特許情報に関する言語生成モデルを活用した知的財産創造手法の開発

Development of Intellectual Property Creation Method Using Language Generation Model on Patent Information

Shigeaki Onoda

Graduate School of Information Engineering, Toyama Prefectural University
t855005@st.pu-toyama.ac.jp

Wednesday., 10 4, 2019,
Toyama Prefectural Univ.

今週行ったこと

知財案提案システム生成

1. 新規データのクロール
2. 特許データローダーの開発
3. 山元さんとの卒研の兼ね合い相談

新規データのクロール

今週は新規データとして量子コンピュータのデータを追加でクロールした

量子コンピュータ特許概要

検索クエリ：quantum computer
取得できた特許数：861

今後の特許キーワードの決め方

これからどの分野を取るかは技術トレンドのリスト？的なものがあれば自動化できる
知ってる方いたら教えてください
例：https://ihsmarkit.com/Info/0119/top-tech-trends-2019.html

データローダーの開発

ニューラルネット系学習のよくある手順

1. データの前処理
2. データセットのロード
3. モデル構築
4. ロスの計算
5. オプティマイザによる重みの更新

この手順で前処理やデータセットのロードは似たような処理が何度も発生するためただ作るだけでなくデータローダーという概念を導入して再利用性の増加・属人性の低下できるようなシステムアーキテクチャにした

データの処理の最適化

上のデータ前処理とデータのロードを簡便化かつ属人性を排除するためデータローダーという概念を用いる

transforms
データの前処理を担当するモジュール

Dataset
データとそれに対応するラベルを 1 組返すモジュールデータを返すときに transforms を使って前処理したものを返す。

DataLoader データセットからデータをバッチサイズに固めて返すモジュール

通信のプロトコルのようにフォーマットを統一することで綺麗なデータパイプラインを構築可能！

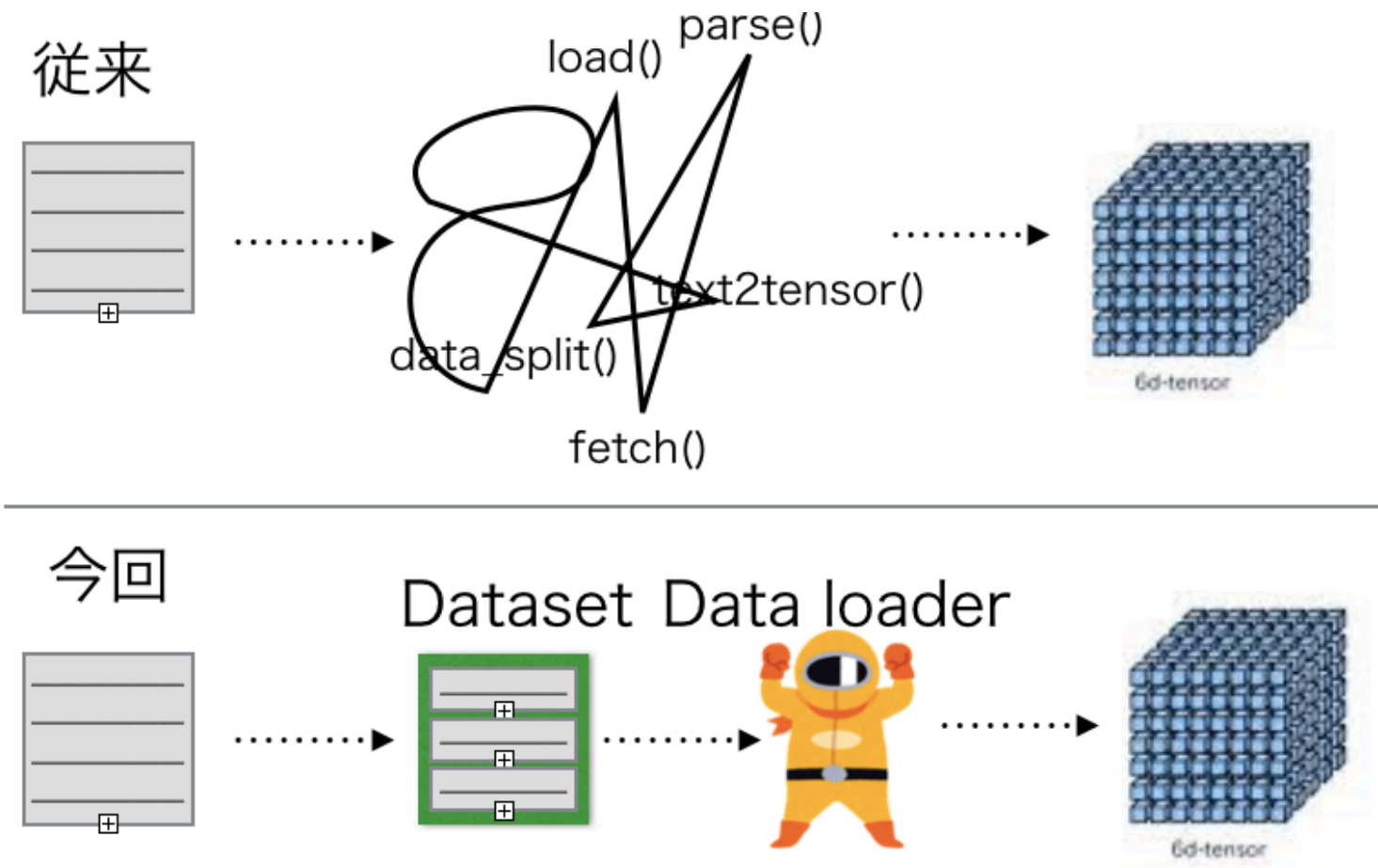

Figure: 1: データローダーの概要

山元君の研究との関連

山元くんの研究で検索結果からの２つの検索ワードからの重要単語？を見つけることにシフトしつつあるので, 小野田の研究との直接的な接地はなくなった
そのためページランク (HITS) を使う場合は自作で作る必要になった

現状の問題点

データローダーから読み込んだデータを学習させようとしたが, 入力信号となるキーワードが多く取りすぎていたことが確認できたため一旦サスペンド中
例えばある特許は 5000 単語に対して, 800 キーワードとキーワードが多すぎる

取るべき対策

- キーワード抽出時に数を絞る（簡単）
- 10 くらいで割ってサブセットを作り交差検証のように複数組み合わせで回す

どちらが良さげでしょうか？

今後やること

ToDo

1. キーワードを絞る
2. データーローダーで読み込んだデータを学習
3. アプリの開発