An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry

Wenxin Jiang*, Nicholas Synovic†, Matt Hyatt†, Taylor R. Schorlemmer*, Rohan Sethi†, Yung-Hsiang Lu*, George K. Thiruvathukal†, James C. Davis*
*Purdue University and †Loyola University Chicago

2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), pp. 2463–2475

発表：山村慎太郎

背景｜事前学習済みモデル（PTM）

- 大規模なデータセットで事前に学習された機械学習モデル

—様々な汎用的なタスクに対応できる

—転移学習やFine-Tuningにより，特定のタスクに特化したモデルを生成できる

- 学習には高いコストや大規模なリソースが付き物

⇒既存のPTMの**再利用**により解決

—モデルの再利用により学習コストをカット

- PTMは，”Hugging Face”のようなDLモデルレジストリを介して共有される

背景｜Hugging Face 🤗

機械学習モデルやデータセットを提供するプラットフォーム

- 現在**最大**のPTMレジストリ
- PTMの**再利用**の際に盛んに利用される
- ソフトウェアパッケージ(PyPIなど)をモデルにしている

PTM再利用のあり方や課題点などは研究されていない

通常のソフトウェアパッケージにおける，**再利用**に関する研究は進んでいる

⇒**PTM再利用**においても適用できるかどうかはわからない

目的

代表的なPTMレジストリである”Hugging Face”における，**PTMの再利用**に関する実証的研究を行う

アプローチ

・定性的調査：Hugging Face利用者へのアンケート

・定量的調査：STRIDE分析+α

RQ

RQ1

技術者はPTMを**どのように選択**するのか？

RQ2

PTM選択にどのような**属性**を重視するか？

RQ3

PTM再利用における**課題点**は？

RQ4

Hugging Faceにおける潜在的な**リスクと軽減策**は？

アプローチ(RQ1~3)

アンケート手順

①モデル選択方法，PTM属性，PTM信頼性をテーマに，**インタビュープロトコルを作成**

②HuggingFaceユーザを回答者として**リクルート**

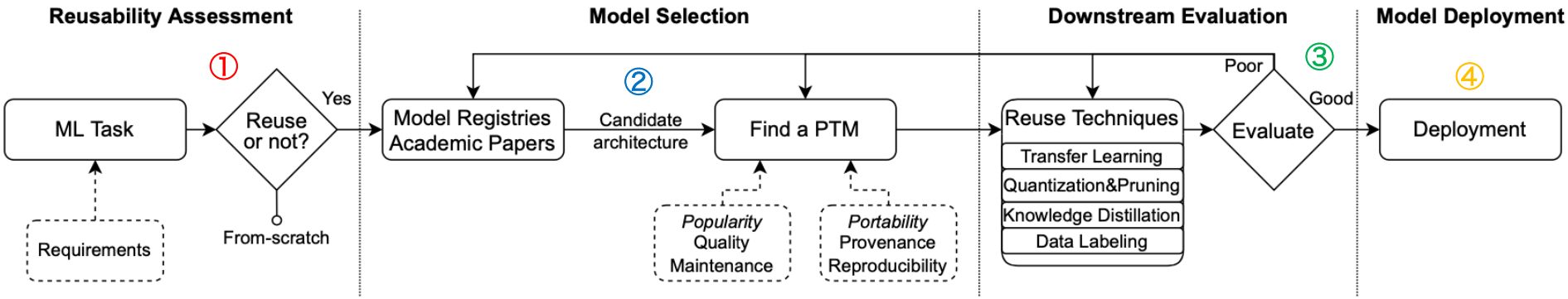
③**インタビュー記録**を書き起こし，テーマに沿ってまとめる

アンケート対象者

Hugging Faceへの<u>モデル投稿数順</u>にユーザをソート，

top50名にコンタクト，**12名**が回答

RQ1｜PTM選択プロセス



モデル選択までの流れ

①再利用が必要・適切かどうか？
②タスクやアーキテクチャに即したPTMを探す
③モデルの評価
④デプロイ

調査結果
- 皆概ね**同様な選択プロセス**を経ていた
- SPに比べ，**評価のやり直し**回数が多い
- GitHubよりも**採用がやりやすい**

⇒Hugging Faceでは，PTMがドメインやタスクレベルで分類されているため

SP:ソフトウェアパッケージ

RQ2｜PTM選択に寄与する属性

ソフトウェアパッケージでも重視された属性

・人気度　・メンテナンス　・品質

⇒人気度(ダウンロード数など)がPTMでは特に重視された

Hugging Face特有の，重視された属性

・Provenance（来歴，実証性）

アーキテクチャや論文からの変更点，学習データセットなど，モデル生成に関する情報

・再現性

論文やモデル説明と同等の結果を出すか

・移植性

動作する環境や計算資源に関する情報

RQ3｜🤗におけるPTMの課題

属性の欠落

PTMに関する，**欲しい情報が書かれていない**ケースが多い

ex)データセット，モデルの詳細，ロバスト性 など

主張と性能の不一致

著者が主張する性能を，実際に発揮してくれないケースがある

⇨**再現性**に関わる

モデルのリスク

プライバシー問題や**悪意を持って学習されたモデル**が存在する

ex)BadNets：

学習データに汚染データを注入，バックドアが設置されたPTMが完成

→誤分類やセキュリティ問題を引き起こす

RQ4 アプローチ（STRIDE分析）

セキュリティに関する脅威を洗い出すための手法

Threat（脅威）
Spoofing（なりすまし）
Tampering（改ざん）
Repudiation（否認）
Information disclosure（情報漏洩）
Denial of Service（サービス拒否）
Elevation of Privilege（特権昇格）

・あるプロセスにおいて，左表の脅威が発生しうるかを考え，それぞれの脅威の分析（発生リスク，シナリオ，対策など）を行う

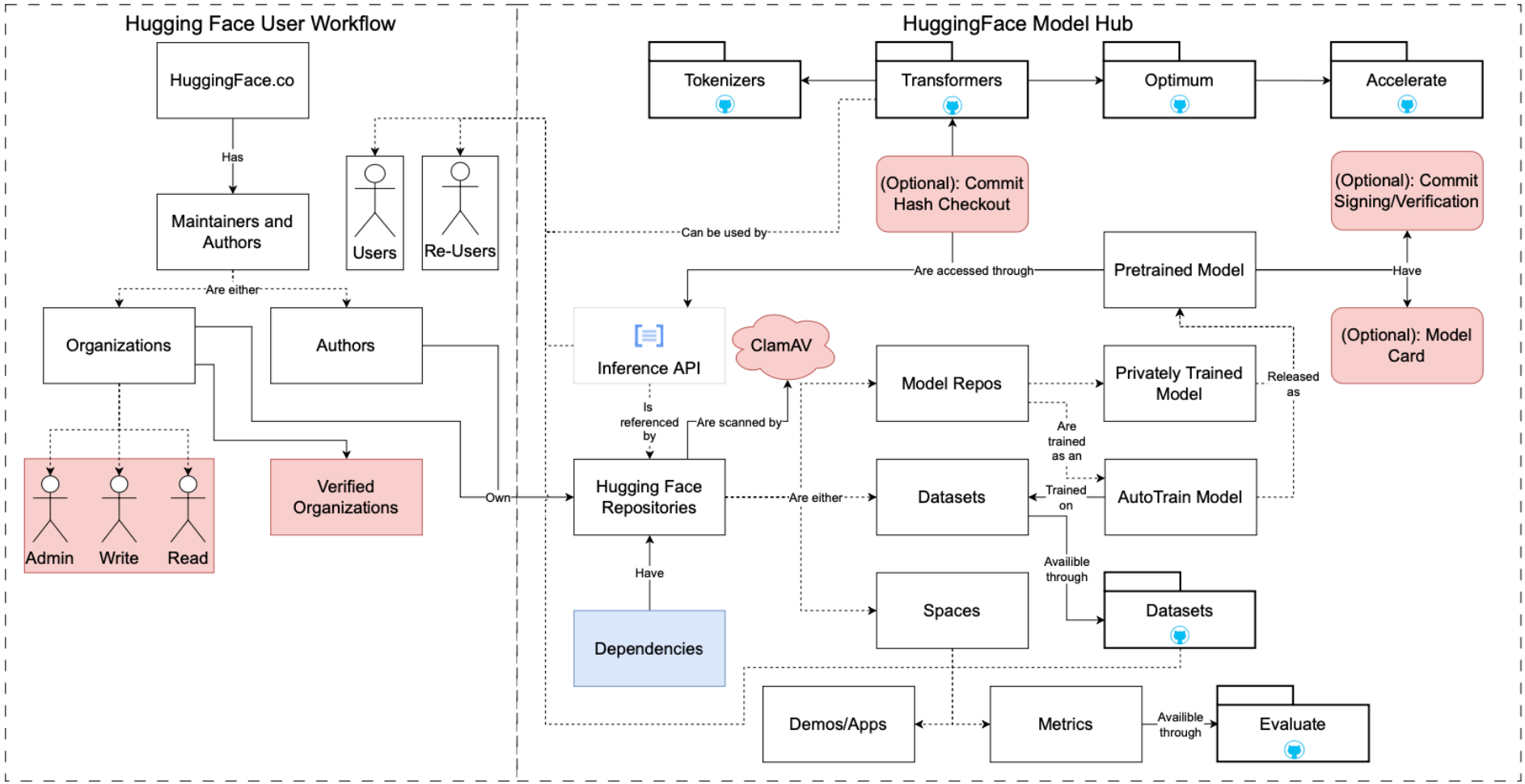
①PTMリポジトリにおけるデータフローに対し，発生しうる脅威とその対策を洗い出す

②Huggingface上のPTMがどの程度その脅威の対策を行っているかを調査

⇒RQ3の課題の定量的な裏付けとなる

データフロー図

データの流れを可視化したもの



RQ4｜リスクと対策

リスクの例と対策方法（一部）

STRIDEリスク	例	対策	現状の詳細
Spoofing(なりすまし)	実在の組織を装い，モデルを提供	組織認証	採用率3.19%
Tampering(改ざん)	一部のモデルを改ざん，依存関係を持つモデルにも影響	?	巨大な依存関係を持つモデルが存在
Repudiation(否認)	情報欠落により，想定と異なる動作をもたらす	PTM文書の充実	機械解析可能な文書を持つPTM0.1%未満

- HuggingFaceにはリスク対策手段が用意されてあるが，**不完全**であったり，**広く普及していない**
- 情報の欠落は，RQ3の**属性の欠落の裏付け**ともなる
- 情報の欠落が起こりやすいのは，モデル情報は**自己申告**であるという性質によるもの

まとめ

- Hugging Faceにおける，PTMの再利用について調査
- PTM選択に至るまでのワークフローを作成
- PTM再利用における課題を特定
  —情報不足や利用におけるリスクなど
- PTM再利用における有用な属性やリスクを調査

所感

- 今後利用するであろうHugging Faceについて少しは勉強できてよかった

- 実際に文書や情報の欠落が多いなら，論文情報などをもとに自動生成すればどうか？

—そんなに役に立たなさそう

- BadNetsなど，機械学習特有のセキュリティ問題は知識として面白かった