The Thieves on Sesame Street are Polyglots — Extracting Multilingual Models from Monolingual APIs

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong & Richard Socher
Salesforce Research
{nkeskar,bmccann,cxiong,rsocher}@salesforce.com

Abstract

Pre-training in natural language processing makes it easier for an adversary with only query access to a victim model to reconstruct a local copy of the victim by training with gibberish input data paired with the victim's labels for that data. We discover that this extraction process extends to local copies initialized from a pre-trained, multilingual model while the victim remains monolingual. The extracted model learns the task from the monolingual victim, but it generalizes far better than the victim to several other languages. This is done without ever showing the multilingual, extracted model a well-formed input in any of the languages for the target task. We also demonstrate that a few real examples can greatly improve performance, and we analyze how these results shed light on how such extraction methods succeed.

1 Introduction

Deploying machine learning models typically involves significant cost, including the expense of data acquisition, data cleaning, and model training and tuning. Recent work by Krishna et al. (2020) has demonstrated that deployed NLP models can be stolen by adversaries by querying victim models with gibberish input data that consists of random sequences of words. In particular, they showed that the following approach is sufficient for stealing text classification and question answering models. First, unlabeled data is created by randomly sampling words from a vocabulary. Second, a deployed API is queried with each random input sequence to obtain a label for each. Third, a pre-trained language model such as BERT (Devlin et al., 2019) is fine-tuned on the victim-labeled gibberish data. The resulting model retains a significant fraction of the victim model's performance without ever seeing a single well-formed input sentence. This process of "stealing" from an API, or "extracting" a local copy of a victim model, is not specific to NLP tasks but rather is a more general phenomenon (Tramèr et al., 2016; Orekondy et al., 2019; Juuti et al., 2019; Milli et al., 2019). Notably, it does not succeed when the extractor model is trained from scratch; a pre-trained model, such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2020), appears to be critical (Krishna et al., 2020).

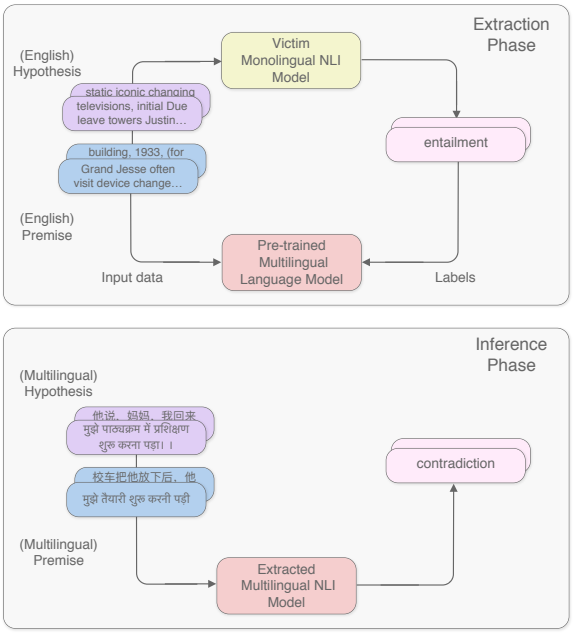

Figure 1: Extraction of multilingual models from monolingual APIs. *(Extraction phase:)* A pre-trained multilingual model is fine-tuned on gibberish data whose labels are queried from a monolingual API. *(Inference phase):* This model is then used for zero-shot cross-lingual transfer on different languages.

The costs for creating and hosting multilingual NLP models can be even greater than for monolingual models. Therefore, extracting a multilingual model is potentially more valuable for an adversary. We demonstrate that it is possible to create

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6203–6207,
November 16–20, 2020.

multilingual models by stealing the task-specific knowledge from a monolingual victim model and extracting it into a new model pre-trained for multilingual language modeling, such as mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2019). These models are similar to BERT and RoBERTa discussed above but extend pre-training and fine-tuning to multiple languages. Even when fine-tuned in one language, say English, these models achieve good zero-shot performance in other pre-training languages. This phenomenon, known as zero-shot cross-lingual transfer, forms the basis of our approach. Combining it with model stealing, or extracting, we demonstrate cross-lingual transfer of task-specific knowledge stolen from a monolingual victim model without collecting a single grammatically correct sentence in any language.

Our investigation has ramifications for the discussion of model APIs as intellectual property and motivates the need to build defenses against such attacks. Since models could be deployed by adversaries in multiple languages without collecting real examples in any, defenses such as watermarking (Szyller et al., 2019) would be rendered useless.

While the reason for the surprising phenomenon is unknown, it is hypothesized (Krishna et al., 2020) that the dynamics of extraction is similar to that of model distillation (Hinton et al., 2015). During model distillation, a (student) model is trained with labels as the outputs of another (teacher) model rather than the ground truth to achieve similar or better performance than the teacher (Furlanello et al., 2018). The success of distillation could thus help explain that of model extraction.

For this short paper, we consider the problem of natural language inference on the multilingual XNLI dataset (Conneau et al., 2018) and show that:

1. Using labels obtained from an English model queried with gibberish English data, a multilingual model can be trained to a high performance on the English task and obtain good zero-shot performance on several other languages.

2. By additionally fine-tuning on 5% of the original English data, we can significantly boost performance on all languages. This post-hoc fine-tuning performs better than mixing the real and gibberish data during extraction.

3. The vocabulary used for generating gibberish data greatly impacts performance. The inability of pre-trained language models to distinguish real and gibberish examples is potential, partial explanation for the success of model extraction.

2 Methodology

We study the problem of natural language inference (NLI): classifying the relationship between a pair of sentences (premise and hypothesis) as either entailment, contradiction or neutral. We focus on this problem given the availability of data in several languages and a history of results on the benchmark (Conneau et al., 2018). We consider the setting where an NLI classification model is available as a black-box. It can be queried with any input data and returns hard labels. Consistent with earlier work, we call this model the *victim* model. We consider a separate model, the *extractor* model, that is trained by extracting task-specific knowledge from the victim model. We aim to study how multilingual pre-training affects the extractor and show that it allows transfer of task-specific knowledge from the victim model to other pre-trained languages. We consider two instances in our experiments: (i) one where the extractor has access to no real data and only queries gibberish, and (ii) extractor has access to some real data in English. Here, we refer to *real data* as data which was also used to train the victim model.

In our experiments, the victim model is trained on the MNLI dataset (Williams et al., 2018). We perform all cross-lingual experiments on the XNLI benchmark (Conneau et al., 2018). This benchmark contains NLI instances in several languages whose test sets were translated by humans using the MNLI dataset. In order to generate gibberish input data, we follow the approach of Krishna et al. (2020). For the hypothesis, we generate sentences of random length by sampling words uniformly from the word-level vocabulary of WikiText-103. The length of the sentence is sampled based on the distribution of lengths in WikiText-103. For the premise, we randomly swap three words of the hypothesis for random words leaving the rest identical. This is to mimic common NLI inputs which have several overlapping words in the hypothesis and premise. For all input sentences, we then perform inference using the victim model and use the hard labels as ground truth labels for the gibberish input data. The gibberish dataset is generated to be the same size as the MNLI dataset ($\sim$392k examples). When training the extractor, we tune the learning rate and maximum iterations, and we use the HuggingFace

Victim	Extractor	Accuracy
RoBERTa-Large	—	90.6
BERT-Large[1]	BERT-Large	76.3
RoBERTa-Large	RoBERTa-Base	74.8
RoBERTa-Large	XLMR-Base	69.0
RoBERTa-Large	RoBERTa-Large	84.3
RoBERTa-Large	XLMR-Large	78.6

Table 1: Development set accuracy on MNLI of various extractor models. [1] Result from Krishna et al. (2020).

library (Wolf et al., 2019).

For both the victim and extractor, we consider variants of the RoBERTa model (Liu et al., 2020). For the victim model, we use the RoBERTa-Large variant and, for the extractor we use the XLMR architecture (Conneau et al., 2019). The former is a language model pre-trained on a large amount of English data whereas the latter is similar but trained on data from over 100 languages. XLMR demonstrates zero-shot cross-lingual transfer: when fine-tuned on one language, say English, it is able perform well on other languages without seeing a single training example in those languages.

3 Experimental Results

Pre-trained multilingual models also succeed at model extraction with gibberish inputs. In Table 1, we present results for development set results for MNLI for the extracted models. Using the multilingual variant of RoBERTa (XLMR) does not appreciably reduce the extraction performance relative to the English-only variant. For the rest of the experiments, we use the XLMR-Large variant and note that the RoBERTa-Large — XLMR-Large pair exceeds the extraction performance reported in (Krishna et al., 2020) using BERT-Large models under identical conditions.

Models extracted with multilingual pre-trained language models perform well on zero-shot cross-lingual transfer. In Table 2, we present results for the zero-shot cross-lingual transfer. The first three rows correspond to the baseline cases in which the models are trained on 100% real English data (MNLI) and tested against the XNLI dataset (Conneau et al., 2018). Next, we include the novel extraction results where no real data is available and training is performed solely on the gibberish data in English. This model has not seen any grammatically correct sentences labeled for the task in English, and no sentences in other languages labeled for the task, yet it is better than a strong BiLSTM baseline from (Conneau et al., 2018). As is observed in other zero-shot cross-lingual work (Conneau et al., 2019; Singh et al., 2019), zero-shot performance on languages similar to English are comparable to the English performance while those languages which are low-resource and dissimilar to English suffer.

Performance of extracted models greatly improves with a fraction of real data. We now consider the case when the adversary has access to some real labeled data. Here, we sample 1, 5, or 10% of the MNLI (English) data and investigate two ways of using it: during extraction by adding it to the gibberish data, or after extraction as another fine-tuning stage similar to supplementary training (Phang et al., 2018). The results show that even a small fraction of real data can significantly improve zero-shot performance. In particular, 5% of the MNLI English training dataset is enough to lift the performance of extraction to that of XLMR-Base for all languages. Further, the results show that presenting data after extraction is better than mixing it during extraction. This is in line with results from Phang et al. (2018); Keskar et al. (2019).

4 Analysis

The dynamics of model extraction are hypothesized to be similar to that of model distillation (Krishna et al., 2020). NLP models ascribe high confidence to gibberish data (Feng et al., 2018). By distilling a model from such queries, the stolen model's decision boundary approximates that of the victim's.

We further validate this hypothesis by demonstrating that the embeddings of pre-trained language models show similar behavior whether the input data is real or gibberish. We compute representations for each example of the MNLI dataset and the gibberish dataset by max-pooling the outputs of the last layer of pre-trained RoBERTa. For 1000 examples from the MNLI development and 1000 gibberish samples, we compute the minimum Inner-Product distance between each example and the MNLI training data. We plot this nearest distance in Figure 2. Overlap in the distribution suggests the distinction between real and fake is difficult to make by embeddings alone. Though gibberish samples appear random, they sufficiently mimic the input distribution to allow distillation from the victim-

Model	% real data	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh
Baselines																
BiLSTM[1]	100	73.7	64.8	67.9	67.7	68.9	68.7	67.7	64.1	65.4	55.7	64.1	64.2	58.4	66.4	65.8
XLMR-Base[2]	100	85.8	73.8	79.6	78.7	77.5	80.7	79.7	72.4	78.1	66.5	74.6	74.2	68.3	76.5	76.7
XLMR-Large[2]	100	89.1	79.8	84.0	83.9	82.9	85.1	84.1	76.9	81.2	73.9	78.1	79.6	73.8	80.8	80.2
XLMR-Large	0	77.0	67.5	71.4	70.7	68.4	71.8	71.5	64.2	68.3	60.9	65.8	67.4	60.3	67.3	67.6
Additional (real and labeled) data from MNLI available *during* extraction																
XLMR-Large	1	82.1	70.0	74.6	75.4	72.7	77.1	75.4	68.5	72.4	64.7	68.8	71.4	64.3	72.1	72.6
XLMR-Large	5	84.9	74.0	77.7	79.0	76.1	79.8	78.3	71.1	75.7	66.7	73.5	74.4	67.5	75.9	75.1
XLMR-Large	10	85.9	74.9	78.9	80.0	77.0	81.1	79.2	72.6	76.5	68.2	73.6	76.0	69.1	75.7	75.8
Additional (real and labeled) data from MNLI available *after* extraction																
XLMR-Large	1	82.9	73.2	76.8	77.7	75.8	78.8	77.9	70.6	75.0	66.7	71.8	74.2	67.4	73.5	74.1
XLMR-Large	5	86.2	75.2	80.1	80.5	78.6	81.3	80.2	72.5	77.9	68.7	74.6	76.1	68.3	76.3	76.3
XLMR-Large	10	87.4	76.1	80.9	80.8	79.1	82.5	81.5	73.6	78.8	69.7	76.0	76.9	69.7	77.6	77.1

Table 2: Test set performance of various models on zero-shot cross-lingual transfer. The baseline models were trained on MNLI (100% real data). The model extraction experiments were performed by training XLMR-Large on gibberish data with additional 0, 1, 5, or 10% of MNLI data provided during or after extraction.
[1] Results from (Conneau et al., 2018), [2] Results from (Conneau et al., 2019)

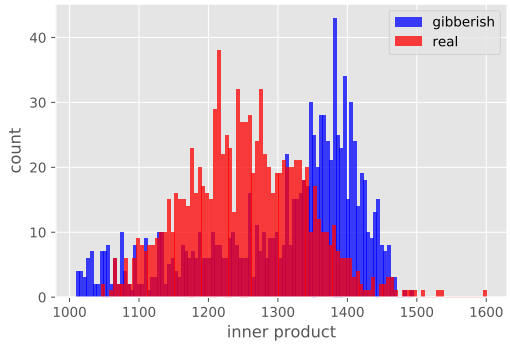

Figure 2: Histogram of lowest distances between embeddings of gibberish and MNLI development set data from the MNLI training data.

teacher model into the local-extracted model.

Finally, we demonstrate that extraction depends heavily on the vocabulary used for random sequence generation and not only on the properties of the models. Instead of using the vocabulary from WikiText-103, we use vocabulary of a dataset derived from papers on COVID-19[1]. The extraction performance drops from 78.6% to random chance. This suggests that model extraction is unlikely to succeed if the domain of the victim model and the input sampling distribution are different. The most common words of the COVID-19 dataset included *influenza, RNA, infection, respiratory, patients, viral* which substantially differ from the more common terms in WikiText-103 such as *television, family, government, military, system*. Whereas our earlier experiments demonstrated domain extension is possible by extracting into a multilingual model, this transfer requires input queries to reasonably mimic the domain of the victim model.

5 Conclusion

We study the problem of extracting multilingual models by stealing from a monolingual model. We query the monolingual victim model with gibberish data. We then use the victim's labels as ground-truth to fine-tune a separate multilingual. This extracts the task-specific knowledge from the victim and transfers it to languages seen by the multilingual model during its own self-supervised pre-training. We show that high accuracy can be obtained on several languages using this approach, and that this performance improves when the extractor has access to a small fraction of real data. We also show that post-hoc fine-tuning on real data is better than mixing real and gibberish data during extraction. We present results underscoring the importance of vocabulary on the extraction performance, and we provide preliminary evidence to support the hypothesis that the dynamics of model extraction are similar to that of model distillation. Our work prompts a deeper investigation into associated topics such as theoretical similarities to distillation, defenses against such multilingual extractions, and improving performance on out-of-domain vocabulary.

[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *EMNLP*.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. *ArXiv*, abs/1805.04770.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. 2019. Prada: protecting against dnn model stealing attacks. In *EuroS&P*.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering, text classification, and regression via span extraction. *arXiv preprint arXiv:1904.09286*.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. 2019. Model reconstruction from model explanations. In *FAT**.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2019. Knockoff nets: Stealing functionality of black-box models. In *CVPR*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N Asokan. 2019. Dawn: Dynamic adversarial watermarking of neural networks. *arXiv preprint arXiv:1906.00830*.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *USENIX*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.