CNNによる商品札文字認識

成沢 淳史[1,a)]　柳井 啓司[1,b)]

1. はじめに

近年, Convolutional Neural Network (CNN) を用いた手法が物体認識をはじめ様々な認識タスクで高い精度を示している. また, CNN を用いて検出及びセグメンテーションを行う研究も報告され幅広い応用がなされている. 文字認識においても例外でなく手書き数字データセット MNIST における認識精度が CNN を用いることでほぼ 100%の認識精度を達成する. さらに, 認識のみならず情景画像中からテキストを検出し読み取りを行うシーン文字認識の研究 [1] が行われ, 検出から認識までを CNN でモデル化する手法が従来よりも良い結果を得ている. 我々はこうした現状から文字認識と物体認識によるシナジーを期待している. しかし, 日本語において上記のような課題に CNN を適用する場合には次の疑問が存在する.

(1) 日本語のように数千に及ぶ多様な文字種の場合に学習が行えるか.
(2) 情景中に見られる文字を含んだ日本語データセットがあるか.
(3) 情景画像中の部分的に文字に見える領域をノイズとして識別できるか.

本研究では以上の疑問に対して調査を行い, 実際にフォントから画像を生成し CNN がどの程度の文字を識別できるのか検証した. また, シーン画像からテキスト領域を推定する実験を行う.

2. 目的

我々は次の二つの実験を行い, CNN の日本語文字識別可能性を検討する.

(1) 日本語フォントから画像の生成を行い, CNN を学習し文字どうしの識別可能性を評価する.
(2) 情景画像中のテキスト領域と背景領域の二値分類が可能な CNN を学習しテキスト領域の推定を行い評価する.

[1] 電気通信大学
[a)] narusawa-a@mm.cs.uec.ac.jp
[b)] yanai@mm.cs.uec.ac.jp

3. 関連研究

まず我々は研究に利用可能な日本語データセットに関して調査を行った. 有料の日本語データセットとして環境研究所 [*1] が販売を行っている. このデータセットは活字に加え手書き文字が含まれる. 1 文字あたり 25 枚程度の画像が 14.7 万枚収録されているため 5880 種類のひらがな, カタカナ, 英数字, 記号, 漢字に対応する唯一の大規模データセットとなる. 一方, 無料で利用可能なデータセットに ETL9[*2] が存在する. 80 年代に産業技術総合研究所の前身となる通商産業省工業技術院が作成した手書きデータであり, 約 3000 種類の文字種に対応する. こうした活字, 手書きとは異なり情景画像中の文字に対してアノテーションを施しデータセット作成に取り組む活動が見られる [2]. 実データを用いることで複雑な背景であっても学習できる可能性が高まるため貴重な資源である. しかし, 実データを有するデータセットは数少なく未だ目的にあったデータセットが必ずしもある状況ではない.

画像から文字を読み取る手段として日本語に対応した OCR が存在する. この OCR はテキスト文字に対して有効であり, 長年の開発により徐々に画像中からテキスト領域を推定する前処理を行うことで文字を正確に読み取れるようになっている. これとは別に情景文字認識サービスを NTT Docomo が Web API[*3] として開発者向けに提供を行っている.

4. 手法

4.1 CNN の構造

実験に用いる CNN のアーキテクチャーは caffe [3] のリポジトリに含まれる cifar-10-quick.prototxt を参考にした. またテキスト領域推定時には完全結合層は使わず, 最後の畳み込み層の出力をソフトマックス正規化したものを利用する. また, 完全結合層を全く使わずシーン文字のスポッティングの手法 [5] が提案されており本研究においても参考にしている. 本論文では図 1 のようなネットワークを利

[*1] urlhttp://naturecode.ddo.jp/home/
[*2] urlhttp://etlcdb.db.aist.go.jp/etlcdb/etln/etl9/etl9.htm
[*3] https://dev.smt.docomo.ne.jp/?p=docs.api.page&api_name=character_recognition&p_name=api_reference

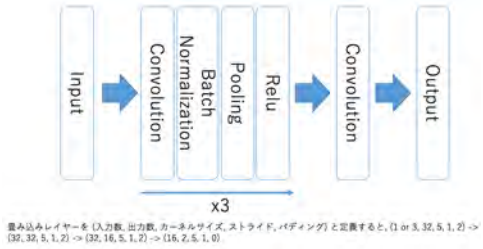

図 1　ネットワーク図

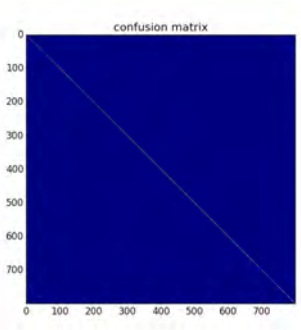

図 2　文字の混合分布行列

用し実験を行う.

4.2　データセット

情景文字を含む画像として野菜など扱う大型のスーパーにて値札の画像を収集した. この画像データは今後我々が研究を展開していく対象である. 今回は収集した 77 枚の画像に対して値札領域にバウンディングボックスを付与した後, 値札領域を前景, それ以外を背景とし MSER [6] アルゴリズムにより画像パッチを収集し検出での二値分類モデル学習に利用する.

またスーパーで見られる文字を特定するためネットスーパーの商品情報に注目し, 1407 種類の商品名をテキストデータとした. このテキストデータから文字の種類を求めた結果, 半角, 全角の区別なしで 799 種類の文字を特定した. この 799 文字に対して日本語フォント 92 種類から回転, スケール変換を加え, 各 500 枚画像を生成し全体で 40 万枚の画像データセットとした.

5.　実験

ここでは 2 つの実験の内容と結果を記載する.

5.1　CNN による日本語文字識別

図 1 のネットワーク構造を用いて文字の種類を 799 種として学習を行い, 図 2 のような相関関係から似ている文字を探すことを行う. ソフトマックス出力のソートを行い似ている文字を探した結果, 数字の”5”と英字の”S”が似ているなど判明した. また, いとへんのような部首が同じ文字は似ていることも判明した.

図 3　左 : 入力, 中央 : 特徴マップ, 右 : ラベリング後

5.2　テキスト領域推定

値札が写る商品画像から値札領域を前景, それ以外を背景とし MSER アルゴリズムを適用し注目領域からイメージパッチを取得する. このイメージパッチをデータセットとし 2 クラス分類器を学習した. 学習したトレーニング画像に対する適用結果を図 3 に示す.

6.　考察

日本語の多様な文字種に対して CNN は十分な精度を発揮した. CNN の特徴表現により文字単体をベクトルで表せるようになり単語同士の動的マッチングにも利用出来ると考えられる. また CNN を用いた領域推定のアルゴリズムは我々の研究において, テキスト領域獲得の有効な手段であると実験から評価できる.

7.　終わりに

我々は文字認識が物体認識と CNN という共通の土俵の上でうまく調和するだろうと考えている. 今回の結果から文字領域の情報が物体認識の結果に作用するような研究が展開できると考えている.

参考文献

[1] M. Jaderberg. *Deep Learning for Text Spotting*. PhD thesis, University of Oxford, 2015.

[2] 瞳佐藤, 崇宏松田, 佑輝池田, 雅一岩村, 浩一黄瀬. 大規模日本語情景文字データセット opu-jst-1(一般セッション, パターン認識・メディア理解における組み合わせと統合, 文化の振興と教育). *Technical report of IEICE. PRMU*, Vol. 113, No. 493, pp. 49–54, mar 2014.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[4] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[5] A. Zamberletti, I. Gallo, and N. Lucia. Augmented text character proposals and convolutional neural networks for text spotting from scene images. In *Asian Conference on Pattern Recognition*, 2015.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, Vol. 22, No. 10, 2004.