Model Extraction Attacks on DistilBERT

Amro Salman & Ayman Saeed
Department of Electrical and Electronics Engineering
University of Khartoum
Khartoum, Sudan
{AmroAdil420,Ayman.saeed.moh}@gmail.com

Khalid N. Elmadani
Department of Computer Science
University of Cape Town
Cape Town, South Africa
ahmkha009@myuct.ac.za

Sharief Babiker
Department of Electrical and Electronics Engineering
University of Khartoum
Khartoum, Sudan
sfbabikir@uofk.edu

Abstract

This paper investigates model extraction attacks, where an adversary can train a substitute model by collecting data through query access to a victim model and steal its functionality. We use DistilBERT as the victim model due to its compact size and faster processing speed compared to BERT. The results demonstrate the effectiveness of the model extraction attack and show that there is a relation between its success and the similarity between the training data of the victim model and the attacker's queries. The study provides important insights into the security of DistilBERT models and inform the development of more robust defense mechanisms against such attacks.

1 Introduction

Model Extraction Attacks have emerged as a pressing concern in the landscape of machine learning security (Krishna et al., 2019; Keskar et al., 2020; He et al., 2021). As shown in figure 1, these attacks manifest when an adversary, through query access, gathers data from a victim model. This data is then used to train a parallel, substitute model, effectively replicating and potentially misusing the functionality of the original.
DistilBERT (Sanh et al., 2019) is a small, fast, cheap, and light Transformer model (Vaswani et al., 2017) based on the BERT architecture Devlin et al. (2018). We use DistilBERT as the victim model because it is smaller and faster than the `BERT-base` model, and retains most of its functionality. Hence, DistilBERT is cheaper to deploy, and using it as a victim model will be more realistic.

2 Background

Our investigations are based on earlier works that discuss several extraction attacks. Krishna et al. (2019) studies model extraction in NLP using an adversary who reconstructs a local copy of a victim model with query access. They extract models on four NLP tasks using query generators and task-specific heuristics, finding they are accurate even when trained with nonsensical inputs. Keskar et al. (2020) took a step further by demonstrating the creation of multilingual models by extracting knowledge from a monolingual model. He et al. (2021) proposed a two-stage attack is proposed on BERT-based APIs, which depends on the domain closeness between the victim's data and the attacker's queries. The extracted model crafts adversarial examples that can be transferred to the victim model.

5th Deep Learning Indaba Conference (DLI 2023).

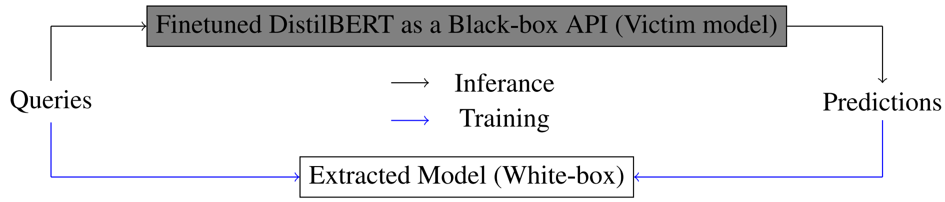

Figure 1: Model Extraction Attack: We pass queries to the victim model and use its predictions with the queries to train the extracted model.

3 Methodology

For this study, we employ extraction attacks by training two DistilBERT victim models. One of these models is trained on the Ag News dataset (Zhang et al., 2015), while the other is trained on the TrustPilot Reviews dataset (Hovy et al., 2015).
We conduct experiments under two scenarios. In the first scenario, we assume the attacker knows that the victim model is a DistilBERT model, so we train three extracted models based on DistilBERT using a different dataset for each one of them. The first dataset is the same dataset that is used to train the victim model with the original labels replaced by the predictions of the victim model ($D_A = D_V$) and the second one is a dataset from the same domain with the same size as the training dataset with labels extracted from the victim model ($D_A \neq D_V$, 1x), and the last one is a dataset from the same domain with a size equal to five times the size of the training dataset with labels extracted from the victim model ($D_A \neq D_V$, 5x).
In the second scenario, we assume the attacker does not know the victim model, so we test four models with different sizes, `BERT-base`, `BERT-small` (Turc et al., 2019), `TinyBert` (Jiao et al., 2019), and a model with the same architecture as `DistilBERT` but trained from scratch (`BERT-base` $>$ `DistilBERT` $>$ `BERT-small` $>$ `TinyBert`).
It should be noted that we use the Yelp polarity (Zhang et al., 2015) to extract the model trained on the reviews' dataset and Yahoo Answers (Zhang & Lecun, 2015) to extract the model trained on the news' dataset.

4 experimental setup

4.1 Datasets

Table 1 shows the datasets we used. More information about the datasets is provided below.
Trustpilot Sentiment dataset (Hovy et al., 2015): It contains reviews associated with a sentiment score on a five-point scale. In this paper, we only use two scores, score one as a negative review and score five as a positive review.
Yelp Polarity dataset (Zhang et al., 2015): Yelp dataset is a document-level sentiment classification dataset. The original dataset is on a five-point scale, while the polarised version assigns negative labels to the rating of 1 and 2 and assigns positive labels to 4 and 5.
AG News dataset (Zhang et al., 2015): AG news corpus is a dataset that's mainly used for topic classification task which is to predict the topic label of the document. It has four different topics in total. We only used three topics: “Sports”, “Business” and "Sci/Tech" for our research purposes.
Yahoo Answers Topics dataset (Zhang & Lecun, 2015):Yahoo answers topics covers 10 different topics.however, We used only three topics of them for our studies: “sports”, “Business & Finance” and "Science & Mathematics"
Trustpilot and **AG News** are used for training the victim and for extraction. While **Yelp Polarity** and **Yahoo Answers** are used for extraction only.

4.2 Models

Our research study focuses on developing a victim model that is based on fine-tuning a DistilBERT model Sanh et al. (2019) for a sequence classification task.
To extract information from the victim model, we employed two different approaches:
The first approach involved training the attacker multiple times by fine-tuning several models

Table 1: Statistic of sentiment analysis and topic classification datasets.

Dataset	Train	Validation	Test	Task
Trustpilot	22,142	2,767	2,767	sentiment analysis
Yelp Polarity	520K	-	-	sentiment analysis
AG News	7,100	887	887	topic classification
Yahoo Answers	590K	-	-	topic classification

Table 2: Accuracy and Loyalty of the victim models and the extracted models among different datasets. #Q is the number of queries.

Model	**#Q**	**AG News**		**Trustpilot reviews**	
		Accuracy	Loyalty	Accuracy	Loyalty
Victim		93.91%	-	88.00%	-
DistilBERT					
$D_A = D_V$		94.58%	98.64%	88.20%	95.20%
$D_A \neq D_V$	1x	89.96%	90.64%	87.60%	91.40%
$D_A \neq D_V$	5x	91.88%	92.21%	87.97%	93.40%
BERT-base					
$D_A = D_V$		94.92%	98.08%	88.70%	94.20%
$D_A \neq D_V$	1x	90.64%	88.27%	87.10%	91.30%
BERT-small					
$D_A = D_V$		93.01%	97.74%	87.50%	93.30%
$D_A \neq D_V$	1x	88.95%	86.69%	85.00%	89.50%
TinyBERT					
$D_A = D_V$		93.46%	95.94%	88.60%	93.70%
$D_A \neq D_V$	1x	85.56%	85.79%	86.40%	90.30%
DistilBERT from scratch					
$D_A = D_V$		92.44%	93.79%	87.70%	91.60%
$D_A \neq D_V$	1x	66.62%	79.93%	80.40%	82.80%

including DistilBERT Sanh et al. (2019), BERT-base Devlin et al. (2018), BERT-small Turc et al. (2019), and TinyBERT Jiao et al. (2019). This approach allowed us to observe the behavior of the attacker model when it is pretrained and have a prior understanding of the language.
The second approach we employed was to train a DistilBERT model from scratch using the extracted dataset. This approach helped us to gain insights into the performance of a new model on the same task and evaluate its effectiveness in performing attacks

4.3 Hyper-parameters

We finetune all models for 4 epochs with a learning rate of $5e-5$, and a batch size of 16 example per batch. This study employed Huggingface libraries[1] for training purposes.

5 Resutls

Table 2 shows the results. We report the accuracy and loyalty of each model. Loyalty is the accuracy while using victim predictions as reference labels, it represents a measure of how successful is our extracted model in mimicking the behavior of the victim model. The Table contains multiple blocks. The victim block represents the results of the victim model, while each block from the rest of the blocks represents the results of the corresponding model as the extracted model.
In all our models used for extraction, we can see that the accuracy and the loyalty of the extracted model are higher when $D_A = D_V$, which indicates that the success of model extraction depends on the similarity between the training data of the victim model and the attacker's queries.
Additionally, we can see that the extracted model that is based on the same architecture of the victim model (DistilBERT) outperforms the victim model when $D_A = D_V$, also known as self-distillation

[1]https://huggingface.co/

Furlanello et al. (2018). Overall, for pre-trained extracted models, we observe that more models' parameters lead to higher accuracy. The only exceptions are `BERT-small` and `TinyBERT`. Although `TinyBERT` has fewer parameters, it extracts the victim better than `BERT-small`. We argue that this is because `TinyBERT` was trained with Knowledge Distillation, similar to the victim (DistilBERT).
Also, the DistilBERT extracted models trained with 5 times the size of the training dataset (5x) outperforms the ones trained with the same size of the training dataset (1x), indicating that increasing the number of queries leads to improved extraction performance.
Finally, The extracted models based on pre-trained DistilBERT have higher accuracy and loyalty compared to the corresponding extracted models that are trained from scratch, emphasizing the importance of using pre-trained models.

6 Conclusion

Our study provides a comprehensive analysis of extraction attacks on DistilBERT models, revealing potential vulnerabilities associated with these models. Furthermore, The results obtained have significant implications for the security of DistilBERT models and inform the development of more robust defense mechanisms against such attacks. Overall, our study contributes to the growing body of research on adversarial attacks in Natural Language Processing and underscores the importance of continued investigation into model security.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pp. 1607–1616. PMLR, 2018.

Xuanli He, Lingjuan Lyu, Qiongkai Xu, and Lichao Sun. Model extraction and adversarial transferability, your bert is vulnerable! *arXiv preprint arXiv:2103.10013*, 2021.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of WWW, pages 452–461.*, 2015.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. The thieves on sesame street are polyglots-extracting multilingual models from monolingual apis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6203–6207, 2020.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366*, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Xiang Zhang and Yann Lecun. Text understanding from scratch. 02 2015.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*, September 2015.