Adversarial Attacks Using Model Stealing

Using Active Learning to Steal Target Models

Author: Steffano Psathas - Contact: s.g.psathas@student.tudelft.nl - Supervisors: Stefanie Roos, Jihue Huang, Chi Hong

Background

- **Adversarial attacks:** add an perturbation to an image to make a classifier misclassify the image.
- **Model stealing:** Using a substitute model to function like the target model.
- Model stealing requires a lot of real-world examples

figure 1: Example of a base image (left) that gets classified correctly and an altered image (right) which gets misclassified

Research Question

- Assuming that there is an unlabeled dataset from real-world examples, how can you use a subset of the dataset to allow for less target model querying during model stealing while maintaining the accuracy of the substitution model?

Methodology

- **Active Learning**

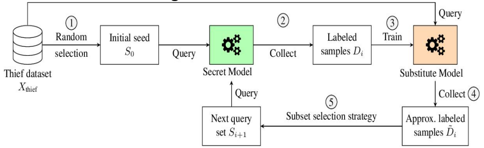

figure 2: ActiveThief algorithm depicted in a diagram

- **Seeding Strategy:** Cluster seeding
- **Subset Selection Strategy:** Uncertainty, clustering, uniform, and combinations
- **Stopping Criterion:** Uncertainty-based stopping

The evaluation and the results

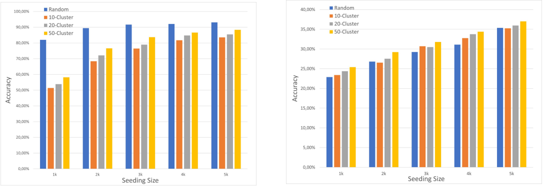

figure 3: Random seeding versus cluster seeding using different number of clusters with different seeding size on MNIST dataset (left) and FashionMNIST dataset (right).

Seeding strategy:

- **MNIST:** Random Seeding performs best.
- **FashionMNIST:** Cluster seeding performs best.
- **Cluster Seeding**: More cluster increases accuracy.

figure 4: uniform, entropy and clustering strategies versus random sampling strategy on FashionMNIST dataset.

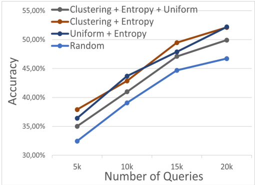

figure 5: combination of several strategies versus random sampling strategy on FashionMNIST dataset.

Subset Selection Strategy:

- Random Seeding gets outperformed by al strategies.
- Combining strategies gives a performance boost
- The combination of entropy, clustering and uniform underperforms compared to other combinations.

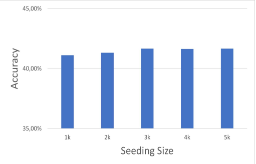

figure 6: Effect of seeding size with random seeding and entropy strategy on FashionMNIST dataset.

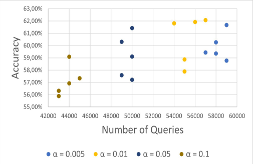

figure 7: Effect of α in the stopping criterion on FashionMNIST dataset. Using the whole dataset results in an accuracy of 66.52%

Seeding Size:

- Having a higher starting seeding size does not influence final accuracy.

Stopping Criterion:

- Higher α leads to earlier convergence, but lower accuracy.
- α-values cluster together on the graph.

References

[1] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples.CoRR, abs/1602.02697, 2016.
[2] S. Pal, Y. Gupta, A. Shukla, A. Kanade, S. Shevade, and V. Ganapathy, “ActiveThief: Model Extraction Using Active Learning and Unannotated Public Data”, *AAAI*, vol. 34, no. 01, pp. 865-872, Apr. 2020.

Discussion

Seeding Strategy:

- Could differ between datasets because cluster size might not be high enough.
- More Clusters increase computation costs

Subset Selection Strategy:

- The combination of entropy, clustering and uniform underperforms because clustering and uniform counteract each other.

Seeding Size:

- Seeding size has almost no influence because the information gain is done with the subset selection strategies
- Some strategies can have impact because of their reliance of earlier used data points.

Stopping Criterion:

- α should be between 0.01 and 0.05 because that gives the best query to accuracy relation.

Conclusion

- ActiveThief is a viable solution to decrease the number of queries needed for model stealing.
- The use of cluster seeding can have impact on the seeding accuracy when using enough clusters.
- Combining subset selection strategies gives a performance boost.
- Higher starting seeding size does not influence final accuracy.
- The stopping criterion can be used to early halt the algorithm.

Future Work

- Effectiveness on different datasets.
- The effectiveness of cluster seeding on the final accuracy.
- More subset selection strategies and combinations can be explored.
- Different uncertainty measures can be used for the stopping criterion.