Can’t Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders

Zeyang Sha1 Xinlei He1 Ning Yu2 Michael Backes1 Yang Zhang1
1 CISPA Helmholtz Center for Information Security 2 Salesforce Research

Motivation
Model stealing attacks aim at stealing the functionality of the target model. So far, model stealing attacks concentrate on the supervised classifiers, i.e., the model responses are prediction posteriors or labels for a specific downstream task. The vulnerability of self-supervised image encoders is unexplored.

Method

Conventional attacks will use MSE loss to optimize the surrogate encoder. The loss function can be formulated as:

\[ L_{MS} = \sum_{k=1}^{N} \left( \sqrt[p]{E_T(x_k)^2 - E_S(x_k)^2} \right) \]

The loss function will make different embeddings of same images closer to each other. In this way, the surrogate model will behave similarly as target model.

Cont-Steal attacks will first generate different views of given images using RandAug. Then, the cont-steal loss will try to enforce the surrogate embedding of an image close to its target embedding (defined as a positive pair) and also push away embeddings of different images irrespective of being generated by the target or the surrogate encoders (defined as negative pairs).

Conclusion

We introduce the first model stealing attacks against image encoders:
- We pioneer the investigation of the vulnerability of unsupervised image encoders against model stealing attacks. We discover that encoders are more vulnerable than classifiers;
- We propose Cont-Steal, the first contrastive learning-based stealing attack against encoders that outperforms the conventional attacks to a large extent; Extensive evaluation shows that the advantageous performance of Cont-Steal is consistently amplified in various settings, especially when the adversary suffers from zero information of the target dataset, limited amount of data, or restricted query budgets.