P-1-7:CPU環境で高速に動作する軽量OCR「NDL古典籍OCR-Lite」の開発

国立国会図書館　青池 亨

概要

□ これまで国立国会図書館（NDL）が開発・公開してきたOCRは動作環境にGPUを必須としてきた
→ GPUの準備や環境構築を行わないと動作しないため、**誰でも使えるとは言い難かった**

□ 一方で、「みんなで翻刻データ」等の加工・利用によるデータセット構築の成果を積み上げてきた

✓ 構築したデータセットを利用して、レイアウト認識及び文字列認識を軽量化する代替手法を検討

✓ 開発した軽量なOCRを更に使いやすくするため、Windows/Mac/Linux向けデスクトップアプリを作成

NDLにおけるこれまでのOCR開発と課題

➢ 明治期以降の活字の図書・雑誌資料→NDLOCR（ver.2.1）※主に委託で開発、納品後にNDLで一部改良

➢ 古典籍資料→NDL古典籍OCR （ver.3）　※「みんなで翻刻データ」等オープンデータを利用して内製開発

両OCRとも、動作環境にGPUを必須する点が短所

- NDL目線での課題：
職員が手元でスキャンした資料やピンポイントで必要な画像を即席でテキスト化する用途には不向き
- OCRユーザ目線での課題（実際にNDLに寄せられたフィードバックから）：
「環境構築方法が難解」「Colaboratory版は便利だが、サービス側の仕様変更で時々動かなくなるのが困る」
「海外の研究者向けにMac対応してほしい」「障害当事者が自身で活用できる形態で公開してほしい」

汎用的なCPU環境で高速に動作するOCRの開発はNDL内外にニーズがある

軽量なレイアウト認識・文字列認識手法の検討

「**モデル（パラメータ）サイズ**」、「**認識性能**」及び「**寛容型OSSライセンス**」の3点を重視

レイアウト認識手法

学習及び検証用データセット：
- **NDL-DocL（1,219画像分、公開済）**
- **みんなで翻刻の対象画像から新規作成（1,389画像分、未公開）**

テスト用データセット：NDL-DocL評価用（210画像分、非公開）

手法の名称	mAP_50	パラメータサイズ
RTMDet-tiny	0.947	4.9M
RTMDet-small	0.957	8.9M
YOLOv9-S	0.952	7.1M
YOLOv9-M	0.948	20M
Cascade Mask R-CNN (backbone:ConvNeXt-T)	0.905	28M

※YOLOv9については、公開当初の実装は非寛容型のGPLライセンスであるが，原著者らによるMITライセンスによる再実装が進められていたため候補とした

文字列認識手法

学習及び検証用データセット：
・OCR学習用データセット（みんなで翻刻）（502,065行分、公開済）

テスト用データセット：
・OCR学習用データセット（みんなで翻刻）（21,218行分、公開済）

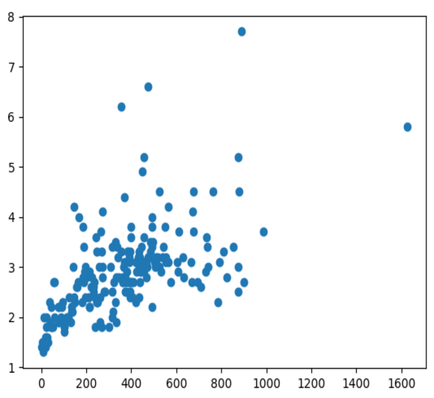
手法の名称	F値	パラメータサイズ
PARSeq	0.861	29M
PARSeq-tiny	0.854	8.8M
TrOCR-small	0.888	99M

今回の目標上、認識性能が少し低下してもCPUで高速に動作することが重要と考えられたため「**RTMDet-small**」及び「**PARSeq-tiny**」を採用した
この結果、NDL古典籍OCR ver.3の手法（表中グレー部分）と比較して、大幅なモデルサイズの縮小を実現した

NDL古典籍OCR-Lite全体でのテキスト化性能

処理速度

NDL職員事務用端末上で
平均値2.9秒，分散値0.78
（詳細は論文参照）



画像中に含まれる文字数（横軸）
処理時間（秒）（縮軸）

認識性能

みんなで翻刻データの特定のプロジェクト成果物を利用してテキスト認識性能の評価を実施

- NDL古典籍OCR-LiteのF値の中央値は：0.894
- NDL古典籍OCR ver.3のF値の中央値：0.920

性能劣化の一因
- レイアウト認識結果の矩形が上下方向に若干小さく推論される事象を確認

→公開版のNDL古典籍OCR-Liteでは検出矩形の上下に高さの2%分マージンを付与

UIの作成及びマルチOS対応

Windows/Mac/Linux向けデスクトップアプリケーションを開発・公開

是非お試しください！