Privacy-preserving Neural Representations of Text

Maximin Coavoux – Shashi Narayan – Shay B. Cohen

University of Edinburgh – ILCC

EMNLP 2018 – Brussels

Context: Privacy and Neural Networks

- Machine learning uses data (e.g. UGC) susceptible to contain private/sensitive information
  - Privacy risks when collecting data, releasing data, releasing model, ...
  - User perspective: use machine learning based services but avoid sharing personal data unnecessarily
  - Data controller: accountability for the safety of personal data

- Privacy-related vulnerability example (Carlini et al., 2018)
  - Sample from pretrained language model to reconstruct sentences from the training set and discover 'secrets' in training data
  - → The parameters of a released pretrained model may expose private information

Privacy and Neural Networks: NLP

- Private information **explicitly** stated in text:
  - Name, phone number, email address, medical information, credit card number . . .
  - can be preprocessed out of training data

Privacy and Neural Networks: NLP

- Private information **explicitly** stated in text:
  - Name, phone number, email address, medical information, credit card number ...
  - can be preprocessed out of training data
- or **implicit**, i.e. predictable from linguistic features of text
  - age, gender (Schler et al., 2006)
  - native language (Malmasi et al., 2017)
  - authorship (Shrestha et al., 2017)
  - ...

    "[...] language is a proxy for human behavior, and a strong signal of individual characteristics" (Hovy and Spruit, 2016)

  - implicit information cannot be easily removed from text

Privacy and Neural Networks: NLP

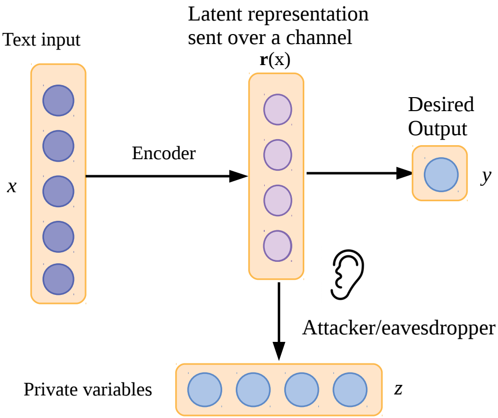
- Private information **explicitly** stated in text:
  - Name, phone number, email address, medical information, credit card number . . .
  - can be preprocessed out of training data
- or **implicit**, i.e. predictable from linguistic features of text
  - age, gender (Schler et al., 2006)
  - native language (Malmasi et al., 2017)
  - authorship (Shrestha et al., 2017)
  - . . .

    "[. . .] language is a proxy for human behavior, and a strong signal of individual characteristics" (Hovy and Spruit, 2016)

  - implicit information cannot be easily removed from text
- textual input $\approx$ demographic characteristics of author

Privacy and Neural Networks: Research Questions

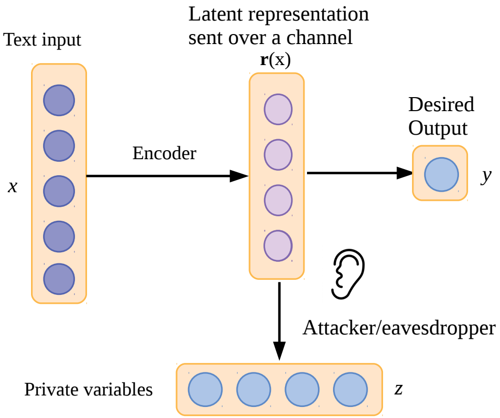
- If an attacker eavesdrops on the hidden representation of a neural net, what can they guess about the input text?
- Can we improve the privacy of the latent representation $\mathbf{r}(x)$?



Scenario:

- Text classifier (topic, sentiment, spam, etc..) shared across several devices:
  1. Text-to-vector encoder
  2. Classifier itself
- Latent representation intercepted by attacker and exploited to recover private information about the text

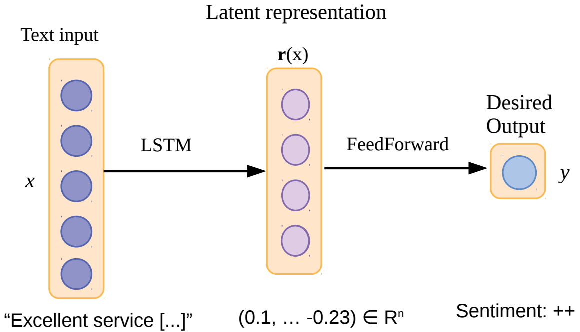
Contributions



1. Measuring the privacy of neural representations with the ability of an attacker to recover private information
2. Improving the privacy of neural representations using adversarial training

Measuring Privacy: Target Model

- x: text input (sequence of tokens)
- $\mathbf{r}(x) = \mathrm{LSTM}(x)$: latent representation
- y: text label (topic, sentiment, etc) predicted by feedforward net

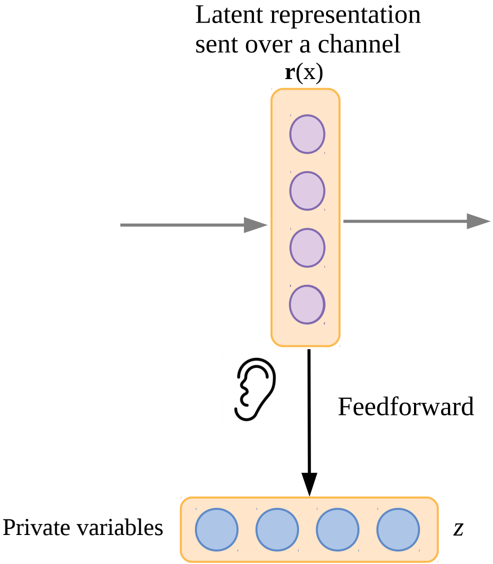


Measuring Privacy: Attacker's Setting – Classifier

- Attacker's model: feedforward net

$$P(\mathbf{z}|\mathbf{r}(x)) = \text{FeedForward}(\mathbf{r}(x))$$

- Target private variables:
  - age and gender of author
  - named entities that occur in the text
- Representation is private if the attacker cannot recover these variables accurately
- Note: a 'private' representation should resist any type of classifier; we only experiment with a tuned feedforward net



Measuring Privacy: Attacker's Setting – Dataset

- The attacker needs to train a model on a dataset of $(\mathbf{r}(x), \mathbf{z})$ pairs.
- Can use the dataset of the text classifier if available
- Otherwise, the attacker can construct a dataset from:
  - Any collection of texts annotated with private variables $\left\{(x^{(i)}, \mathbf{z}^{(i)})\right\}$, e.g. scraped from social networks
  - The encoder function $\mathbf{r}$ of the target classifier, assumed to be publicly available

How well can an attacker predict private variables from latent representations?

- Trustpilot dataset (Hovy et al., 2015):
  - sentiment analysis on users' reviews
  - divided in 5 subcorpora depending on location of author
  - private variables: self-reported gender and age of authors

	Most frequent label		Attacker	
	Gender	Age	Gender	Age
TP (Denmark)	61.6	58.4	62.0 (+0.4)	63.4 (**+5.0**)
TP (France)	61.0	50.1	61.0 (+0)	60.6 (**+10.5**)
TP (Germany)	75.2	50.9	75.2 (+0.4)	58.6 (**+7.9**)
TP (UK)	58.8	56.7	59.9 (**+1.1**)	61.8 (**+5.1**)
TP (US)	63.5	63.7	64.7 (**+1.2**)	63.9 (+0.2)

- **The latent representations contain a signal for private variables even though they were not trained to.**
  LSTM incidentally learns private variables

Improving the Privacy of Latent Representations

- Problem statement: learn an LSTM that produces
  - **useful** representations (contain information about text label)
  - **private** representations (contain no information about private variables)
- We introduce two methods based on **adversarial training** (+ third method based on distances, not in this talk, see paper)
- ⚠ both objectives (privacy and utility) contradict each other since some of the private variables might be actually correlated with the text labels.
- Improving privacy might come at a cost in accuracy → **tradeoff**

Defense Method 1: Adversarial Classification

- We simulate an attacker at training time who predicts private variables from latent representations and optimizes:

$$\mathscr{L}_{\text{attacker}} = -\log P(\mathbf{z}|\mathbf{r}(x))$$

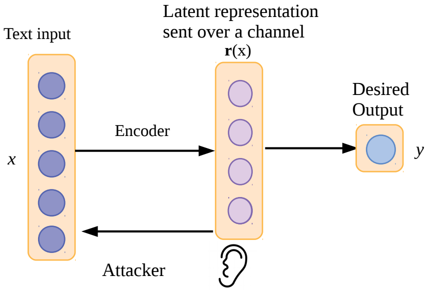
- The main model has a double objective:
  - Maximize the likelihood of the text label (maximize utility)
  - Confuse the attacker (maximize privacy) by updating the parameters of **r**

$$\mathscr{L}_{\text{classifier}} = -\log P(y|x) \; - \mathscr{L}_{\text{attacker}}$$

- Both agents have their own parameters (similar to GANs):
  - Attacker only updates its feedforward net parameters but cannot modify the parameters of **r**
- To evaluate privacy, a new attacker is trained from scratch

Defense Method 2: Adversarial Generation

- Limitation of adversarial classification: you must know in advance which private variables you need to obfuscate



- Instead of maximizing the likelihood of the private variables, the adversary optimizes a language model objective:

$$\mathcal{L}_{\text{attacker}} = -\log P(x|\mathbf{r}(x))$$

$\rightarrow$ learn to **reconstruct the full text** x from its latent representation $\mathbf{r}(x)$

- The objective of the main classifier stays the same:

$$\mathcal{L}_{\text{classifier}} = -\log P(y|x) \; - \mathcal{L}_{\text{attacker}}$$

Experiments: Datasets

Datasets	private variables
Sentiment Analysis	
Trustpilot, reviews (Hovy et al., 2015)	age, gender of author
Topic Classification	
AG news (Gulli, 2005)	named entities
DW news (Pappas and Popescu-Belis, 2017)	named entities
Blog posts (Schler et al., 2006)	age, gender of author

Experiments: Results

- Privacy measure: 100 – accuracy of attacker (higher is better)
- Evaluation of effect of defense methods on (i) accuracy (ii) privacy (model selection on development accuracy)
- Main result: defense methods **improve privacy** with a (mostly) small cost in accuracy.

Corpus	Standard		1. Adversarial classifier		2. Adversarial generation	
	Acc.	Priv.	Acc.	Priv.	Acc.	Priv.
Sentiment						
TP Germany	85.1	32.2	-0.6	-0.3	-1.3	**+0.6**
TP Denmark	82.6	28.1	-0.2	**+4.4**	-0.1	**+6.0**
TP France	75.1	41.1	-0.8	**+0.7**	-1.4	**-6.4**
TP UK	87.0	39.3	-0.5	**+0.9**	-0.2	+0.2
TP US	85.0	33.9	-0.1	**+2.6**	-0.2	**+1.8**
Topic						
AG news	76.5	33.7	**-14.5**	**+14.5**	+0.2	**-7.8**
DW news	44.3	78.3	**-5.7**	**+21.7**	**+5.9**	**+13.1**
Blogs	58.3	40.8	-0.8	**+3.4**	**+1.1**	**+0.9**

Conclusion

- Latent representations for texts contain a signal for private information
- Measure privacy of latent representation by the ability of an attacker to recover private information from it.
- Improve representation privacy with defense methods based on adversarial training
- `github.com/mcoavoux/pnet`

Conclusion

- Latent representations for texts contain a signal for private information
- Measure privacy of latent representation by the ability of an attacker to recover private information from it.
- Improve representation privacy with defense methods based on adversarial training
- `github.com/mcoavoux/pnet`

Thank you for your attention!

References I

Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *CoRR*, abs/1802.08232, 2018. URL http://arxiv.org/abs/1802.08232.

A. Gulli. The anatomy of a news search engine. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pages 880–881, New York, NY, USA, 2005. ACM. ISBN 1-59593-051-5. doi: 10.1145/1062745.1062778. URL http://doi.acm.org/10.1145/1062745.1062778.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://anthology.aclweb.org/P16-2096.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 452–461, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3. doi: 10.1145/2736277.2741141. URL https://doi.org/10.1145/2736277.2741141.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. A Report on the 2017 Native Language Identification Shared Task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Nikolaos Pappas and Andrei Popescu-Belis. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I17-1102.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs - Papers from the AAAI Spring Symposium, Technical Report*, volume SS-06-03, pages 191–197, 8 2006. ISBN 1577352645.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/E17-2106.