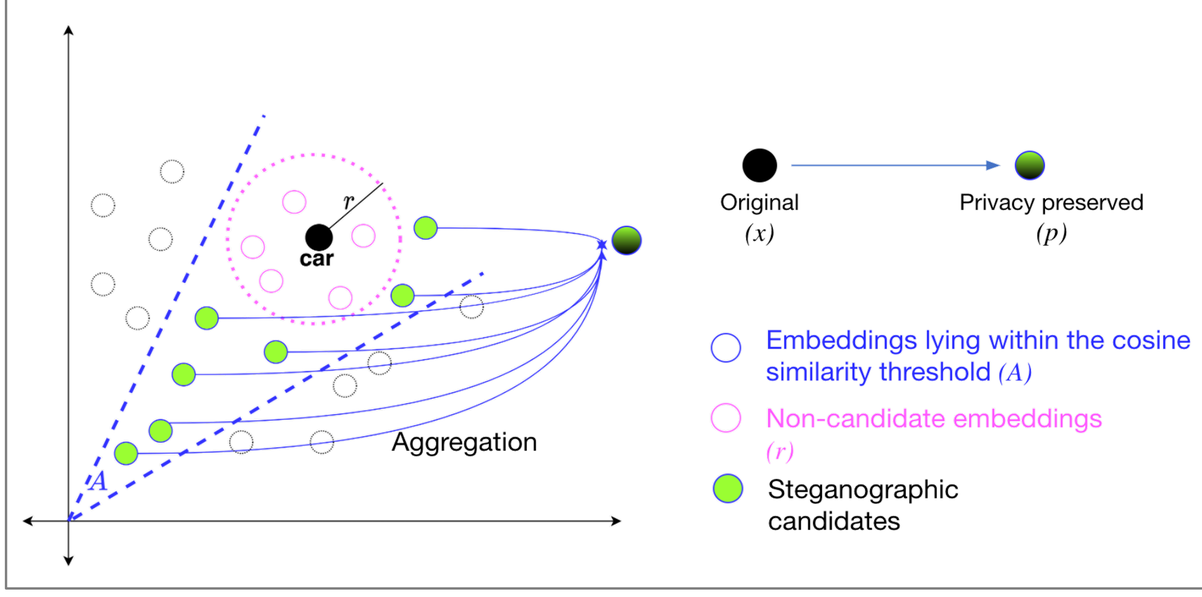
Vector Space Steganography (VSS)

Key Idea: Create privacy-preserving representations such that the algorithm

- can be applied as a post-processing step
- is interpretable
- retains utility



- Obfuscate original representation using steganographic candidates generated from the vector space geometry
- Candidates generated by reversing the steganalysis criteria described for multidimensional vectors
- Aggregate candidates to generate modified representation



Proposed Research Objectives

How effectively can steganographic models of privacy preservation obfuscate spurious associations (implicit and explicit) without compromising data utility?

- Design VSS applicable on-top-of LM representations to hide spurious associations
- Evaluate the representation utility post-VSS across general purpose LMs
- Evaluate the provable privacy guarantees of VSS

Types of Attacks

Type	Knowledge
Black-Box	x'
Grey-Box	$x' + g$ $g \subset [c, LM, A, r]$
White-Box	$x' + c + LM + A + r$

x : Original Representation
x' : Private Representation
c : Cover Text
A : Cosine Similarity Threshold
r : Distance Threshold
LM : Language Model

Takeaways

- Spurious associations learnt by neural language models can lead to privacy risks
- Propose vector space steganography, using the vector representation geometry to hide original representations
- Protects training data privacy and maintains data utility at minimal cost
- Quick implementation on-top of any state-of-the-art language models

Selected References

- Carlini, N.; Deng, S.; Garg, S.; Jha, S.; Mahloujifar, S.; Mah- moody, M.; Thakurta, A.; and Trame`r, F. 2021. Is Private Learning Possible with Instance Encoding? In *2021 IEEE Symposium on Security and Privacy (SP)*, 410–427. IEEE.
- Coavoux, M.; Narayan, S.; and Cohen, S. B. 2018. Privacy- preserving Neural Representations of Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1–10.
- Huang, Y.; Song, Z.; Chen, D.; Li, K.; and Arora, S. 2020. TextHide: Tackling Data Privacy in Language Understanding Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1368–1382. Online: Association for Computational Linguistics.
- Jiang, N. 2008. A Novel Analysis Method of Information Hiding. In *2008 Congress on Image and Signal Processing*, volume 5, 621–625.

PURDUE UNIVERSITY Discovery Park

CERIAS