

日本清華同方OCR技術のご紹介



日本清華同方
Japan Tsinghua Tongfang, Inc

目 次

OCR現状

当社OCRの全体構成図

属性説明

技術説明

補足説明

効果図

OCR現状

- OCRとはOptical Character Recognition

光学文字認識ということ

- 現状技術

従来のOCRでは、テンプレートマッチングと特徴量の方法が使われており、近年フレームワークは主にTesseractが使われている。明確な標準テキストを認識するには問題無し。

**少し複雑なケースでは（認識角度が真正面でない場合や手書きの状況など）、
正解率が非常に低く、解析時間も長い**

当社OCRの全体構成図

フレームワーク (PaddleOCR)

位置検出最先端技術
DBアルゴリズム*

文字認識最先端技術
CRNNアルゴリズム*

独自アルゴリズムより
さらに精度アップ

文字抽出

文字認識

演算速度 0.0844秒
認識率 96.158%

画像

テキスト

属性説明

●当社技術の特徴

独自アルゴリズム

- 一、認識精度・ロバスト性及び認識速度を高めるためには、**独自の学習モデル蒸留技術**を開発
- 二、OCR主流技術の巨大の学習モデル「**ResNet34_vd**」から**13.29倍**圧縮し（**MobileNetV3**ファインチューニングモデル）上で、**認識精度は96.1538%、認識速度も3.2倍高速化の0.0844秒**に達成。
- 三、独自データ拡張により**不規則な手書き**の文字や様々な角度の文字の認識場面に置かれても**ロバスト性も大きく向上**

学習データ

- 一、学習データは100万+日本語を用いた
- 二、各種フォーマット。（スーパーや飲食店・ホテルの領収書，電車切符）
- 三、**独自データ拡張（学習データに存在しないスタイルや文字の認識角度・照明条件を模擬）**

属性説明

ベースライン

- ・ OCR主流技術と同じ，文字列位置検出と認識の2段階アルゴリズムを採用：
文字列位置検出は最先端のDBアルゴリズム
文字認識は最先端のCRNNアルゴリズム

ラベリング・学習フレームワーク

- ・ 最先端の深層学習OCRフレームワーク「PaddleOCR」を利用
ラベリング、学習、配布。**システムの拡張性は高い。**

属性説明

インファレンスエンジン

- ・ インファレンスエンジン（学習済みの学習モデルの利用）はCPU、もしくは、GPU（おすすめ）の利用が可能、
- ・ お客様のニーズに応じて学習データの追加や、配布は柔軟
特殊フォーマット、手書きなどに対応可能。

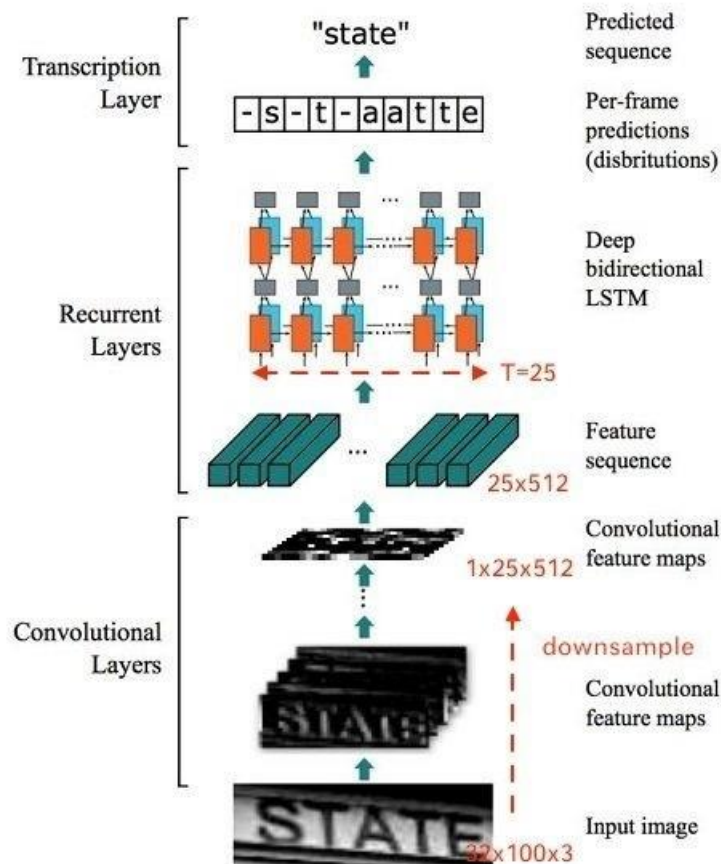
今後の拡張

- ・ 技術提案について、他の言語への適用も可能。
例えば：英語、フランス語、中国語、韓国語など。

技術説明（概要イメージ）

OCR主流技術
2段階アルゴリズム

文字列位置検出（DB）
文字認識（CRNN）



* 画像はGitHubより

補足説明

1. 文字列位置検出は最先端のDBアルゴリズム

Differentiable Binarization algorithm 微分型2値化アルゴリズム

論文リンク : <https://arxiv.org/abs/1911.08947>

2. 文字認識最先端のCRNNアルゴリズム

CNN+RNN+CTC

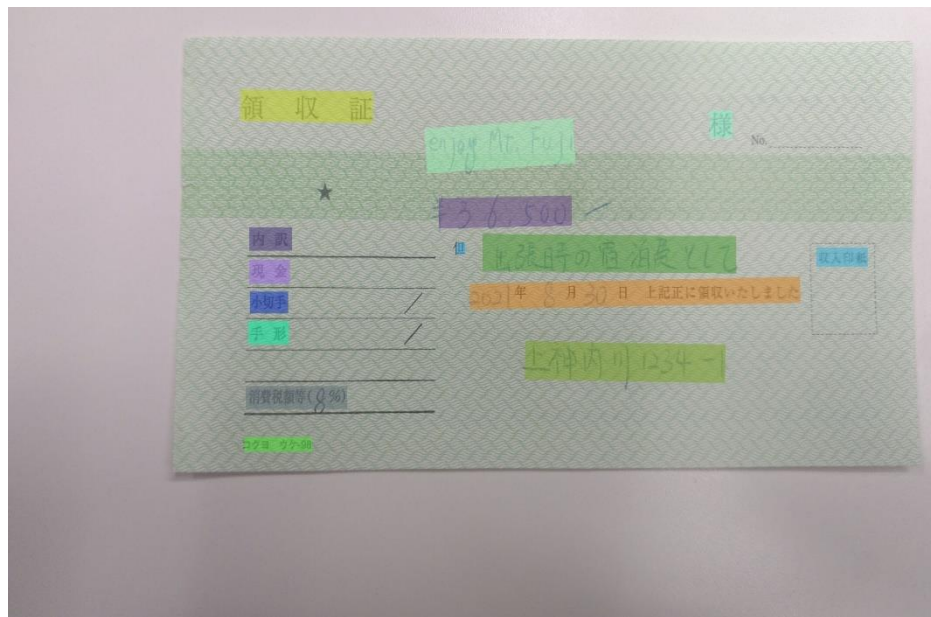
CNN: Convolution Neural Network 特徴抽出

RNN: Recurrent Neural Networks 再帰型ニューラルネットワーク

CTC: Connectionist Temporal Classification コネクショニスト時間分類

論文リンク : <https://arxiv.org/abs/1507.05717>

効果図



領収証

様

enjoy Mt. Fuji

36,500

出張時の宿泊費として

収入印紙

2021年8月30日上記正に領収いたしました

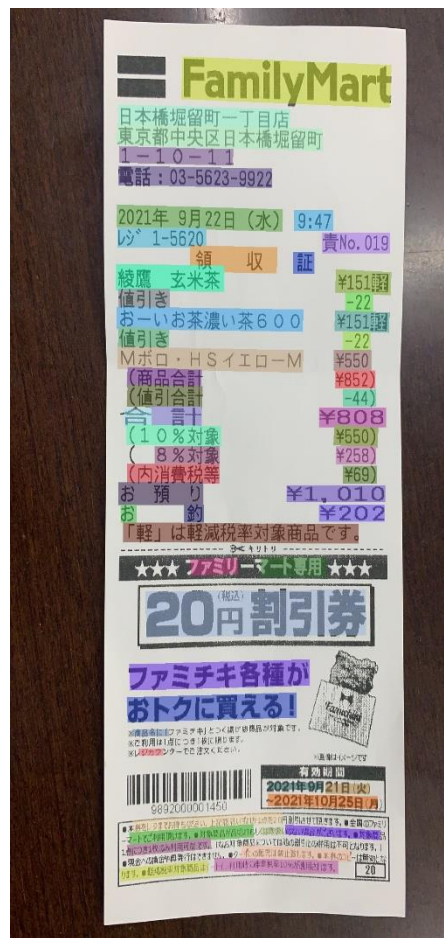
上神内川1234-1

内訳
現金
小切手
手形

消費税額等(8%)

コクヨ ウケ-08

效果图



ご清聴ありがとうございました