Exploratory Reconnaissance of the Machine Learning Security domain

Security of Large Pretrained Language Models in NLP

Long Phan (lnp26)

January 17, 2022

Since the releasing of the famous paper ”Attention is All You Need” [7], there has been a widely adaption of Transformers Architecture in Natural Language Processing (NLP) and Natural Language Understanding (NLU) tasks. The architecture which has encoder and decoder blocks (in which consitses of layers of attention mechanism over word embdding to understand a representation meaning of an input) has been proved to achieve state-of-the-art on many of the NLP/NLU tasks including classification or sequence-to-sequence generation.

Following is the releasing of BERT [2] and other big pretrained encoder models (RoBERTa [4], Electra [1], etc.) These models, leveraging transfer learning, are pre-trained on a large corpus of either genreative domain or specific domain to learn a better presentation of the trained domain before interned on downstream task. For example, BERT was trained on a large corpus of english data and acheive multiple state-of-the-art results on downstream classification task that include natural English text input. On the other hand, BioBERT [3] utilizing BERT checkpoint and continously training on Biomedical data achieves state-of-the-art results on all English Biomedical Data. PhoBERT [5], using the RoBERTa architecture and re-trained from scratch, trained on a large Vietnamese text data and also achieves great results in Vietnamese domain classification tasks.

These impressive results from pretrained models and its application in downstream task has been open a new era for Natural Language Processing in which a lot of large pretrained models are emerging. There also has been promising startup and organization like HuggingFace [1] creating pipeline, datasets and template architecture for the purpose of training large pretrained models in NLP. Pretraining NLP models are now a norm of almost all NLP tasks from Named-Entity-Recognition (NER), Relation Extraction (RE) for more genretive tasks live summarization or Question Answering (QA).

However, the security of these pretrained models are often overlooked because the topic and the fields are still very new, Security of these models can range from vulnerable to failure of system or Personal Identifiable Information. Large Pretrained Model are very heavy and take a lot of resources on any system to run. These problems can be a target for attackers to exhaust the resources of any platform that host these models then therefore create a failure in the system. Effective Pretrained models usually have around more than 100M+ parameters, making the computation require a lot of GPU resources; yet if not effectively engineered, attackers can send request with a very large input or modify the batch size of lower level engineering code so that make the GPU out-of-memory and runtime crashing.

Another security issue is that the data of pretrained models are too large to make sure that there are no Personal Indentity Information (PII) in it. These data are usually from Wikipedia or C4 [6] which are crawling over the internet. PII issue can be simple as a street address or transportation of a person to crucial as an SSN/Passport number. Hackers can exploit these knowledge that a model learned to steal the identity of any specific customer. There has been limited studies in the studies of how much can a pretrained models remember such of the PII knowledge in their parameters.

References

[1] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555, 2020.

[2] 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.

[3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746, 2019.

[1] https://huggingface.co/

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[5] Dat Quoc Nguyen and Anh Tuan Nguyen. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, 2020.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.