[シンポジウム]

人工知能と台湾総督府文書

目加田　慶　人

一、はじめに

昨今「人工知能（AI)」や「ディープラーニング（深層学習)」として見かけない日はない技術と台湾総督府文書とをつなぐ学際的なプロジェクトは、ディジタル・ヒューマニティズプロジェクト（DHP）として 2015 年に始まった。学際的な研究プロジェクトが成功するためには、それぞれの役割を着実に実施することは当然として、研究に対する文化・価値が異なる研究者がお互いの研究を理解し認めるだけの交流が必要となる。本プロジェクトは、それぞれの考えや研究成果を定期的に議論し、それらを他大学との研究連携につなげるなど継続的に活動しているものである。

本稿では、台湾総督府文書の文字画像を対象として、文字画像を入力したときにその文字コードを出力する深層学習の開発状況について報告する。

二、画像認識のためのニューラルネットワーク

画像認識における深層学習は「畳み込みニューラルネットワーク(Convolutional Neural Network、CNN)」が使われることが多い。その学習では、畳み込みと呼ばれる演算によって獲得される多数の画像特徴とそれらの関連をネットワークとして定義し、入力された画像がしかるべきカテゴリに分類されるような画像特徴の種類とそれらの関連の強さを獲得するものである。ネットワークの出力は各カテゴリに対する信頼度となり、この信頼度の高いカテゴリに入力画像を分類するものである。2000 年ごろから画像分類への能力の高さが知られるようになったと共に、グラフィックスプロセッシングユニット（GPU）を利用することで高速に畳み込み演算を実行できるようになったことから、広範に利用されるようになった。画像分類のための CNN において、認識性能を左右するのは画像特徴の種類など、ネットワークを定めるパラメータである。その獲得のために、あらかじめ正解カテゴリが付与された学習データを用意し、その学習データの分類性能が高くなるように最適化することでパラメータは決定される。したがって CNN の性能を向上させるには、表現力の高いネットワークの構造と、そのパラメータを決定するのに十分で質の高い学習データを準備することが大切である。また、学習された CNN の性能を評価するためには、学習データとは異なるデータ（テストデータ）を用意し、これらを分類したときの正しく認識された割合である認識率で評価される。

三、CNN による文字認識

画像認識の研究分野において、文字認識に関する研究の歴史は古く、1970 年代に開発された郵便番号を自動で読み取り区分する機械など、実

4を9に誤る　9を4に誤る　2を7に誤る　3を7に誤る

図1　手書き数字認識における認識誤りの例

用化されているものも多い。文字認識は、文字画像から各文字種に特有な特徴が得られるように特徴抽出をおこない、それに基づき文字画像を分類するための分類器を構築する課題である。数字であれば10カテゴリの分類問題になるがそれですら簡単な問題ではなく、例えばMNIST[1]と呼ばれる手書き数字データセットから6万文字を学習データとして利用し、認識率が98.3%の分類器を構築できたとしても図1のような分類誤りが存在する[1]。文字認識の研究の進展にMNISTや手書き漢字を集めたETL文字データベース[3]が大きく寄与したように、本研究プロジェクトにおいても台湾総督府文書から文字画像データセットを作成することから始め、約17万文字、2600字種となっている。ただし、文字認識研究のために作られたデータセットではないため、片仮名が全体の2割強を占めるなど文字種毎にデータ数が偏る。また、文字画像の平均サイズは縦23画素、横21画素と、漢字を認識対象とした場合には十分な解像度とは言えない。

本プロジェクトで利用しているCNNの構造は、学習データ数が十分でないことを考慮して畳み込み層を2層とGoogLeNet[4]で利用されているInception V1 moduleを4層つなぎ合わせたネットワークとした。また、文字画像データセットは4対1に分割し、前者を学習データ、後者をテストデータとした。一般に学習データのバリエーションを増やすために「データ増強」と呼ばれる方法が用いられる。これは学習データに平行移動や回

1　なお、このデータセットに対する世界最高水準の認識率は99.75%である[2]。

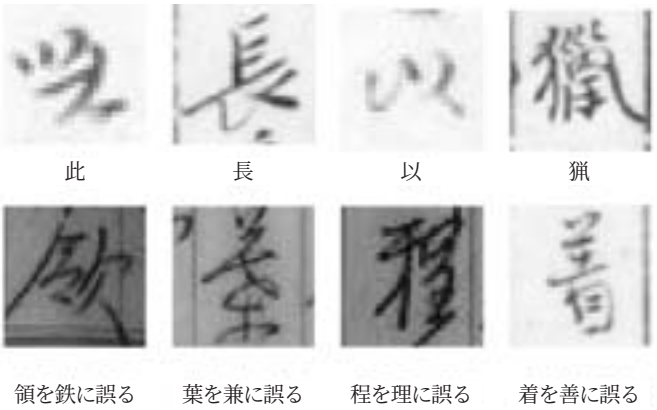

図 2　認識結果の例（上段：認識正解例、下段：認識失敗例）

転などの微小な変動を加えてデータを水増しする方法である。本研究では、上記で述べたデータ数の偏りの影響を少なくするために、学習データに含まれる数が少ない文字種に対しては変動量を大きくしてデータ数を増加させた。認識結果の例を図 2 に示す。認識に失敗した例であっても、正解文字種の概形に類似した文字種に誤っていることがわかる。約 3.2 万文字のテストデータに対する認識率は 90.0%であり、データ増強を行わない場合が 87.7%であることからその効果が確認できる。学習データに 50 画像以下しか含まれていない文字に限った認識率は、データ増強により 3.1%の認識率の向上が認められたものの 77.0%であった。出現頻度が低い文字の認識率を向上させるための更なる工夫が必要である。

四、まとめと今後の展望

本稿では、深層学習に基づく台湾総督府文書の自動解読を目的とした研究プロジェクトの概要を報告した。翻刻された文書画像から各文字の領域を抽出し、抽出された文字画像領域と翻刻結果を学習データとした。出現

頻度が低い文字種に対して重点的にデータ増強をおこない、深層学習による分類器を構築した。約 16 万文字を利用した実験により、信頼度が 1 位の文字が正解である割合は 90%であった。信頼度が 10 位までに正解が含まれる割合は約 98%であり、現時点であっても学部学生が翻刻をする際の支援には十分であると考える。

深層学習による文字認識の性能を向上させるには、より多くの学習データを利用できることが望ましいことから、データ整備を継続する。また、単一の文字を認識するだけでは台湾総督府文書を翻刻しただけであり、文書群を理解したとは言えない。自然言語処理による文字認識誤り訂正や、単語間の関係や文書間の関係の抽出など、情報技術を利用した文書群の理解を目標に文系研究者との共同を継続したいと考える。

参考文献

[1] THE MNIST DATABASE of handwritten digits, http://yann.lecun.com/exdb/mnist/

[2] Sara Sabour, Nicholas Frosst, Geoffrey E Hinton, Dynamic Routing Between Capsules, NIPS 2017, https://arxiv.org/abs/1710.09829, 2017

[3] ETL 文字データベース，http://etlcdb.db.aist.go.jp/

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CVPR 2015, https://arxiv.org/abs/1409.4842, 2015