画像と非画像データの共起的学習のための深層ニューラルネットワークの開発

Deep Neural Network for Co-occurence Learning of Image and Non-image Data

2171020

研究代表者　東京電機大学　理工学部　准教授　日　髙　章　理

［研究の目的］

本研究では，画像データと共起的に発生する他形式のデータ（例えば音声）を相補的に学習する深層学習手法を開発することを目的とする。第一に，特徴ベクトルの配列変換に基づく非画像データに対する畳み込みニューラルネットワーク（Convolutional Neural Network : CNN）の開発を行った。第二に，画像と共起して発生する非画像データを深層学習に利用するための半自動的データ作成手法の開発を行った。

［研究の内容，成果］

A. 非画像データ識別のための Quadratic CNN

まず，特徴ベクトルの配列変換に基づく非画像データに対する CNN の開発を行った。提案法の全体像を図 1 に示す。まず，Kクラス識別問題のために用意された任意のD次元特徴ベクトル$\mathbf{x}$（図ではD=21）が与えられたとき，末尾に定数項 1 を付け加えた拡張ベクトル $\hat{\mathbf{x}}$ を取る。次に，$\hat{\mathbf{x}}$ 自身のテンソル積を取り，D+1行D+1列の特徴量行列Xを作る。この行列を$(D+1)\times(D+1)$ ピクセルの画像行列とみなし，図 1 右半分に示される一般的な構造の CNN への入力とした。入力ベクトル$\mathbf{x}$が学習用データであるとき，$\mathbf{x}$に対応する教師ベクトル$\mathbf{y}$が用意される。これはK次元の One-hot ベクトル（K個の成分のうちいずれか 1 個のみが 1 で残りが 0 となるベクトル）であり，$\mathbf{x}$がk番目のクラスに属するとき，$\mathbf{y}$の第k番目の成分が 1 となる。図 1 の CNN は，$\mathbf{x}$を変形して得たXを入力として$\mathbf{x}$の教師ベクトル$\mathbf{y}$（すなわち$\mathbf{x}$の所属クラス）を予測しようとする写像となる。

本研究では，CNN のネットワーク構造は，用いるデータによらず一律に図 2 の形とした。すなわち，入力層の後に畳み込み層と活性化層のペア（＝Convolutional block）が 2 つ続き，その後に全結合層と出力層が続く形となる。各層のパラメータ（特徴マップのサイズとチャネル数）は図 2 のブロック中の計算式によって自

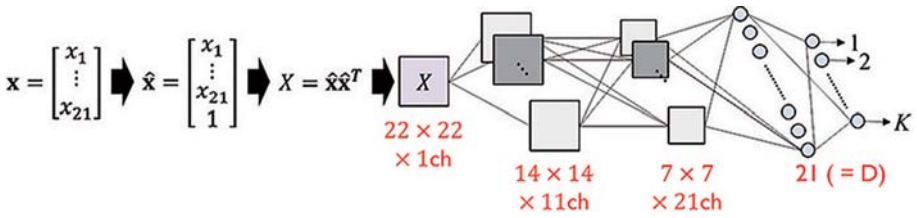

図 1　D＝21 次元特徴ベクトルに対する K クラス識別用 CNN（Quadratic CNN）

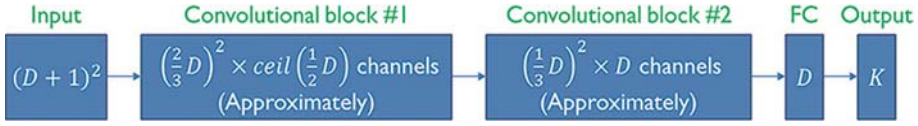

図２　提案 CNN の構造とパラメータ決定式

動決定した。

テンソル積で得られる特徴量行列Xの各成分X_{ij}は，元の特徴ベクトル$\mathbf{x}$の要素x_iとx_jの積x_ix_jとなる。また，定数項１を付加した拡張ベクトルを用いることにより，Xには特徴ベクトル$\mathbf{x}$の元のままの要素もすべて含まれている。このようにXは特徴ベクトル$\mathbf{x}$のすべての１次項と２次項を含んだ特徴表現であることから，提案する CNN を Quadratic CNN（Q-CNN）と呼ぶ。

Xの各成分は互いにスケールが不統一となるため，安定した識別のため各成分を正規化する。ここでは次の３つの正規化手法を比較した。

（N1）$X_{ij} \leftarrow \mathrm{sign}(X_{ij})\sqrt{|X_{ij}|}$

（N2）$X_{ij} \leftarrow \dfrac{X_{ij}}{\|X\|_2+\varepsilon} = \dfrac{X_{ij}}{\sum_{i,j}\sqrt{X_{ij}}+\varepsilon}$

（N3）$X_{ij} \leftarrow \dfrac{X_{ij}}{\|\hat{\mathbf{x}}\|_2+\varepsilon}$

ここでX_{ij}は行列Xの第i行j列の要素，εは微小な正の定数である。（N1）の符号付きで平方根を取る方法と，（N2）の行列 L2 ノルムによる正規化法は，行列状の特徴量に対する既存手法である。また，（N3）のベクトル L2 ノルムで各要素を正規化する方法は，本研究の提案手法である。

表１に本研究で用いたKクラス識別問題の標準的なベンチマークデータの諸元を示す。各データにおいて，全データのうち 2/3 を学習用，残りの 1/3 をテスト用に用いた。学習用データとテスト用データの分割は無作為に行い，異なる分割で５試行の学習・テスト実験を行った。

表２に，従来の識別器であるカーネルサポートベクターマシン（Kernel SVM）および多層パーセプトロン（MLP）と提案法（Q-CNN）

表１　データの諸元

データ ID・データ名	データ数	D	K
1. Segment	2,310	19	7
2. DNA	2,586	180	3
3. Splice	3,175	60	2
4. Waveform	5,000	21	3
5. Satimage	5,104	36	6
6. HTRU2	16,259	8	2
7. Bank additional	41,188	20	2
8. PUC-Rio	165,633	18	5

出典：UCI ML Repository

表２　テスト識別率（５試行の平均値）

ID	Kernel SVM	MLP	Q-CNN (N1&N2)	Q-CNN (N3)
1	96.7 %	91.9 %	93.9 %	96.0 %
2	94.9 %	92.2 %	52.0 %	93.8 %
3	91.2 %	84.0 %	89.1 %	93.3 %
4	86.6 %	83.6 %	84.3 %	85.9 %
5	91.7 %	87.2 %	88.3 %	88.2 %
6	97.9 %	96.9 %	90.9 %	97.5 %
7	90.9 %	88.7 %	88.7 %	90.9 %
8	--.- %	28.6 %	92.5 %	99.2 %

の識別率を示す。提案 CNN では上記の正規化（N3）を用いたときが最も性能がよく，従来の最高レベルの識別器であるカーネル SVM と同等の認識精度を示した。なお，データ８（PUC-Rio）についてはカーネル SVM の学習が著しく遅く，４万秒経過しても学習が終了しなかったため，その時点で学習を打ち切った。

カーネル SVM の最適化処理は並列化が難しいため，その計算時間は主に CPU のシングルコア演算性能に依存する。一方，CNN の最適化は高度に並列化が可能であり，GPU のような並列計算ユニットによる高速化の恩恵を受けることができる。表３にカーネル SVM を Intel 製 CPU（Core i7-5930），Q-CNN を Nvidia 製 GPU（Geforce GTX1080）で学習した際の計算時間を示す。データ数が多く，かつ次元数が多くないデータに対しては，カーネル SVM と比

表３　１試行あたりの学習時間（秒）

ID	Kernel SVM on CPU	Q-CNN（N3）on GPU
1	61[s] : 96.7 %	79[s] : 96.0 %
3	134[s] : 91.2 %	205[s] : 93.3 %
5	239[s] : 91.7 %	153[s] : 88.2 %
6	269[s] : 98.0 %	110[s] : 97.5 %
7	3,946[s] : 90.9 %	308[s] : 90.9 %
8	＞40,000[s] : --.- %	658[s] : 99.2 %

べて Q-CNN の学習時間の方が著しく高速となった。このことから，提案法はビッグデータに対するスケーラビリティを有すると言える。一方，提案法は特徴次元Dに対して$O(D^2)$ の計算量となるため，次元が多い場合には特徴量削減などの工夫が必要となる。

B. 画像と音声の共起的学習のための半自動的データ生成手法の開発

深層学習研究の隆盛により，ここ数年で機械学習に基づく画像認識や音声認識の技術が著しく発達している。人の発話音声から単語や文章を自動認識する音声認識問題では，認識精度が既に実用水準に達しており，Google，Microsoft，IBM などがクラウド型の商用サービスを提供している。また，動画中における顔検出についても，OpenFace などのオープンソースアプリケーションによって高精度で安定的な動作が可能になってきている。

一方で，これらの技術は機械学習によって実現されるものであるため，高い認識性能を得るためには学習用の教師情報付きデータを手動で収集する必要がある。例えば，英語などのプレーンテキストを人工音声として生成する Text to Speech というタスクでは，学習用データとして「文章テキスト」と「それを人間が読み上げた音声ファイル」のペアが大量に必要となる。実用的な音声品質を得るためには，少なくとも延べ数十万単語（数十時間）にのぼる学習用データが必要となり，データ作成に多大なコストが掛かることとなる。また，発話時の唇の動きから発話テキストを予測する Lip reading（読唇テキスト認識）のタスクにおいても，深層学習に基づく認識手法が高い性能を示しているが，その学習には話者の口元を映した画像と，それに対応する正解情報（発話テキスト）のペアが多数必要となる。

他方，既存の音声認識サービスや顔認識アプリケーションが既に高い認識性能を実現していることから，それらの認識結果を Text to Speech や Lip reading などのタスクの学習に利用することが出来れば，データ作成のコストを著しく削減できるものと期待される。そこで本研究では，取得が容易な生データ（ネット上の動画等）に既存の音声認識手法および顔認識手法（可能な限り低工数で済むもの）を適用し，それぞれの認識結果と元データ（動画）とを時間的に同期させて得られるデータセットを他の深層学習タスクに利用する枠組みを開発する。ここでは，図３に示すように半自動的に作成した学習用データを音声合成および読唇テキスト認識の２つのタスクの学習に用いた。

B1. 半自動作成データによる音声生成の学習

ここでは既存の音声認識サービスで動画音声を認識させ，その出力（タイムスタンプ付きの発話単語列）から発話時の音声を再構成する深層 NN の学習について述べる。音声認識サービスには Google のクラウドサービスである Google Clould Speech API を用いた。また，Text to Speech（すなわちテキストからその読み上げ音声を生成するタスク）を行う深層 NN

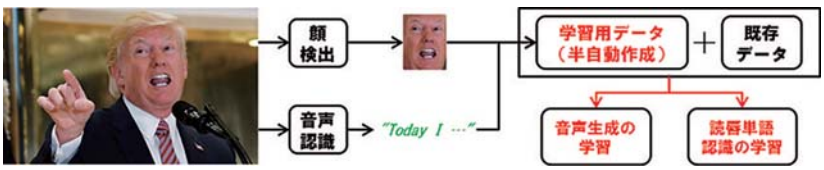

図３　取得が容易な生データからの低工数の学習用データ作成

には，Deepvoice3 法を用いた。

Text to Speech タスクのための学習用データとしていくつかのデータセットが公開されており，本研究ではベースとして LJSpeech Dataset を用いた。LJSpeech Dataset は，7 つのノンフィクションの本から一節を読む 1 人の女性話者の短いオーディオクリップからなるスピーチデータセットである。その音声ファイル数は 13,100 本，総単語数は 225,715 語，総文字数は 1,308,678 文字，合計再生時間は 23 時間 55 分 17 秒である。

本研究では，この LJSpeech Dataset に加えて図 3 の工程で得られた自作の学習データを用いて DeepVoice3 の学習を行い，どの程度の合成音声品質が得られるのかを検討した。ここでは自作データの元動画として，Youtube 上で入手できる第 45 代米国大統領 Donald John Trump のスピーチ動画 5 本を用いた。それらの動画を Google Clould Speech API に入力し，Trump の発話内容（タイムスタンプ付きの発話単語列）を出力させ，任意の連続する W 単語（W＝1, 2, 3, 4, 5）の開始時刻と終了時刻に対応する時間区間の音声を元動画から切り出し，延べ 3，785 個の短時間音声ファイル（最短 0.45 秒，最長 35.03 秒）を作成した。このとき，1〜5 語の単語列が DeepVoice3 への入力（読み上げさせたいテキスト）となり，対応する音声ファイルが教師情報（生成目標）となる。

表 4 に，ベンチマークに用いる LJSpeech データセットおよび自作データセットの諸元を示す。実験では，DeepVoice3 を LJSpeech データのみで学習した場合と，自作データを加えた統合データで学習した場合で，音声生成の結果がどのように変化するかを確認した。LJSpeech のみで学習した場合，当然ながら生成される音声は読み上げを行った女性の声のみとなるが，統合データで学習した場合，生成結果のうち一部は明瞭に Trump 大統領の声質に聞こえる音声となった。また，声質は女性のものでありながら，イントネーションが Trump 風に変化したように聞こえる生成結果も見られた。このことは，特定人物の発話映像を無造作に集めただけのデータから極めて低工数・低コストで半自動的に作成した学習用データにより，その人物の声質でテキスト読み上げを行う深層 NN を学習することが十分に可能であることを意味する。

表 4　Text to Speech 学習用データの諸元

	LJSpeech	自作	統合
話　者	女性 1 名	男性 1 名（Trump）	2 名（男女 1 名）
データ数	13,100	3,785	16,885
単 語 数	225,715	3,805	229,520
合計時間	23 : 55 : 17	03 : 08 : 32	27 : 03 : 49
最長時間	10.10 秒	35.03 秒	35.03 秒
最短時間	1.11 秒	0.45 秒	0.45 秒

B2. 半自動作成データによる読唇認識の学習

ここでは，先ほどの音声生成で用いた Trump の動画から OpenFace 法によって抽出した顔画像を用い，唇周辺の動き情報からそのときの発話内容（単語列）を予測する再帰型ニューラルネットワーク（Reccurent Neural Network : RNN）の学習を行った結果について述べる。OpenFace 法は動画の各フレームに含まれる顔を検出し，検出された顔の正規化画像や顔パーツの Landmark（特徴点）座標などを自動出力する。本研究の場合，Google Clould Speech API の認識結果に含まれる各単語の発話時刻と OpenFace の認識結果のフレーム時刻を対応させるだけで，「ある単語を発声している最中の Trump の顔特徴点座標の時系列」（すなわち顔パーツの動き情報）を取り出すことが出来る。ここでは音声認識の結果で出現頻度が多かった 4 単語（and, the, to, we）について，その単語を発声している際の Trump の顔特徴点座標時系列のみからどの単語を発話しているかを識別する学習を行った。

ここでは認識モデルとして，時系列を扱う RNN の一種である Bidirectional Long-Short Term Memory（Bi-LSTM）法を用いた。各単語それぞれ無策に選んだ 15 サンプルをテスト用とし，残りを学習用とした。表 5 にテストサ

表 5　Bi-LSTM の単語識別率

単語	and	the	to	we
出現数	124	158	162	80
訓練数	109	143	147	65
テスト数	15	15	15	15
識別率	67.3 %	48.0 %	36.7 %	6.0 %

ンプルに対する 10 試行の平均単語識別率を示す。4 クラス問題における期待識別率は 25% であるが，ここでは 4 単語のうち 3 単語で期待値を大きく超える識別率となった。このことは，音声生成の場合と同様に，無造作に収集した発話映像から極めて低工数・低コストで半自動的に作成した学習用データにより，Lip reading の精度向上を図りうることを意味する。

［今後の研究の方向，課題］

本研究では，下図のように動画中で共起する画像と音声の相互変換を実現することを大目標に設定していたが，画像から音声を生成する NN モデルのアーキテクチャの開発に難航し，残念ながら実現には至っていない。しかし，こ

図 4　画像と共起する情報の変換学習

こ半年程で NN のアーキテクチャを自動設計・最適化する Neural Architecture Search（NAS）という技術が著しく発達しており，これにより入力を画像，出力を音声とする NN の内部構造を自動的に最適化する道筋が立った。今後はこの方法で大目標の達成を図っていく。

［成果の発表，論文等］

［1］ A. Hidaka and T. Kurita, “Convolutional Neural Networks as General Nonlinear Classifiers”, ISCIE SSS’17, pp. 95-96, Nov. 2017.

［2］ 今井大貴，日高章理，“深層学習による音声生成等の学習データの半自動的作成手法の開発”，第 51 回計測自動制御学会北海道支部学術講演会，pp. 41-42，2019 年 3 月.