Attacks on Victim Model! A defense strategy

Abstract:

This paper talks about two defense strategies which could be implemented in the victim model in order to avoid the model being extracted by the adversary. The whole process is explained using the pre trained NLP model which is a BERT model. The attacker need not have any training data for model extraction. Instead, can extract the model just by simply passing a query which has random sequence of words along with task specific approach. Such random sequences of words are used to effectively extract the victim model in varying NLP tasks such as question answering, NLI (Natural Language Inference) etc. Using queries which would cost a few hundred bucks for the adversary to reconstruct the model and performs little worse than the victim model. In this paper, two strategies, membership classification and API watermarking are used to protect the model from attacker. This approach has succeeded only in protecting against the naïve ones not the advanced ones.

The model is extracted using the APIs where the attacker can issue large number of queries. Using the results which has input and output, one can train the local model and over the process the attacker were able to recreate the original model. By doing so sensitive information about the training data could also be extracted which will lead organizations to potential risk. For instance, the recent success in the contextualized representation for transfer learning in the NLP could lead to these kinds of problems. Basically, the adversary can pass the nonsensical random sequence of words to the BERT based victim model and use these outputs from the victim model to fine tune the local BERT model.