Supplementary material for ACTIVETHIEF: Model Extraction Using Active Learning and Unannotated Public Data

Soham Pal[*],[1] **Yash Gupta**[*],[1] **Aditya Shukla**[*],[1] **Aditya Kanade**[†],[1,2]
Shirish Shevade[†],[1] **Vinod Ganapathy**[†1]
[*] [†] [1]Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India
[2]Google Brain, USA
{sohampal,yashgupta,adityashukla,kanade,shirish,vg}@iisc.ac.in

This document includes supplementary material for the paper "ACTIVETHIEF: Model Extraction Using Active Learning and Unannotated Public Data", to appear in the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20).

Details of secret datasets

The details of image and text datasets are given in Tables 8 and 9 below. 20% of the training dataset was set aside as the validation set, wherever a validation set was unavailable. For the text datasets, we truncate the vocabulary to the 5K most frequent words in the dataset. Out of vocabulary words are replaced with a `OOV` token, and sentences are prepended with a `START` token. All sequences are padded to a maximum length of 300 with a special `PAD` token.

Table 8: Details of image datasets

	MNIST	CIFAR-10	GTSRB
Resolution	28x28	32x32	32x32
Channels	1	3	3
# Train samples	48K	40K	31K
# Valid samples	12K	10K	8K
# Test samples	10K	10K	12K
# Classes	10	10	43

Table 9: Details of text datasets

	MR	IMDB	AG News
Dictionary	5K	5K	5K
# Train samples	7.7K	20K	96K
# Valid samples	1.9K	5K	24K
# Test samples	1K	25K	7.6K
# Classes	2	2	5
Mean sequence length	20	237	32

[*]All three authors contributed equally.
[†]All three authors contributed equally.

Training regime for the substitute model of Papernot et al. (2017)

We use a value of $\lambda = 0.1$, as recommended in the paper. The number of initial samples and augmentation steps is adjusted, as shown in Table 10. This is done to ensure that the number of queries made to the secret model is 9.6K (for MNIST and CIFAR-10) and 13.76K (for GTSRB), allowing for a fair comparison to ACTIVETHIEF, where we use a total budget of 10K (20% of the queries being used for validation, and the remaining 80% being used for training).

Table 10: Hyperparameters used for the training of the substitute model of Papernot et al. (2017)

	MNIST, CIFAR-10	GTSRB
λ	0.1	0.1
Augmentation steps	7	6
Initial samples per class	15	10
Total queries made	9.6K	13.76K

The Shapiro-Wilk test

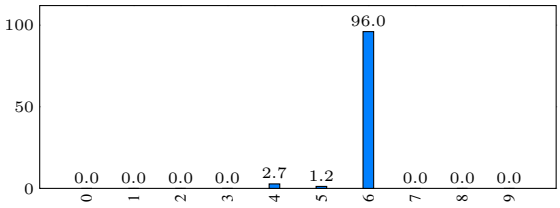
Consider a situation in which a client makes n queries $x_1, x_2, \ldots x_n$, which are subsequently classified as belonging to the classes $y_1, y_2, \ldots y_n$. The minimum distance values are computed as: $d_i = \min_{j<i, y_j=y_i} \|x_i - x_j\|_2$, where d_i is vacuously set to 0 where required. Using these distance values, the following test statistic is computed:

$$W(D) = \frac{(\sum_{i=1}^{n} a_i d_{(i)})^2}{\sum_{i=1}^{n}(d_i - \bar{d})^2}$$

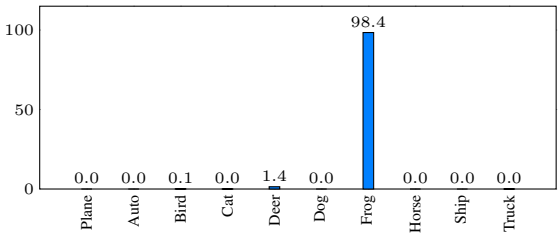
where $D = \{d_i\}_{i=1}^{n}$, and $d_{(i)}$ refers to the i^{th} order statistic of D, and the values of a_i are functions of the i^{th} expected order statistics of i.i.d. normally distributed random variables. When $W(D) < \delta$, PRADA rejects the null hypothesis and claims that an attack has been detected. For our experiments, we use a value of $\delta = 0.9$.

Distribution of labels predicted by the secret model for uniform noise samples

We generate data by sampling from a multidimensional version of the $U[0, 1]$ uniform distribution. Note that this is a SNPD dataset, and corresponds to the simple equation-solving attack of Tramèr et al. (2016). The outputs produced by the secret model are recorded, and then averaged. The frequency of the resulting values are reported for the MNIST and CIFAR-10 datasets in Figure 5. We observe a similar distribution for the GTSRB dataset, but we omit the corresponding results as there are 43 classes in the dataset. We speculate that due to the lack of samples from certain classes (Digits 0-3, and 7-9 in MNIST), the secret model is unable to classify them correctly, leading to poor agreement.



(a) MNIST



(b) CIFAR-10

Figure 5: The distribution of labels (frequency in %) assigned by the secret model to uniform noise (SNPD) input.

References

Papernot, N.; McDaniel, P. D.; Goodfellow, I. J.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *AsiaCCS*.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction APIs. In *USENIX Security Symposium*.