Assessing Unintended Memorization in Neural Discriminative Sequence Models

Mossad Helali, Thomas Kleinbauer, Dietrich Klakow

Assessing Unintended Memorization in Neural Discriminative Sequence Models

Mossad Helali[0000−0001−7490−5011], Thomas Kleinbauer, and Dietrich Klakow

Spoken Language Systems Group
Saarland Informatics Campus
Saarland University
Saarbrücken, Germany
{mhelali,thomas.kleinbauer,dietrich.klakow}@lsv.uni-saarland.de

Abstract. Despite their success in a multitude of tasks, neural models trained on natural language have been shown to memorize the intricacies of their training data, posing a potential privacy threat. In this work, we propose a metric to quantify unintended memorization in neural discriminative sequence models. The proposed metric, named d-exposure (discriminative exposure), utilizes language ambiguity and classification confidence to elicit the model's propensity to memorization. Through experimental work on a named entity recognition task, we show the validity of d-exposure to measure memorization. In addition, we show that d-exposure is not a measure of overfitting as it does not increase when the model overfits.

Keywords: Named Entity Recognition · Natural Language Understanding · Privacy

1 Introduction

Neural networks have become prevalent in numerous machine learning tasks in general and in natural language processing in particular. An issue that has been identified with neural models, however, is that they tend to memorize their training data [7,10,2]. Memorization raises severe privacy concerns in cases where such models are trained on datasets that contain sensitive information such as credit card numbers, passwords, etc. If such models are deployed e.g. on smartphones [5] or as a service [4], they give attackers access to the memorized sensitive information.

The focus of this paper is on *unintended* memorization, which occurs when models retain information that are orthogonal to the learning task. For example, for the task of named entity recognition (NER) on a dataset of emails, memorizing passwords that appear in the dataset is unintended. Existing work focuses on

This research has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 3081705 – COMPRISE (https://www.compriseh2020.eu/)

neural generative sequence models, such as language models and machine translation models, and uses model perplexity to quantify unintended memorization [2]. In this paper, we propose a metric for discriminative models where perplexity cannot be utilized. The idea is to give practitioners the means to assess the degree of memorization in models intended for deployment, allowing them, e.g., to choose hyper-parameter settings that minimize privacy-threatening information leakage.
Our main contributions are:

- A method for quantifying memorization in discriminative models. This involves inserting specifically designed ambiguous phrases into the training set of the model and analyzing the model's confidence with respect to the created phrases. The proposed metric is named *d-exposure* (for discriminative exposure).
- An experimental validation of the proposed definition on a competitive neural NER model and benchmark dataset. As in previous work, we find that exposure increases with the number of repetitions of inserted phrases in the training set. In addition, we confirm that d-exposure is not a measure of overfitting as unintended memorization does not increase when the model starts to overfit.

2 Related Work

In one of the earliest studies on memorization in neural networks, Zhang et al. [10] show that neural networks have the capability to fit data with random labels, meaning that state-of-the-art models are at risk of memorization. Song et al. [7] present a method to create neural models that memorizes the training data with no noticeable difference in utility. This raises concerns because utility is often the main criterion for deciding which model to deploy. In their analysis of memorization, Arpit et al. [1] show that memorization is not only dependent on the model, but also on the dataset. While these works are important in the analysis of memorization, they do not provide a quantitative method for gauging the depth of the problem.

The first work on assessing *unintended* memorization in neural models on language tasks was by Carlini et al. [2]. To assess unintended memorization, they define a metric, named *exposure*, that is based on comparing the perplexity $\mathrm{Px}(s)$ of a random phrase s inserted into the training set with the perplexities of other phrases from the same random space. The basic tenet is that a significantly lower perplexity of the inserted phrase vs. those of the other random phrases signals that the neural model has unintentionally memorized that phrase. Specifically, for a random phrase s inserted into the training set of a model θ, exposure is defined as:

$$\mathbf{exposure}_\theta(s) = -log_2 \mathop{\mathbf{Pr}}_{r \in \mathcal{R}}[\mathrm{Px}_\theta(r) \leq \mathrm{Px}_\theta(s)] \tag{1}$$

where $\mathcal{R}$ is the random space of all such phrases. Note that high memorization, i.e., low perplexity, is reflected by high exposure values. The authors test

their definition empirically and conclude that memorization is not directly linked to overfitting but rather to the learning process itself, making memorization a prevalent issue in state-of-the-art neural models. However, the authors' approach is limited to generative sequence models because the definition of exposure is based on perplexity. We show below how a similar line of reasoning can be utilized for an exposure measure on discriminative models.

Another related notion to the problem of memorization is membership inference attacks, where an attacker tries to infer whether a set of samples belong to the dataset of a trained model. Truex et al. [9] have done an extensive analysis on how such attacks can be carried out and on the vulnerability of the models under attack. Though membership inference is related to our work, there are notable differences between the approaches. First, the goal of our work is to measure the model's propensity to leaking information, not analyzing whether the model can be attacked. For example, for an overfit model, membership inference probability increases [6], while memorization is not correlated with overfitting. Moreover, calculating exposure is a simpler procedure that does not involve building shadow datasets and attack models as in membership inference.

3 Approach

The existing definition of exposure in [2] is inapplicable to discriminative models because it is based on perplexity, which is not supplied by discriminative models. Instead, such models output for each class a level of confidence that the input word belongs to that class. This motivates a definition of exposure *per class* as it can behave differently for each class. While exhaustive enumeration of perplexity is inefficient [2], it is feasible to enumerate the model's confidence for all words in each class because these are in the magnitude of only a few thousands, depending on the the dataset.

Intuitively, memorizing is the opposite of generalizing. A good model will classify an unambiguous sentence with high confidence. For example, in the sentence "I prefer Germany", the last word should clearly be labeled as a location in an NER task. Polysemous words, however, may constitute different named entities depending on the context. For instance, the word "Jordan" could refer to a person (e.g., Michael Jordan), a location (e.g. the country of the same name), or an organization (e.g. The Jordan Company). If some of these cases appear with roughly the same frequency in the training data, an ambiguous test sentence, such as "I prefer Jordan", should thus be classified with low confidence. Even adding the same sentence to the training data should not change this – unless the model tends to memorize sentences. In other words, an unexpected high confidence in the classification of an ambiguous sentence hints at the possibility of unintended memorization in a given model. We base our definition of d-exposure on this notion and follow the general procedure given by Carlini et al. [2].

3.1 d-exposure for Discriminative Models

Given a fixed phrase that has a word $\boldsymbol{s}$ with multiple possible class labels, we insert the phrase in the training set with $\boldsymbol{s}$ labeled as C_i and train the model θ. d-exposure for class C_i is then given by:

$$\textbf{d-exposure}_{\theta,C_i}(s) = -log_2 \underset{w \in C_i}{\textbf{Pr}} [\text{conf}(w) \geq \text{conf}(s)] \tag{2}$$

where $\text{conf}(s)$ is the confidence returned by the model when labeling $\boldsymbol{s}$. Therefore, d-exposure has a value $\in [0, log_2|C_i|]$ with $|C_i|$ denoting the number of words that are labeled only as C_i. Maximum d-exposure is obtained when $\boldsymbol{s}$ has the highest confidence (high memorization) and vice versa. Note that this is the case if all words are assigned the correct class. If $\boldsymbol{s}$ is labeled incorrectly, however, d-exposure is defined to be zero. On the other hand, if other words in C_i are incorrectly labeled, they are treated as having lower confidence than $\boldsymbol{s}$, because the model classified the ambiguous phrase correctly while failing to correctly classify the clear one. We apply the same process for other entity classes in the dataset and calculate d-exposure of the model as:

$$\textbf{d-exposure}_{\theta}(s) = \frac{1}{N} \sum_{C_i} \text{d-exposure}_{\theta,C_i}(s) \tag{3}$$

where N is the number of classes. This definition allows one to ignore classes that are considered irrelevant for the task at hand. For example, if one is interested in measuring the memorization of their model on the names of persons and locations only, one could simply compute d-exposure for these two classes. Recall that the purpose of the metric is to guide the choice of model settings before deployment. Which phrases and classes to consider are choices made by the user.

4 Experimental Validation

In this section, we experimentally test the proposed definition of d-exposure in order to: (1) show its validity as a measure of unintended memorization in discriminative models, and (2) demonstrate that d-exposure is not linked to overfitting. We show our results on a named entity recognition task as an example of discriminative models.

4.1 Setup

We conduct our experiments on CoNLL-2003 [8], a popular NER dataset in English. In our experiments, we focus on the tags: S-PER, S-ORG, and S-LOC. We discard S-MISC to decrease the variability as including it would lead to the inserted phrase having multiple correct labels (based on the definition of S-MISC). Table 1 shows statistics of these classes in CoNLL-2003 dataset. The first column is the number of unique entities that belong only to the respective class ($|C_i|$); the second column is the number of unique entities that have more

Table 1. Statistics of the chosen classes in the training set of CoNLL-2003.

Label	Exclusive	Overlapped	Frequency
S-ORG	1001	101	3836
S-PER	949	19	2316
S-LOC	937	97	6101

Table 2. Number of occurrences of the chosen entities in the training set of CoNLL-2003.

Word	S-PER	S-ORG	S-LOC
Williams	7	8	0
Chelsea	5	6	0
Melbourne	0	4	5

than one possible label (i.e. candidates for $\boldsymbol{s}$); the third column is the frequency of each class in the training set.

For the inserted phrase, we choose the ambiguous format: "There are many people who like _____", which allows entities of the three types to fill the blank. For the chosen entities, "Williams" was inserted as S-ORG, "Chelsea" as S-PER and "Melbourne" as S-LOC. We chose these entities because their occurrences in the training set are more balanced than others. Table 2 shows the number of occurrences of these entities as each class. That said, we found out that the general behavior of d-exposure does not change based on the chosen entities, as long as they are not highly imbalanced towards one class, nor does it change based on the format, as long as it is ambiguous.

For the model, we use a BiLSTM with GloVe embeddings, SGD optimizer, dropout (50%) and learning rate decay, implemented with Targer[1], a neural tagging library [3]. This model achieves an F1 of 90.0 on CoNLL-2003 dataset.

4.2 Repeated occurrences in the Training Set

In this experiment, we test whether d-exposure increases with the number of times the chosen phrase appears in the training set. The intuition is that the more the model sees the sentence, the higher the incentive to memorize it. For this matter, we insert the chosen sentences 4, 16, 64, 128 and 256 times and observe d-exposure for each category. Figure 1 shows the effect of the number of repetitions of the inserted sentence on d-exposure. As expected, d-exposure generally increases with the number of repetitions, implying that repeated occurrence of a sentence in the training set tends to produce higher memorization. Another observation is that d-exposure does not behave the same in all classes. Rather, it is much lower for S-LOC than the other two. This validates our claim that exposure is to be measured per-class as different classes occur in different contexts but the exact reasons for the differing behavior require further investigation. In additional experiments with other model architectures not detailed here, we found the same general trend in the curves but the behavior of S-PER and S-LOC reversed. Table 3 shows d-exposure evaluated at different epochs (columns) and number of repetitions (rows) for the three classes. The first row is the value of d-exposure when the phrase is not inserted in the training set.

[1] https://github.com/achernodub/targer

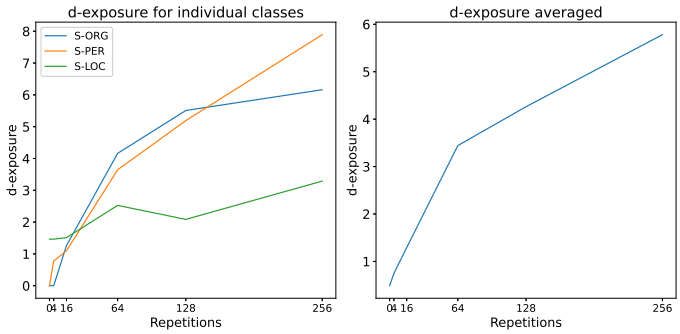

Fig. 1. d-exposure vs. repetitions for individual classes and averaged on CoNLL-2003 at 150 epochs.

4.3 Overfitting

In this experiment, we observe the behavior of d-exposure against overfitting. We conduct this analysis to confirm that exposure is not a measure of overfitting but rather of memorization. If it was so, we expect it to reach its maximum value for all classes when overfitting begins or to keep increasing while the model is overfitting. To make the model overfit, we train it only on 10% of the training data, increase the number of epochs to 250 and disable learning rate decay and dropout. Figure 2 shows the results when the phrases are repeated 16 times. d-exposure increases as the model is learning and stops increasing when overfitting begins. In addition, maximum d-exposure for S-LOC (8.0) or S-ORG (8.1) is not reached at any point. Recall that the maximum d-exposure for a class C_i is $log_2|C_i|$, where $|C_i|$ is the number of entities belonging only to that class. For S-PER, however, maximum d-exposure (7.5) is reached only at stages where the model has not yet overfit. Therefore, we conclude that d-exposure is not correlated with overfitting and for the case of S-PER, the model has higher memorization. Similar results were found for different numbers of repetitions.

Table 3. d-exposure for the classes S-LOC, S-PER, and S-ORG for different numbers of repetitions (rows) and epochs (columns).

	S-LOC						S-PER						S-ORG					
	25	**50**	**75**	**100**	**125**	**150**	**25**	**50**	**75**	**100**	**125**	**150**	**25**	**50**	**75**	**100**	**125**	**150**
0	1.76	1.42	1.51	1.33	1.55	1.46	0.00	0.00	0.00	0.31	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	1.87	1.68	1.47	1.71	1.57	1.46	0.00	0.43	0.48	0.59	0.72	0.78	0.00	0.00	0.00	0.82	0.00	0.00
16	1.87	1.74	1.72	2.00	1.78	1.51	0.49	0.58	0.82	0.98	0.79	1.10	0.00	1.12	0.67	1.28	1.28	1.26
64	2.68	2.27	2.30	2.47	2.68	2.52	2.71	1.40	3.57	2.47	4.47	3.64	2.38	3.14	2.96	2.89	4.16	4.16
128	3.27	3.00	3.50	2.51	2.09	2.09	4.61	4.50	3.94	4.76	4.80	5.19	2.24	4.11	5.01	4.24	4.68	5.51
256	4.52	3.03	3.43	3.57	3.83	3.29	8.89	8.31	8.31	9.89	7.89	7.89	3.46	4.21	6.16	5.21	6.06	6.16

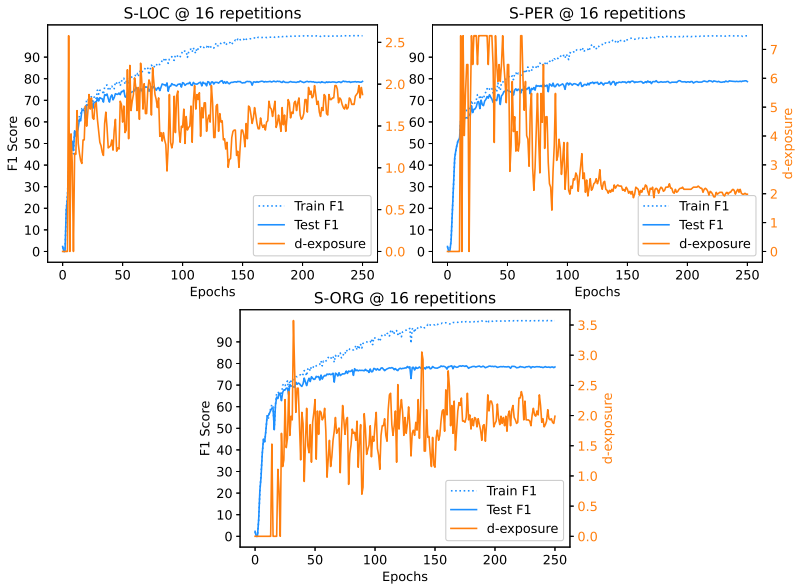

Fig. 2. d-exposure vs. overfitting for S-LOC, S-PER, and S-ORG on CoNLL-2003.

5 Conclusion and Future Work

In this work, we presented a measure of unintended memorization in discriminative neural models. It is inspired by previous work on generative sequence models but offers an approach for tasks where measuring perplexity is not feasible. The core idea is to identify the exposure of potentially private data with confidence assessments of model predictions. We show how ambiguous sentences can be employed towards that goal in a named entity recognition task. One limitation of this methodology is that it can only be applied to NER classes that share some linguistic materials with at least one other class.

We performed a number of in-depth experiments to illustrate the effectiveness of our new metric for assessing model memorization. While we focus on one task here, with a reduced number of NE labels, we are nevertheless able to confirm the findings of the previous work on exposure for generative sequence models. In particular, these are 1) higher d-exposure values for repeated insertions of a test phrase into the training data; and 2) independence of d-exposure from model overfitting. The first finding confirms that the number of occurrence of a phrase in the training data, the expected memorization of that phrase in the model, and the proposed metric all correlate positively. The second finding sets our approach apart from methods such as membership inference attacks which are prone to significant performance drops for overfitted models.

In the future, we plan to perform similar validation experiments for other natural language processing tasks as well. The definition of what constitutes an "ambiguous phrase" for each task poses a challenge but is a necessary step in the proposed methodology. For the NER task addressed here, a number of additional experiments are conceivable as well, e.g., going beyond single-word entities.

With a powerful metric now in place, an even more interesting future step will be the exploration of principled ways in which counter-measures for model memorization could be realized. Ultimately, assessing a potential information leakage is only the first step, supporting the prevention or confinement of such leakages must be the goal to aspire to.

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., Lacoste-Julien, S.: A closer look at memorization in deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 233–242. ICML'17, JMLR.org (2017)
2. Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., Song, D.: The secret sharer: Measuring unintended neural network memorization & extracting secrets. In: Proceedings of the 28th USENIX Security Symposium. pp. 267–284. USENIX Association, Santa Clara, CA, USA (August 14–16, 2019)
3. Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: Targer: Neural argument mining at your fingertips. In: Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics (ACL'2019). Florence, Italy (2019)
4. Hesamifard, E., Takabi, H., Ghasemi, M., Wright, R.: Privacy-preserving machine learning as a service. Proceedings on Privacy Enhancing Technologies **3**, 123–142 (06 2018). https://doi.org/10.1515/popets-2018-0024
5. Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. In: NIPS Workshop on Private Multi-Party Machine Learning (2016)
6. Long, Y., Bindschaedler, V., Gunter, C.A.: Towards measuring membership privacy (2017), http://arxiv.org/abs/1712.09136
7. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. p. 587–601. CCS '17, Association for Computing Machinery, New York, NY, USA (2017)
8. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 142–147 (2003)
9. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing (Early Access) (05 February, 2019)
10. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017)