

What do we know about Hugging Face?
A systematic literature
review and quantitative validation of qualitative claims

Barcelona, Spain Jason Jones, Wenxin Jiang, Nicholas Synovic,
George K. Thiruvathukal, and James C. Davis

ESEM '24, October 24–25, 2024

背景

ソフトウェアパッケージレジストリ(SPR)

- パッケージの作成者と再利用者をコネクトし、ソフトウェアの再利用を促進してきた。

事前学習済みモデル(PTM)が普及

- ディープラーニングモデル(DLモデル)の開発コストのカット。

Hugging Face

- 代表的なPTMレジストリの1つ

問題

PTMを再利用する上で

- PTMパッケージレジストリは、従来のSPRをモデルにしている.
- PTM固有のコンポーネント(学習の重み, 学習データセット等) も含まれている.

PTM再利用に関する体系的な文献レビューがない

- 体系的なレビュー：システマティックレビュー
 - 文献の選定や評価において、一定の基準を用いて、より信頼性の高い結論を導き出す

目的

PTM再利用に関する体系的な文献レビューを実施し、
先行研究の主張を要約する。

- 現時点でHugging Faceについて研究されていることまとめ。

定性的な過去の研究の主張等を定量的に測定する

- 指標を作成し、曖昧性を無くす。
- Hugging Faceと従来のソフトウェアレジストリの比較を容易にする。
 - 将来の研究者向け

リサーチクエスチョン

RQ1 : Hugging Faceでのパッケージの再利用に関する主張は何か.

- これらの主張を分析することにより, 定量化可能な指標に変換する.
- 比較可能性が向上する.

RQ2 : Hugging Faceにおけるパッケージ再利用 (PTM) に関する定性的な主張は、定量化された場合にも成り立つのか？

RQ1：パッケージ再利用に関する主張

1. Hugging Faceのパッケージ再利用に関する12の定量化可能な主張を特定した.
2. 主張は、小規模及び大規模な定量的測定、定性的調査、インタビュー、ケーススタディの5つの定量化可能なカテゴリに分散.
3. デザイントレンド、文書化、理解、選択の考慮事項といった項目は定量化不可能であった.

シネマティック文献レビュー

1. 研究の特定
2. 研究の選択
3. 研究の質の評価
4. データ抽出
5. データ分析

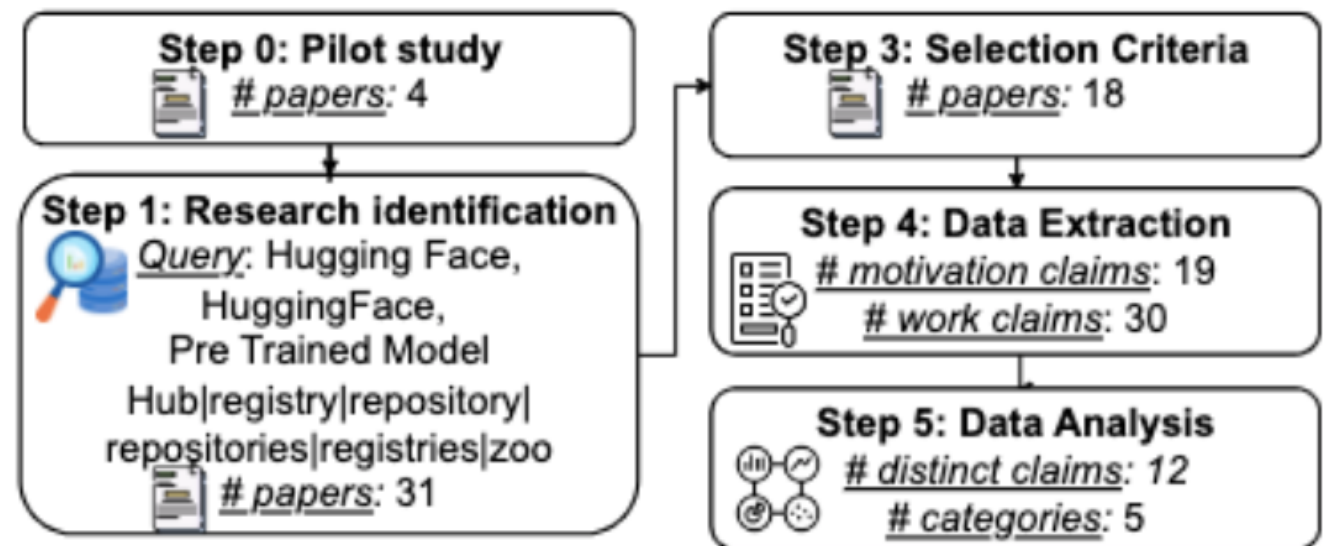


Figure 4: Systematic literature review process and results.

Step3：選択基準

- モデルの再利用性の説明を含む.
- PTMの再利用の実践、課題、または影響に焦点を当てている.
- 査読付きジャーナルまたは会議の進行物に掲載された論文のみ.
- 過去5年以内(2019-2024年)の出版物.

分析と結果

Step4：抽出された主張を2つのカテゴリに分類

- 動機の主張：なぜこの研究を実施したか等.
- 研究の主張：実際に得られた結果から導かれる主張.

5つのカテゴリと12の主張

テーマ	主張 (Claim)	主張の根拠 (Claim Basis)	出典 (Works)
設計の傾向 (Trends in Design)	人気モデルは少数の貢献者によって所有されている	大規模測定	[14]
	Transformersライブラリは、PTMの作成および下流再利用の容易さを高める	大規模測定、ケーススタディ	[14, 31, 32, 56]
	+ TransformersライブラリはPTM進化のプロセスを改善する	ケーススタディ	[56]
	フォークされたりポジトリは、軽微な脆弱性を引き起こす	大規模測定	[32]
選定の考慮事項 (Selection Considerations)	+ モデルアーキテクチャ、性能、再現性、移植性などDL特有の属性が選定に影響する	インタビュー	[28]
	+ 「人気」は「保守性」や「品質」よりも選定に影響を与える	インタビュー	[28]
リポジトリのライフサイクルと保守 (Repository Lifecycle and Maintenance)	モデルは時間とともに保守され、人気が高く大規模でドキュメントが充実しているものほど保守されやすい	大規模測定	[14]
ドキュメントと理解 (Documentation and Understanding)	Hugging Face上のモデル情報は十分に文書化されていない	小規模・大規模測定、アンケート調査	[21, 32, 38, 50]
	+ ドキュメントの質はモデル選定に影響を与える	アンケート調査、ケーススタディ	[13, 38, 50]
	モデル名は一貫性がなく、アーキテクチャを適切に表現していない場合がある	大規模測定、アンケート調査	[27]
	データセットの文書化は、その人気と相関している	小規模測定	[58]
下流利用 (Downstream Usage)	Hugging Faceは指数関数的に成長しているプラットフォームである	大規模測定	-

+ は、「大規模定量的な根拠に基づかない主張」

RQ2：定性的な主張の検証

1. PTMの80%以上で、transformersライブラリが好まれており、PyTorchを上回っている。
2. 従来より早いペースでパッケージの回転がされている。
3. モデルの人気度と派生モデルは相関があるが、別の要因が影響を与えている。
4. ドキュメントの品質と人気度に相関があり、これがモデル選択において重要視されている。

C1

トランスフォーマーはPTMの作成と再利用のアクセシビリティを向上させる。

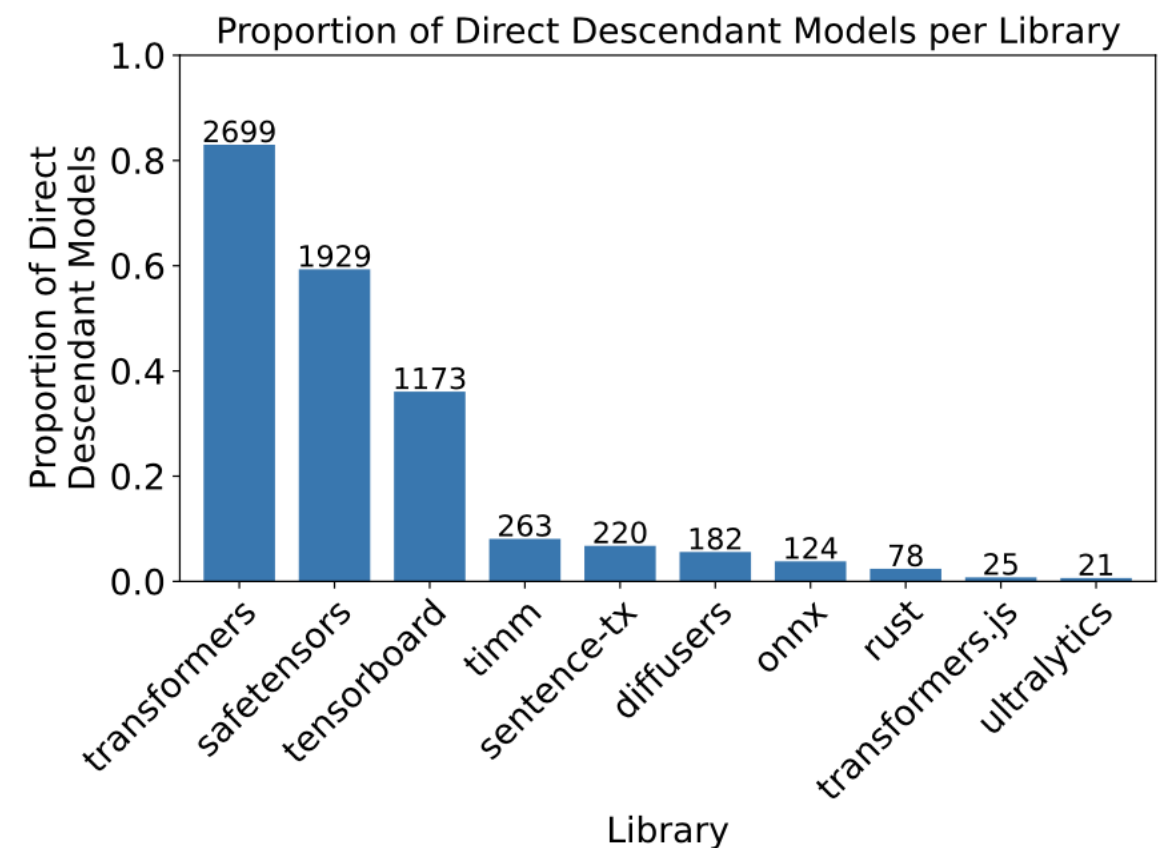
- 指標：ライブラリから子孫モデルへの保存率

- 考慮する要素:

- 評価するライブラリ L
- ライブラリ L を使用し、かつ他のライブラリも使用する基本モデルの集合 B
- 基本モデル $b \in B$ の直接の子孫の集合 D_b
- 子孫 $d \in D_b$ がライブラリ L をサポートしている場合に1を返し、そうでない場合は0を返す関数 $S(d, L)$

- 保存率の計算式:

$$P_L = \frac{\sum_{b \in B} \sum_{d \in D_b} S(d, L)}{\sum_{b \in B} |D_b|}$$



C2

人気のあるモデルは、使用される可能性が高い

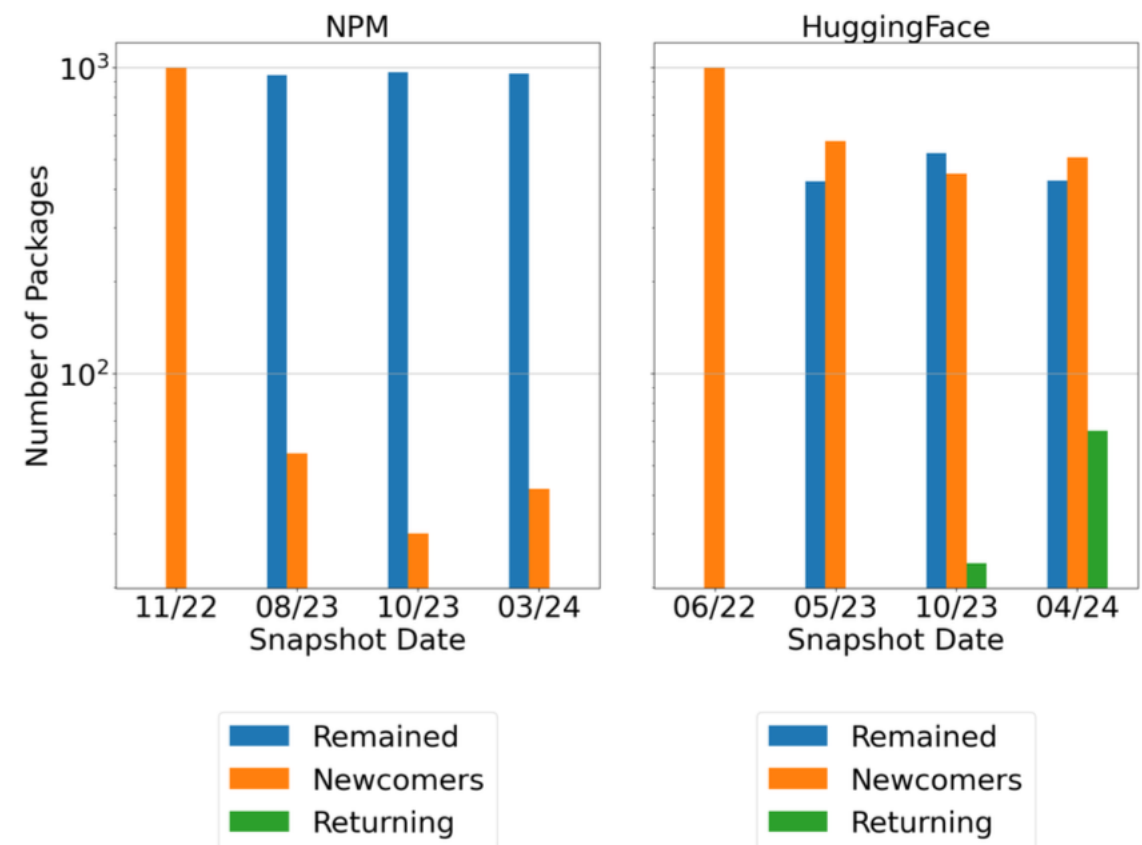
- 指標1：トップPTMの売上高
- 指標2：トップパッケージの子孫数

- Ruturningの値が大きい

- トップモデルが急速に入れ替わっている

- 指標2について

- 人気のモデルほど派生モデルが多く、正の相関が見られた。



ダウンロード数トップ1000の売上高

C3

ドキュメントの品質がモデル選択に影響

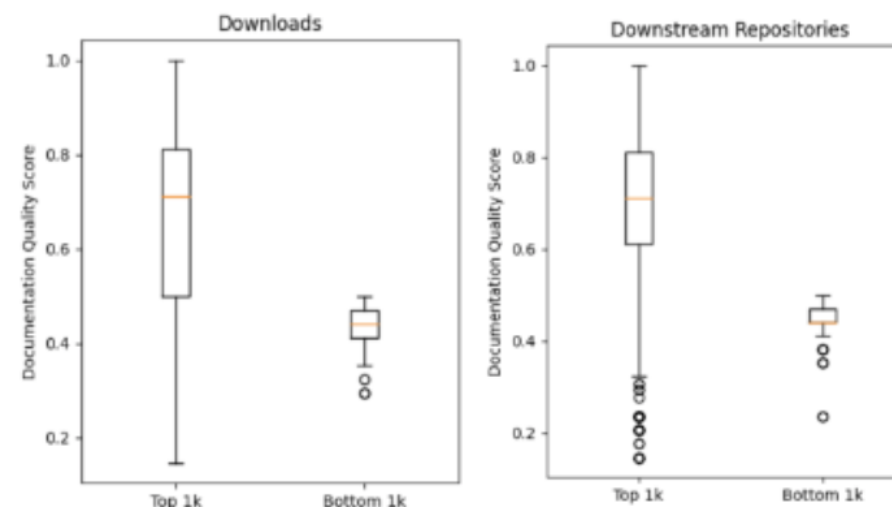
- 指標：ドキュメントの質(モデルカード)

1. モデルカードの完全性

- モデルの説明, 制限, 使用方法, トレーニング, 評価の項目が存在するか
Open AIのGPT-4に問い合わせ, 真偽を0~5点で評価.

2. メタデータの可用性

- Hugging Faceによるダウンロード数, ライク, PeaTMOSSデータセットによるマッピングの下流依存関係を考慮し, 人気度を評価, 正規化等の処理をして(0~1)で評価.



結論・今後の取り組み

PTM再利用に関する知見を体系的に整理し、
従来のソフトウェアパッケージレジストリとの比較

- PTM固有の再利用プロセスや課題を明らかにした

人気度がモデルの選択と再利用の重要な要素

- しかし、唯一の決定要因ではない。

モデルの再利用性と適応性に影響を与える深層学習固有
の属性を明らかにし、定量化するためのさらなる研究

- アーキテクチャや再現性

今後の取り組み・所感

過去の分析結果が短期間で陳腐化してしまう
リスクがある

- リアルタイムで追跡が可能な分析基盤が必要である.

Hugging Faceのような大規模モデルレジストリの
急速な進化は定量化評価を困難にする

- モデルのバージョン管理, バージョンアップの履歴をしっかりとデータで残すことが重要である.

