"What did you say?": Extracting unintentional secrets from predictive text learning systems

Paper by

Gwyn Wilkinson
Dr Phil Legg

3 June 2020

UWE Bristol | University of the West of England

Introduction – Who we are

Gwyn Wilkinson
1st year PhD Student
Computer Science &
Creative Technologies,
Faculty of Environment and Technology
UWE Bristol

Gwyn2.wilkinson@live.uwe.ac.uk

Dr Phil Legg
Associate Professor – Cyber Security
Computer Science &
Creative Technologies,
Faculty of Environment and Technology
UWE Bristol

Phil.legg@uwe.ac.uk
Plegg.me.uk

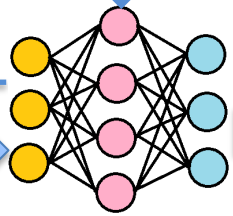
UWE Bristol | University of the West of England

Introduction – Main Idea

- Can a user-trained predictive text model **memorise** a secret…

  – and can we **extract** it?

"My password is Reindeerflotilla"



"Reaction"
"Reestablishment"
"Reindeerflotilla"

UWE Bristol | University of the West of England

Introduction - Roadmap

- Background – Inference Attacks and The Secret Sharer
- Methodology – Model Architecture, Training Data, Attack Design
- Results & Discussion
- Conclusion and Further Work

Model Inversion Attacks

- Fredrikson, Jha, and Ristenpart (2015) *Model inversion attacks that exploit confidence information and basic countermeasures*

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Memorisation & Exposure

- Carlini et al (2019) *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks*

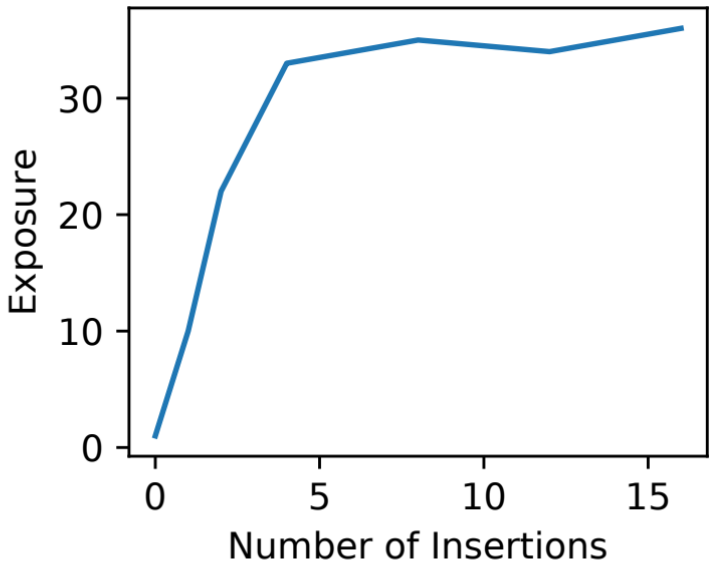
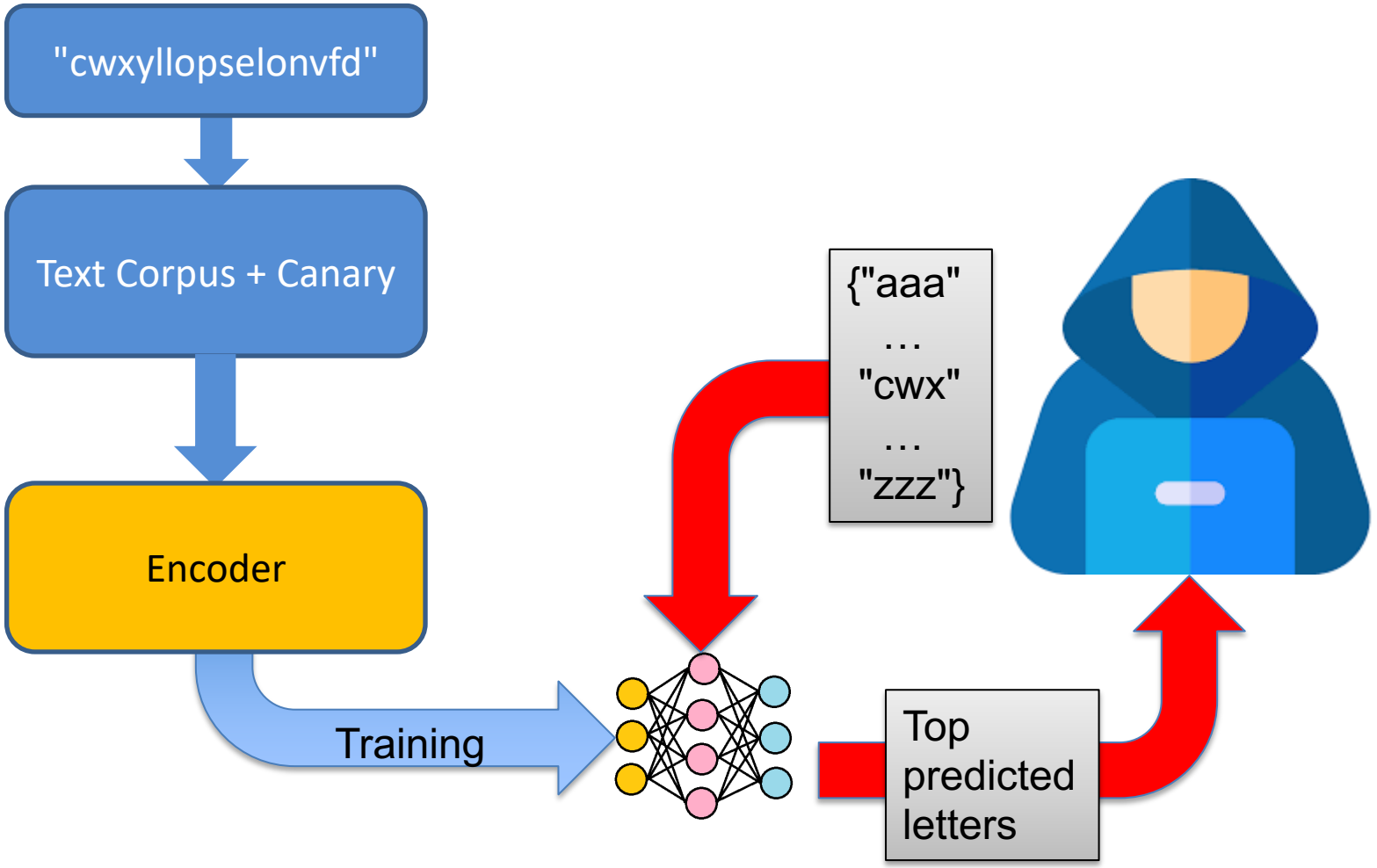


Figure 6: Exposure of a canary inserted in a Neural Machine Translation model. When the canary is inserted four times or more, it is fully memorized.

Our Approach



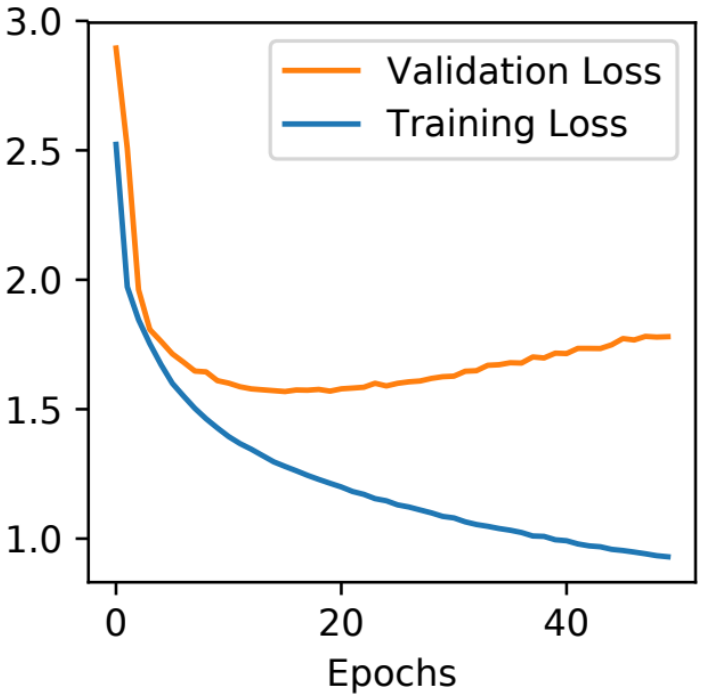
Results

Length (chars)	No. of Words	Algorithm	Candidates	Success%
1000	121	Simple	26.65	90
2000	208	Simple	35.25	95
4000	368	Simple	19.45	60
4000	368	Deep	21952	90
16000	1049	Deep	24389	10

TABLE I

SUMMARY OF RESULTS SHOWING THE SUCCESS OF OUR ALGORITHMS IN EXTRACTING A PASSWORD EMBEDDED IN VARIABLE-LENGTH TEXT CORPORA.

Overfitting?



- How do we define generalisation? What is the validation set?

Results in Perspective

- AES-CBC-256 encryption
- Using strongly random key
- Not brute-forceable on human timescale
  - Small chance to crack it with a 3-5-character search?

Conclusions & Future Work

- Language models are vulnerable to being mined for secrets.
- Mitigations – Sanitisation, Password construction, Model encryption/SMC
- Larger models, real devices
- GDPR Issues?

Thank you!

- Any questions?

UWE Bristol | University of the West of England