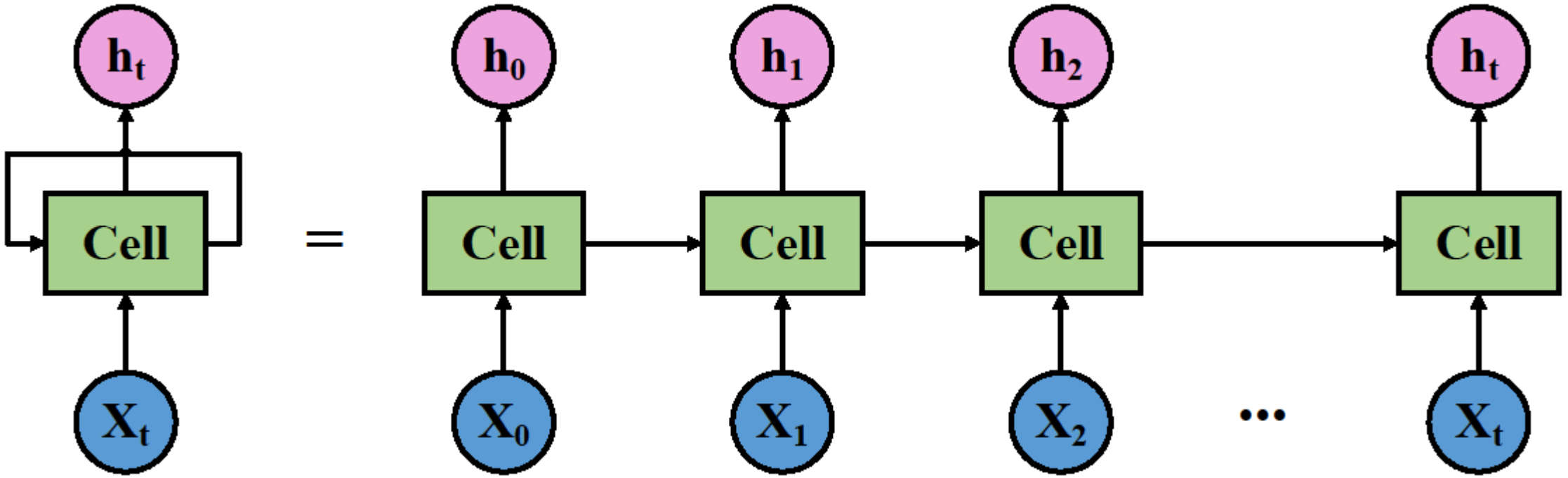
Privacy Leakage in Long Short Term Memory

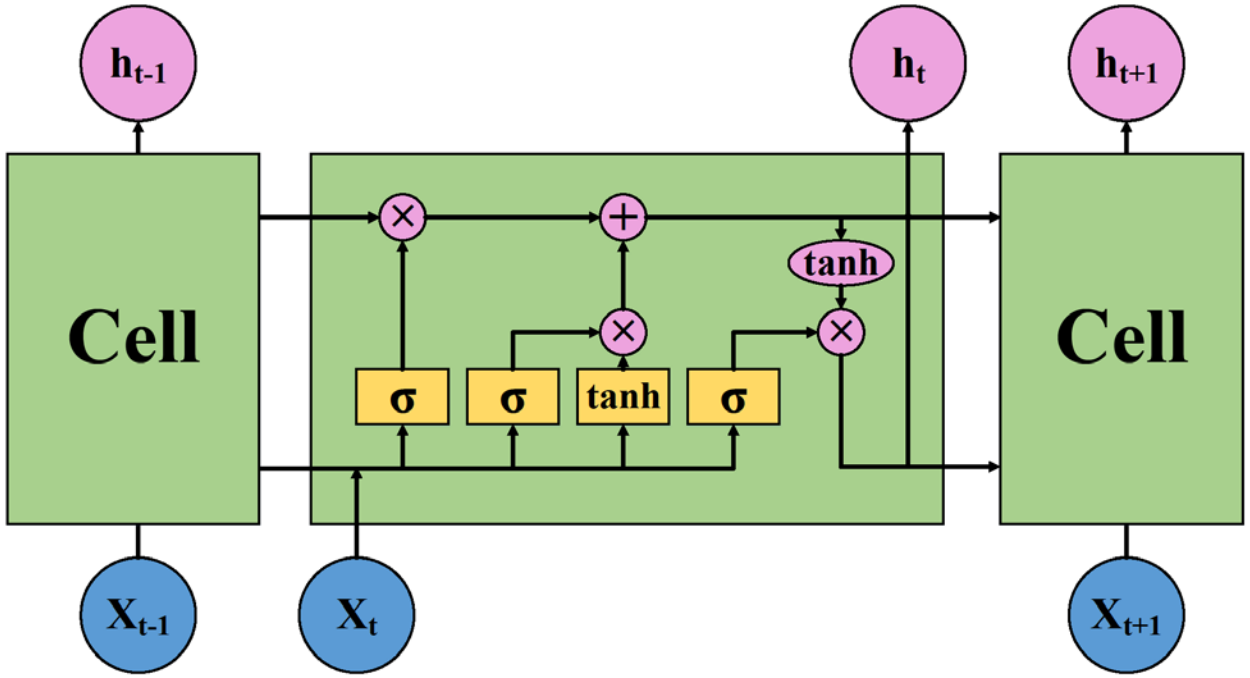
Lun Wang
lun.wang@pku.edu.cn
Peking University
Advisor: Chang Liu, Dawn Song
Thanks: Nicholas Carlini

- Brief Overview of Long Short Term Memory
- Privacy Leakage in Long Short Term Memory
- Targeted Attack on Privacy Leakage
- Conclusion
- References

Brief Overview of Long Short Term Memory



Brief Overview of Long Short Term Memory



$$LSTM : h_t^{l-1}, h_{t-1}^l, c_{t-1}^l \rightarrow h_t^l, c_t^l$$

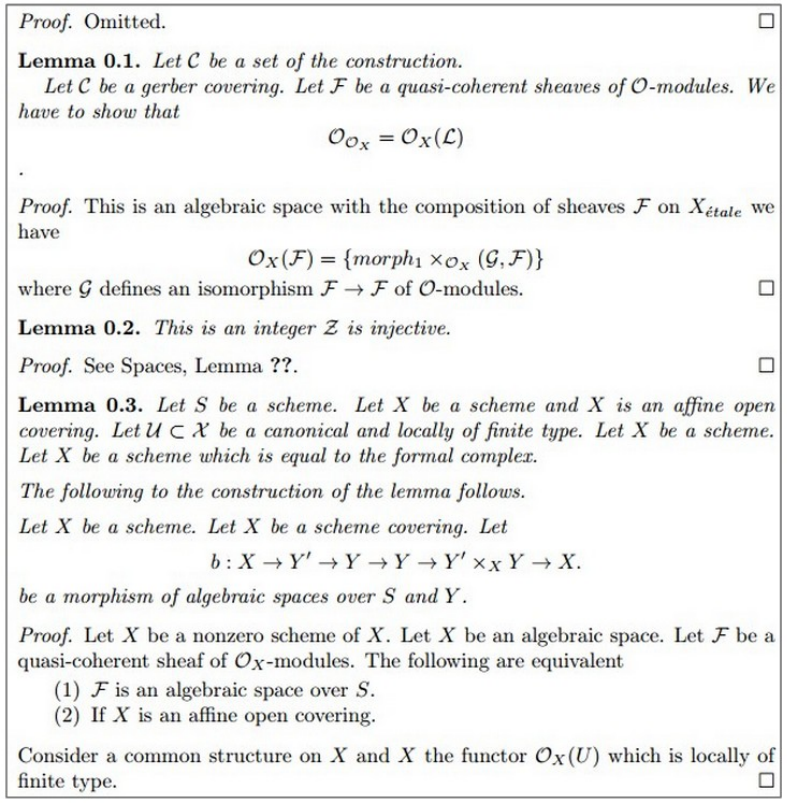
$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} sigm \\ sigm \\ sigm \\ tanh \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$c_t^l = \left\langle f, c_{t-1}^l \right\rangle + \langle i, g \rangle$$

$$h_t^l = \left\langle o, tanh(c_t^l) \right\rangle$$

Brief Overview of Long Short Term Memory

- Language modeling
- Speech Recognition
- Machine Translation



Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma ??. □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

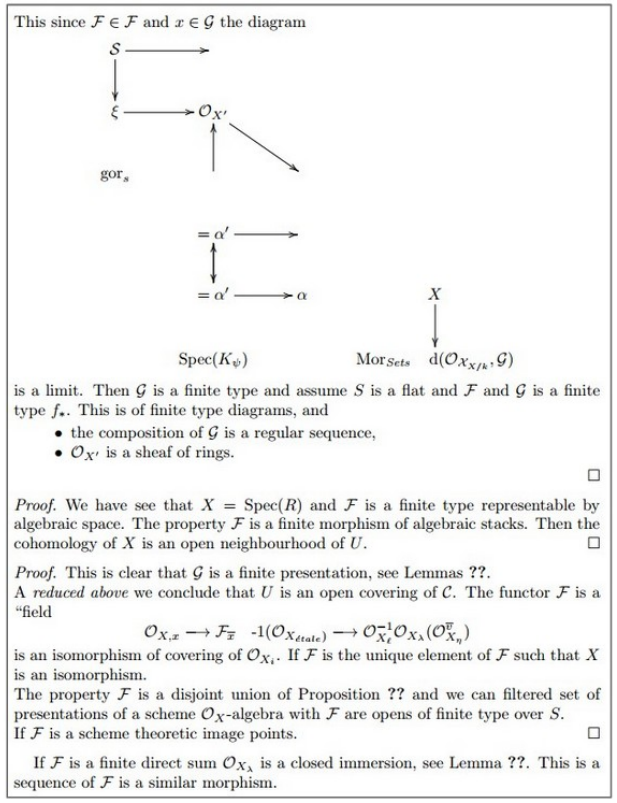
$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □



This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram

$$\begin{array}{ccc} \mathcal{S} & \longrightarrow & \\ \downarrow & & \\ \xi & \longrightarrow & \mathcal{O}_{X'} \\ & & \uparrow \\ \mathrm{gor}_s & & \end{array} \qquad \begin{array}{ccc} & = \alpha' & \longrightarrow \\ & \updownarrow & \\ & = \alpha' & \longrightarrow \alpha \end{array} \qquad \begin{array}{c} X \\ \downarrow \\ \mathrm{d}(\mathcal{O}_{\mathcal{X}_{Ik}}, \mathcal{G}) \end{array}$$

$$\mathrm{Spec}(K_\psi) \qquad \mathrm{Mor}_{Sets}$$

is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type f_*. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \mathrm{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\overline{x}} \quad \text{-}1(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}_{X_t}^{-1}\mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\eta}^{\overline{s}})$$

is an isomorphism of covering of $\mathcal{O}_{X_t}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. □

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

The Unreasonable Effectiveness of Recurrent Neural Networks. Andrej Karpathy.

- Brief Overview of Long Short Term Memory
- Privacy Leakage in Long Short Term Memory
- Targeted Attack on Privacy Leakage
- Conclusion
- References

I. Training Text

- Penn Treebank dataset
- Insertion Rate
- Cardinality

Table 1: Classes of Private Information

Type	Example
SSN	123-45-6789
CCN	1111 2222 3333 4444
Address	Mr Harry Potter Ginger Street, Maple City CA 90230 U.S.A

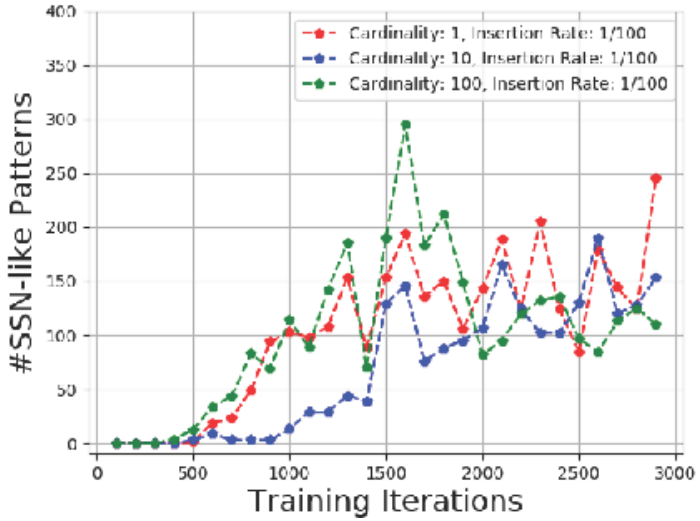
II. Configuration

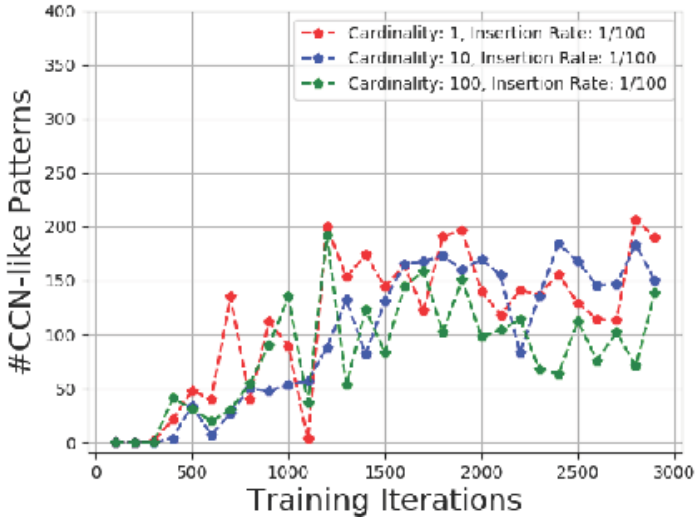
- Ubuntu 16.04.2 LTS
- GeForce GTX TITAN X

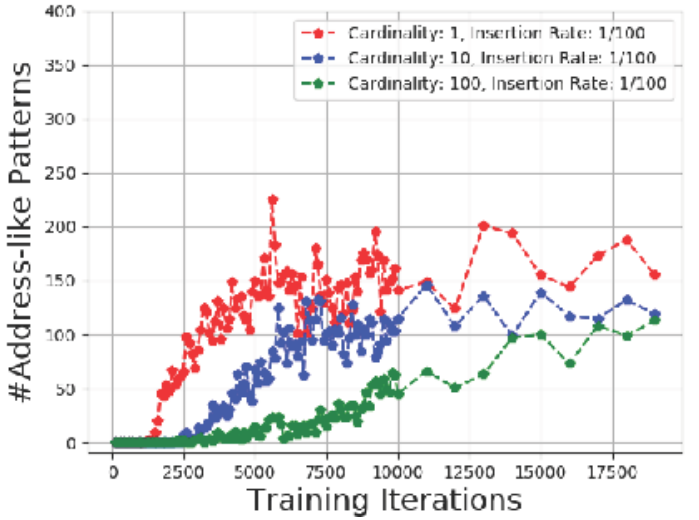
Table 2: Hyper-parameter Setting in Experiments

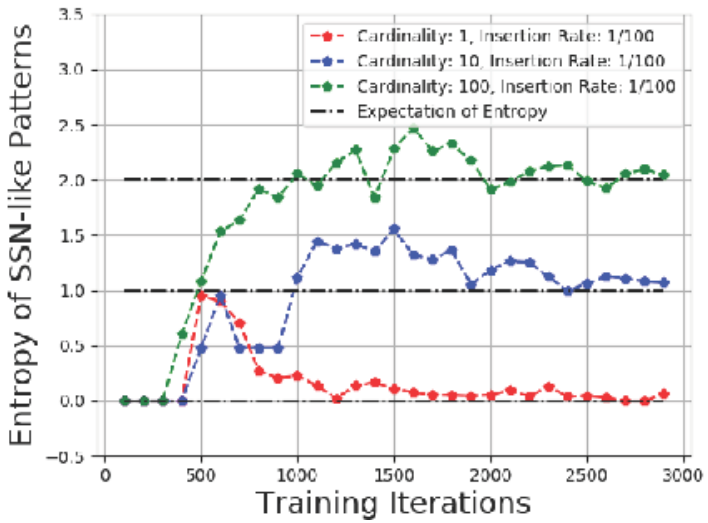
Hyper-parameter	Setting
Layer	3
Hidden Size	512
Training Batch Size	10
Sampling Batch Size	1
Attacking Batch Size	Same as seed length
Unrolled Step	50
Dropout	1
Alphabet Size	Number of different characters in the training text
Initial Learning Rate	0.001
Decay Rate	1
Gradient Clip	5

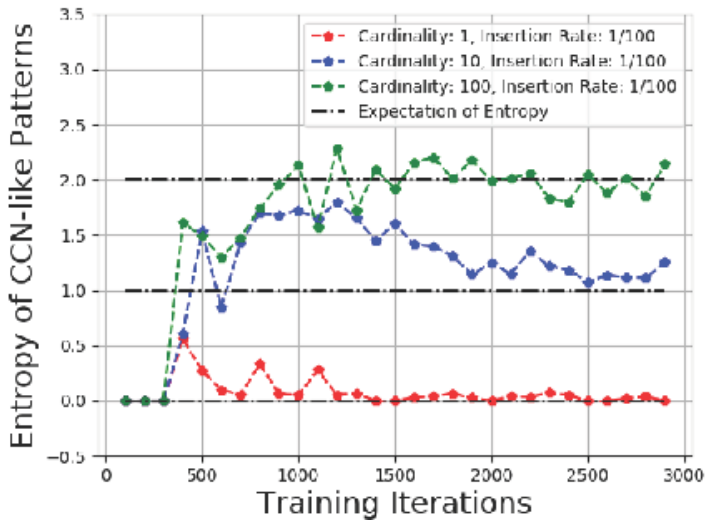
III. Experimental Results

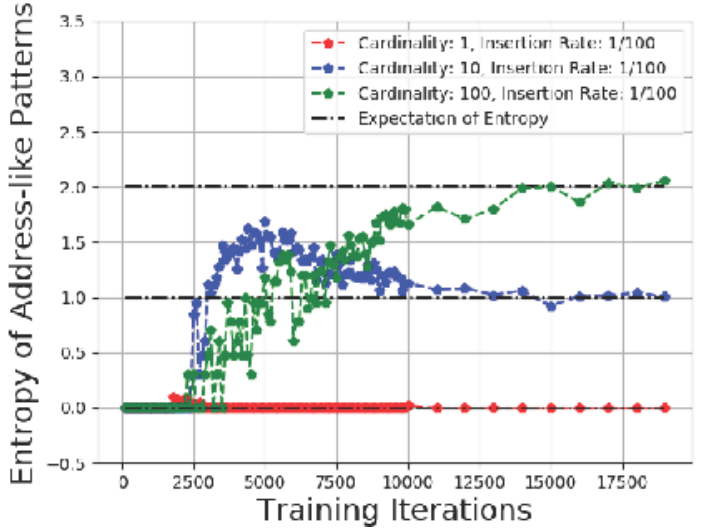




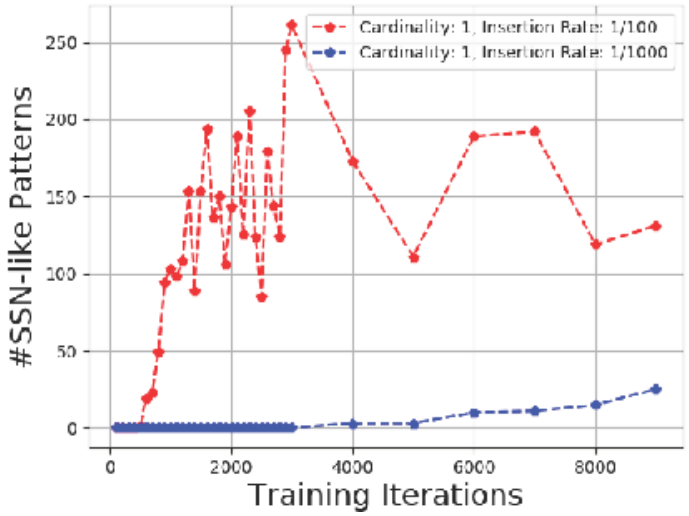


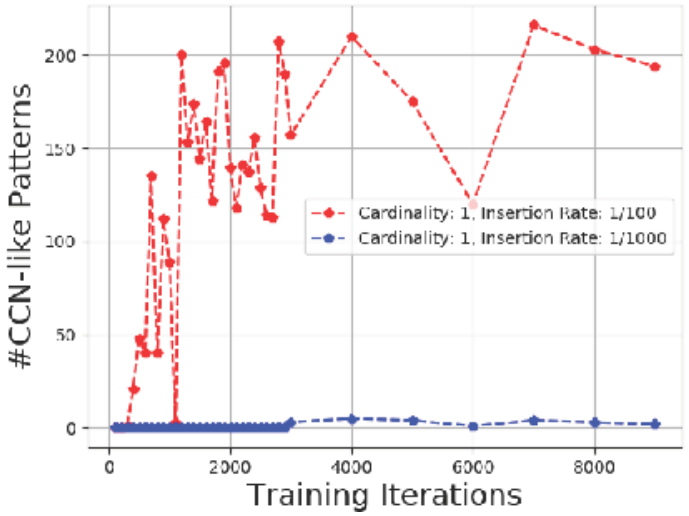


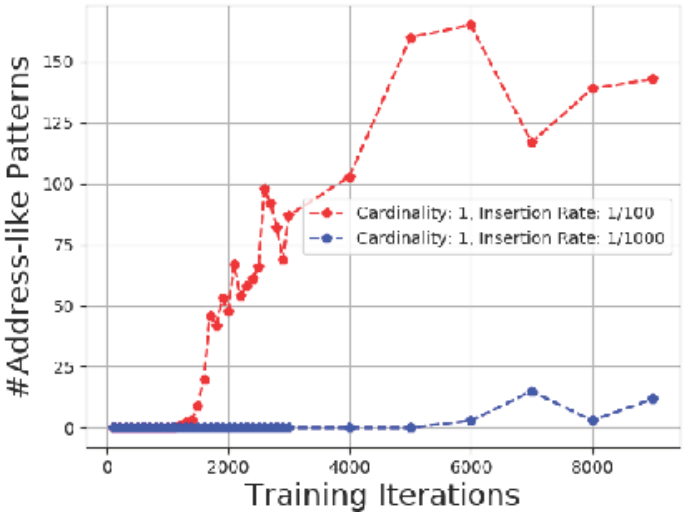




III. Experimental Results

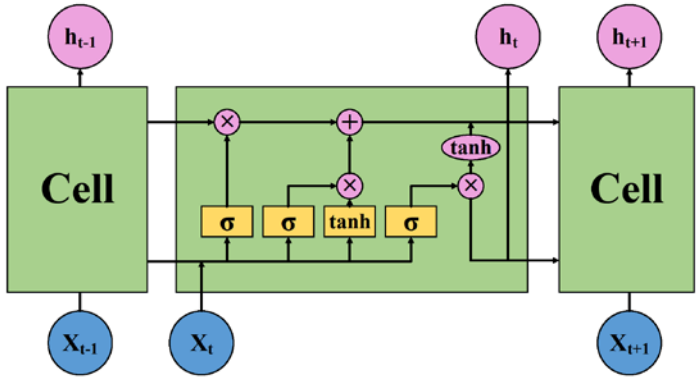






- Brief Overview of Long Short Term Memory
- Privacy Leakage in Long Short Term Memory
- Targeted Attack on Privacy Leakage
- Conclusion
- References

I. Targeted Attack

Attacking Seed → 

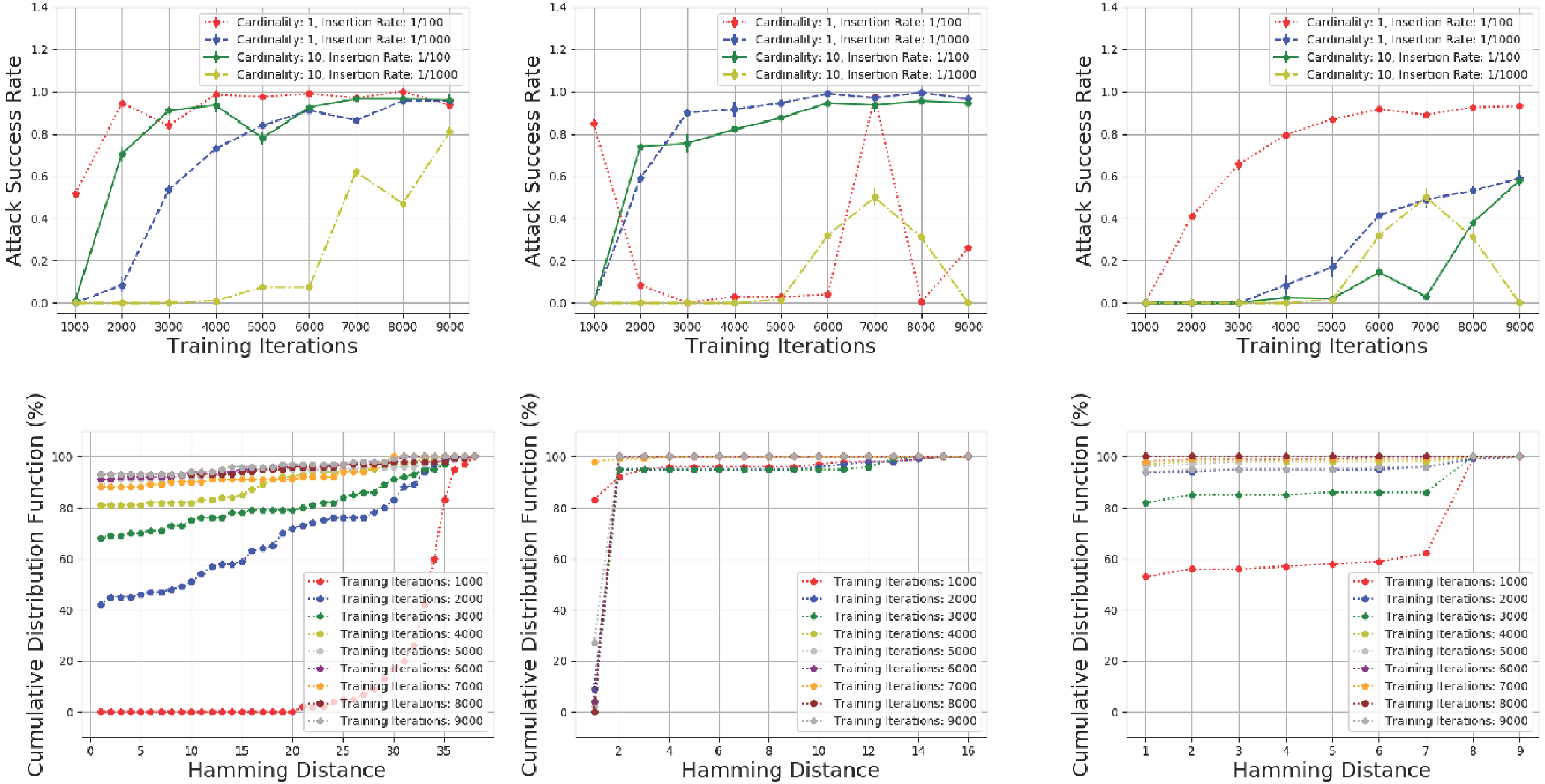
→ Private Information

SSN 123 → ••••••••• → SSN 123-45-6789

CCN 0000 → ••••••••• → CCN 0000-0000-0000-0000

Mr.Harry → ••••••••• → Mr.Harry Potter, Ginger Street, Maple City, CA 90230, U.S.A

II. Experimental Results



- Brief Overview of Long Short Term Memory
- Privacy Leakage in Long Short Term Memory
- Targeted Attack on Privacy Leakage
- Conclusion
- References

Conclusion

In this paper, we explored privacy leakage and targeted attack in LSTM model. The experimental results show that LSTM is likely to give out private information in training. The attack against privacy leakage is also simple and feasible. All the results show that the training set should be properly anonymized to protect privacy of data providers.

- Brief Overview of Long Short Term Memory
- Privacy Leakage in Long Short Term Memory
- Targeted Attack on Privacy Leakage
- Conclusion
- References

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks,5(2):157–166, 1994.
[3] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on
, pages 6645–6649. IEEE, 2013.
[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
[5] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn
treebank. Computational linguistics, 19(2):313–330, 1993.
[6] M. Sundermeyer, R. Schluter, and H. Ney. Lstm neural networks for language modeling. In Thirteenth Annual
Conference of the International Speech Communication Association, 2012.

Q & A