OVERLEARNING REVEALS SENSITIVE ATTRIBUTES SUMMARY NOTE

Mingzhe Du
dumingzhex@gmail.com

1 OVERVIEW

Overlearning means that a model implicitly learns more attributes than expected, which includes two main aspects: 1) outside of the learning objective, and 2) sensitive from a privacy or bias perspective.

This paper shows overlearning in several NLP models and their harmful consequences. In terms of harmful consequences, the author introduces two significant situations: 1) the model's inference-time representation of an input reveals the input's sensitive attributes. 2) a model trained for benign tasks can be transferred (re-purposed) into a model for privacy-violating tasks.

Firstly, The authors point out that overlearning may be intrinsic for some tasks. They show if attributes are censored, the censored models either fail to learn, or still leak sensitive information. Moreover, they exhibit that their de-consoring model is able to retrieve sensitive attributes from overlearned representations, even if these attributes are not present in the training data. Finally, the authors provide an explanation of where and why overlearning happens based on the complexity of the training data.

2 BACKGROUND

2.1 SUPERVISED DEEP LEARNING

This paper defines supervised deep learning as three parts: an input x, a model M, and the target y. In terms of model M, they represent the model $M = C \circ E$, where C is a classifier and E is a feature extractor (i.e. encoder). The representation $z = E(x)$ is passed to C to predict by modeling $p(y|z) = C(z)$.

2.2 MODEL PARTITIONING

Model partitioning splits the model into a local part and a remote part. For privacy, the local part computes a representation, and sends the censored representation to the remote part.

2.3 CENSORING REPRESENTATIONS

The goal is to encode input into a representation z that does not reveal unwanted properties of x, yet is expressive enough to predict the task label y. There are two methods the author introduced which are adversarial training and information-theoretical objective.

2.3.1 ADVERSARIAL TRAINING

Adversarial training involves a mini-max game between a discriminator D trying to infer s from z during training and an encoder and classifier trying to infer the task label y while minimizing the discriminator's success.

$$\min_{E,\,C} \max_{D} E_{x,\,y,\,s} \left[\gamma \log p(s|z = E(x)) - \log p(y|z = E(x))\right]$$

2.3.2 INFORMATION-THEORETICAL OBJECTIVE

Information-theoretical objective tackles this task from another view that transforms the censoring representation into computing the independence between the generated representation z and the censoring attributes s. Since independence is intractable to measure in practice, thus the requirement is relaxed to a constraint on the mutual information between z and s.

$$max\ I(z,y)\ -\ \beta\ I(z,x)\ -\ \lambda\ I(z,s)$$

This objective has an analytical lower bound:

$$E_{x,\ s}[E_{z,\ y}[log\ p(y|z)] - (\beta+\lambda)KL[q(z|x)\ ||\ q(z)] - \lambda E_z[log\ p(x|z,s)]]$$

3 EXPLOITING OVERLEARNING

This paper demonstrates two different ways to exploit overlearning in a trained model M. The inference-time attack applies M to an input and uses M 's representation of that input to predict its sensitive attributes. The model-repurposing attack uses M to create another model that, when applied to an input, directly predicts its sensitive attributes.

3.1 INFERENCE-TIME ATTACK

Assume an adversary can observe the censored representation z but cannot access the original x directly, this attack tries to infer sensitive attributes from z, which are not part of expected outputs. For the very simple models without censoring, the adversary can easily retrieve sensitive information by an inference-time attack. To complete this attack, the adversary firstly needs to prepare a D_{aux} of labeled (x, s), which does not need to have the same distribution with the training data of the target model, and have access to the target model to compute the corresponding $E(z)$. Secondly, training an attack model M_{attack} on the $< E(x),\ s >$ pairs. At the inference time, the adversary finally predict s* from the observed z on M_{attack}(z is what an output from the target model as well).

The next step, this paper considers a more complicated scenario: the representation is censored to reduce the leakage of sensitive information. They trained an adversarial transformer to recover the erased features from $T(z)$ as z^*. Subsequently, the adversary can use $T(z)$ as an uncensored approximation of z to train an inference model M_{attack} and infer s* on M_{attack} as same as the inference-time attack.

Inferring s from representation
Input: Auxiliary dataset D_{aux}, Black-box oracle E, observed z^*.
Construct dataset: $D_{attack} \leftarrow \{(E(x),\ s) \mid (x,\ s)\ \in D_{aux}\}$.
Training: Train attack model M_{attack} on D_{attack}.
Inference: Prediction $s^* = M_{attack}\ (z^*)$.

Table 1: Inferring sensitive data from representation

Adversarial re-purposing
Input: Model M from the original task, transfer dataset $D_{transfer}$ for the new task.
Construct model: $M_{transfer} = C_{transfer} \circ E_l$
Fine tune: $M_{transfer}$ on $D_{transfer}$
Return: Transfer model $M_{transfer}$

Table 2: Adversarial re-purposing

3.2 MODEL-REPURPOSING ATTACK

This attack is based on the idea that converting a model trained for a benign task into a model that predicts sensitive attributes. Conceptually, the adversary is able to use the representation of any layer for this purpose, but the deeper layers may drop more primary information. The transferred

model $M_{transfer} = C_{transfer} \circ E_l$ is fine-tuned on another, small dataset $D_{transfer}$. Through this method, the authors conjecture that learning per se cannot be a regulated purpose of data collection.

De-censoring algorithm
Input: Auxiliary dataset D_{aux}, Black-box oracle E, observed z^*.
Train $M_{aux} = E_{aux} \circ C_{aux}$ on D_{aux}
Initialize transform model T, inference attack model M_{attack}.
for each training iteration do:
Sample a batch of data (x, s) from D_{aux} and compute $z_{aux} = E(x),\ z_{aux} = E_{aux}(x)$
Update T on the batch of $(z,\ z_{aux})$ with loss $\| T(z) - z_{aux}\|_2^2$
Update M_{attack} on the batch of $(T(z),\ s)$ with cross-entropy loss.
end for
return prediction $s^* = M_{attack}(T(z^*))$

Table 3: De-censoring algorithm

4 CONCLUSION

This paper shows that models trained for simple tasks implicitly learn privacy-sensitive concepts unrelated to the labels of the original task, and overlearning is a generic problem in (fully trained) models. Neither mentioned approaches can prevent overlearning, except at the cost of destroying the model's accuracy. Furthermore, attacks is able to recognize sensitive attributes which even not represented in the training data. Authors conjecture that structural complexity of the data is a potential reason for overlearning.

5 IMPROVEMENT

Inspired by the variational information bottleneck, we can add an extra target in adversarial training that aims to forget the original representation from the censored one. We observe that in this paper, the information-theoretical objective has better results than adversarial training, probably because it not only erases sensitive information from the original representation, but also forgets the original representation as much as possible, which is particularly effective for structural information such as images. Therefore, the first follow up point is that a novel adversarial training model:

$$\min_{E,\,C}\ \max_{D}\ E_{x,\,y,\,s}\left[\gamma\ log\ p(s|z = E(x))\ + \beta\ log\ p(x|z = E(x))\ -\ log\ p(y|z = E(x))\right]$$

To measure the similarity of each layer's representation, the authors used the identical model structure of original tasks for retrieving sensitive information. That is a reasonable operation to explore the relationship of two different models, or when and where overlearning will happen in other words, but our final target is how to attack/protect information from a machine-learning-as-a-service. Hence, the second improvement is to treat the target model as a totally black-box, which means we know nothing inside. Based on the black-box, we can train a sensitive inference model. Conceptually, it may reach a higher accuracy since the model architecture is not fixed.

From this paper, authors conjecture that structural complexity of the data is a reason for model overlearning. So model distillation might help to gain a simple student model, avoiding the leakage of privacy. The third idea is that to distill or compress trained models against the inference attack.

From this paper, authors conjecture that structural complexity of the data is a reason for model overlearning. So model distillation might help to gain a simple student model, avoiding the leakage of privacy. The third idea is that to distillate or compress trained models against the inference attack.

Moreover, I think differential privacy may help to prevent from inversion attacks. We could train a bunch of simple models to bewilder the membership of training data. Finally, since VAE can be used for feature censoring, so can GAN naturally.

Table 4: Original tasks v.s. Sensitive transferring tasks.

Dataset	Acc of predicting target y				Acc of inferring sensitive attribute s			
	RAND	BASE	ADV	IT	RAND	BASE	ADV	IT
Health	66.31	84.33	80.16	82.63	16.00	32.52	32.00	26.60
UTKFace	52.27	90.38	90.15	88.15	42.52	62.18	53.28	53.30
FaceScrub	53.53	98.77	97.90	97.66	1.42	33.65	30.23	10.61
Places365	56.16	91.41	90.84	89.82	1.37	31.03	12.56	2.29
Twitter	45.17	76.22	57.97	n/a	6.93	38.46	34.27	n/a
Yelp	42.56	57.81	56.79	n/a	15.88	33.09	27.32	n/a
PIPA	7.67	77.34	52.02	29.64	68.50	87.95	69.96	82.02

Table 5: Improving inference accuracy with de-censoring. δ is the increase from Table 4.

Dataset	Health	UTKFace	FaceScrub	Places365	Twitter	Yelp
ADV $+\delta$	32.55 +0.55	59.38 +6.10	40.37 +12.24	19.71 +7.15	36.55 +2.22	31.36 +4.04
IT $+\delta$	27.05 +0.45	54.31 +1.01	16.40 +5.79	3.10 +0.81	n/a	n/a

REFERENCES

A EXPERIMENTS

In this section, authors utilize a couple of datasets to verify their inferences. To briefly summarize, the introduction of each dataset and corresponding model architectures will not be presented here.

A.1 EXPERIMENTS 1: ORIGINAL TASKS V.S. SENSITIVE TRANSFERRING TASKS

Table 4 shows that the accuracy of original tasks and of inferring sensitive data on different datasets by base, adversarial training and Information-theoretical censoring respectively. The accuracy of inference from the last-layer representations is much higher than random guessing for all tasks when representations are not censored. From that the author obtains a point that models over-learn even in the higher, task-specific layers. Meanwhile, information-theoretical censoring has a better performance on defending inferences, but also damages the accuracy of original tasks more than adversarial training for almost all models.

A.2 EXPERIMENTS 2: GENDER CLASSIFICATION IN SHORT.

Authors introduce an experiment about gender classification in short. Point out that overlearning can cause a model to recognize even the sensitive attributes that are not represented in the training dataset.

A.3 EXPERIMENTS 3: EFFECT OF CENSORING STRENGTH

Figure 1 shows that stronger censoring does not help. Over-stronger censoring cannot effectively prevent inference attacks, but fall down the accuracy of main tasks.

A.4 EXPERIMENTS 4: DE-CENSORING ALGORITHM

Table 5 shows that de-censoring significantly boosts the accuracy of inference from representations censored with adversarial training. The boost is smaller against information-theoretical censoring because its objective not only censors z with $I(z,\ s)$, but also forgets x with $I(x,\ z)$.

Figure 1: Effect of censoring strength.

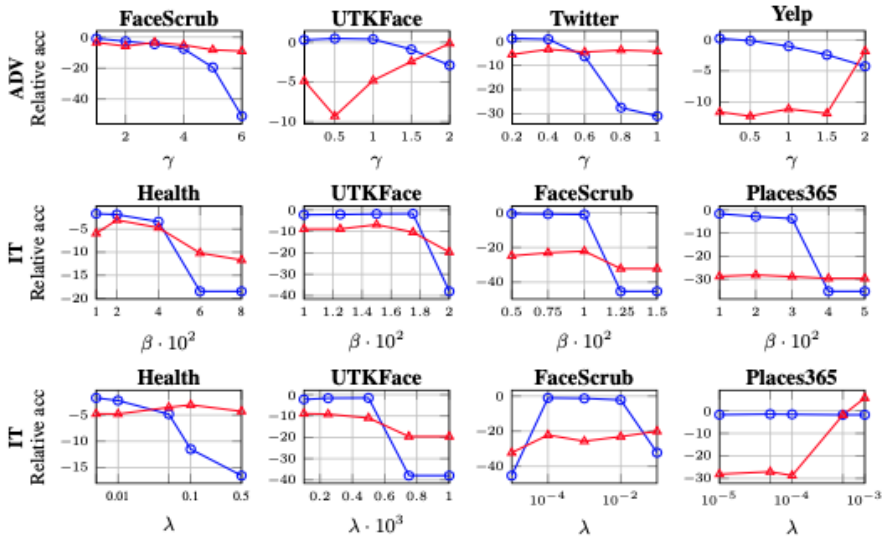

Table 6: Adversarial re-purposing.

$\lvert\mathcal{D}_{\text{transfer}}\rvert/\lvert\mathcal{D}\rvert$	Health	UTKFace	FaceScrub	Places365	Twitter	Yelp	PIPA
0.02	-0.57	4.72	7.01	4.42	12.99	5.57	1.33
0.04	0.22	2.70	15.07	2.14	10.87	3.60	2.41
0.06	-1.21	2.83	7.02	2.06	10.51	8.45	6.50
0.08	-0.99	0.25	11.80	3.39	9.57	0.33	4.93
0.10	0.35	2.24	9.43	2.86	7.30	2.1	5.89

A.5 Experiments 5: Adversarial re-purposing

Table 6 demonstrates that overlearned representations can be picked up by a small set of unseen data to create a model for predicting sensitive attributes.

A.6 Experiments 6: Effect of censoring

Previous work only censored the highest layer of the models. Model re-purposing can use any layer of the model for transfer learning. Therefore, to prevent re-purposing, inner layers must be censored. Table 7 summarizes the results. Censoring lower layers ($conv1$ to $conv3$) blocks adversarial re-purposing, at the cost of reducing the model's accuracy on its original task.

A.7 Experiments 7: When and where overlearning happens

In the last section, the authors discuss when, where and why overlearning happens. Figure 2 shows that lower layers of models A and B learn very similar features. Intuitively, there is little similarity between the low-level features of A and the high-level features of B and vice versa, which is consistent with intuition.

Table 7: Effect of censoring.

Censored on $\gamma = 0.5$	δ_A	δ_B when transferred from conv1	conv2	conv3	fc4	fc5
conv1	-1.66	-6.42	-4.09	-1.65	0.46	-3.87
conv2	-2.87	0.95	-1.77	-2.88	-1.53	-2.22
conv3	-0.64	1.49	1.49	0.67	-0.48	-1.38
fc4	-0.16	2.03	5.16	6.73	6.12	0.54
fc5	0.05	1.52	4.53	7.42	6.14	4.53
$\gamma = 0.75$						
conv1	-4.48	-7.33	-5.01	-1.51	-7.99	-7.82
conv2	-6.02	0.44	-7.04	-5.46	-5.94	-5.82
conv3	-1.90	1.32	1.37	1.88	0.74	-0.67
fc4	0.01	3.65	4.56	5.11	4.44	0.91
fc5	-0.74	1.54	3.61	6.75	7.18	4.99
$\gamma = 1$						
conv1	-45.25	-7.36	-3.93	-2.75	-4.37	-2.91
conv2	-20.30	-3.28	-5.27	-7.03	-6.38	-5.54
conv3	-45.20	-2.13	-3.06	-4.48	-4.05	-5.18
fc4	-0.52	1.73	5.19	4.80	5.83	1.84
fc5	-0.86	1.56	3.55	5.59	5.14	1.97

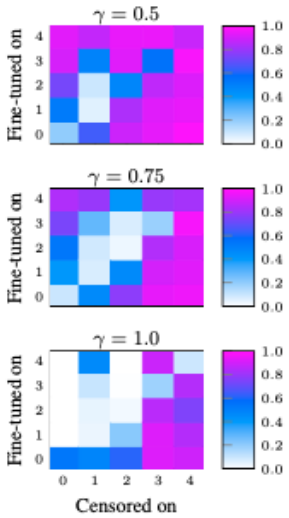

Figure 2: Pairwise similarities of layer representations between models for the original task (A) and for predicting a sensitive attribute (B). Numbers 0 through 4 denote layers conv1, conv2, conv3, fc4 and fc5.