7M-02

環境センシングを活用したマルチモーダル学習の提案

橋本　和幸 †　堀川　三好 †　岡本　東 †

† 岩手県立大学ソフトウェア情報学部

1　はじめに

IoT によるセンサネットワークの構築が進む中，様々な分野でビッグデータ分析や人工知能の適用が期待されている．特にスマートフォンによる周辺環境のセンシングは，利用者の行動分析や新たなマーケティング分野への活用が進められている．また，畳み込みニューラルネットワーク（CNN:Convolutional Neural Network）を用いた画像認識はモビリティ，ものづくり，医療・福祉など幅広い分野へ導入されている．今後，様々なドメインを統合的に処理できる人工知能の登場と，それによる高度なデータ分析が期待されている．

本研究は画像認識の高性能化を目的とし，環境センシングと CNN を組み合わせたマルチモーダル学習を提案する．本稿では環境センシングデータとして BLE(Bluetooth Low Energy) ビーコンの電波受信強度 (RSSI) を用いる．すなわち，特定空間において複数の BLE ビーコンを設置し，物体撮影時に画像と RSSI を収集してデータセットを作成する．また，収集したデータセットを，3 つのマルチモーダル CNN と RSSI を与えずに学習を行う従来の CNN で評価を行い，その結果について報告する．

2　関連研究

マルチモーダル学習の先行研究は教師あり学習と教師なし学習に大別される．教師なし学習においては Ngiam らのマルチモーダル自己符号化器 [1] などが挙げられる．画像とタグ情報を観測変数の一つとして各モデルへ適用し，得られた潜在変数を線形回帰モデルやサポートベクターマシンを用いて分類している．教師あり学習では，特に CNN において，Lin らはマルチモーダル CNN[2] を提案している．Ngiam らと同様に画像とタグ情報を観測変数とし，提案したモデルを用いて画像認識を行っている．

Multimodal Learning Using Environment Sensing
†Faculty of Software and Information Science, Iwate Prefectural University

3　画像認識モデルの提案

マルチモーダル CNN について 3 つの画像認識モデルを提案する．入力層で各 RSSI に重みをつける mCNN-w，入力層でテンソルの連結を行う mCNN-c，そして全結合層の前に連結を行う mCNN-f を提案し，以下に概要を述べる．提案するモデルの構造を図 1 に示す．

3.1　共通の表記

以下で共通する表記について述べる．画像を表すテンソルを x_{ijk}，センサの値を表すテンソルを s_l とする．(i, j, k) は順に（縦，横，チャンネル数）を表し，l は観測したセンサの数を表す．また，ある正整数 n において x^n_{ijk} と s^n_l は同時に得られたサンプルである．X は x_{ijk}, s_l をそれぞれ操作し得られるテンソルを表す．また，f, f' はニューラルネットを θ はそのパラメータとし，$f_\theta(x_{ijk}, s_l)$ は入力 x_{ijk}, s_l に対して各クラスに属する確率を返すものとする．

3.2　mCNN-w

入力層でセンサ値に重みをつけ，画像との和を学習させる．式 (1) の w は各々の s_l に対する重みである，確率的勾配降下法によって w, θ を決定する．

$$X = x_{ijk} + \sum_{m=0}^{l} w_m \times s_m \tag{1}$$

$$f_\theta(x_{ijk}, s_l) = f'_\theta(X) \tag{2}$$

3.3　mCNN-c

入力層で各センサ値を画像と同じ縦横比に整形する．これらと画像を連結した後のテンソルを学習させる．式 (3) の P_{ij} は x_{ijk} と同じ縦横比かつすべての要素が 1 のテンソルであり，式 (4) の $\|$ は各テンソルの連結 (concatenate) を表す．θ は確率的勾配降下法を用いて決定する．

$$X = x_{ijk} \| P_{ij} \times s_0 \| P_{ij} \times s_1 \| \ldots \| P_{ij} \times s_l \tag{3}$$

$$f_\theta(x_{ijk}, s_l) = f'_\theta(X) \tag{4}$$

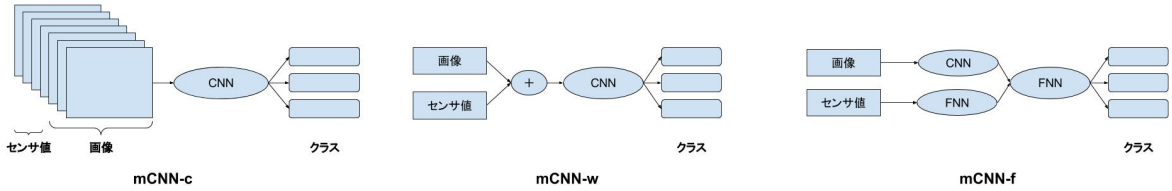

図 1: 3 つの提案する画像認識モデル

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.

3.4 mCNN-f

一般に CNN では出力層の前でテンソルを変形してランクを 1 とし，全結合層を用いる．本手法では画像をもとに CNN を用い，その後ランク 1 にしたテンソルと，センサ値をもとに複数の全結合層を重ねたランク 1 のテンソルを連結し，再び全結合層を重ねる．式 (5) の f'', f''' は中間層を表す．確率的勾配降下法を用いて θ を決定する．

$$f_\theta(x, s) = f'_{\theta_0}\left(f''_{\theta_1}(x) \,||\, f'''_{\theta_2}(s)\right) \tag{5}$$

4 評価実験

4.1 実験目的

提案する画像認識モデルとセンサ値を与えないモデルを評価するため，同一のデータセットを用いて評価実験を行う．岩手県立大学構内に BLE ビーコンを設置し備え付けの消火器・ゴミ箱をどこのものであるかを分類し，その精度を確かめる（図 2）．

4.2 実験方法

データセットの内訳を表 1 に示す．データセットは，ビーコンの設置された環境でセンシングと撮影した後，正規化などの前処理を行う．RSSI のみに分類が依存しないよう，オブジェクトは同じ位置から撮影できるよう選ぶ．また，Android 端末で 1 秒間の RSSI の平均と，同時に撮影した写真の画像を保存する．また，画像は縦横の最大長に合うようゼロパディングし，縦横ともに 640 ピクセルに縮小し，各ピクセルを正規化する．同様に RSSI も正規化する．

作成したデータセットをもとに，各モデルで 2000 回学習を行い精度を比較する．また，各回でテストデータを使い，精度を算出する．

表 1: データセット内訳

		エリア		
		A	B	C
物体	ゴミ箱	757	848	805
	消火器	785	696	914

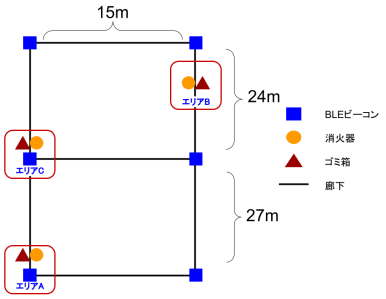

図 2: データ収集を行った屋内環境

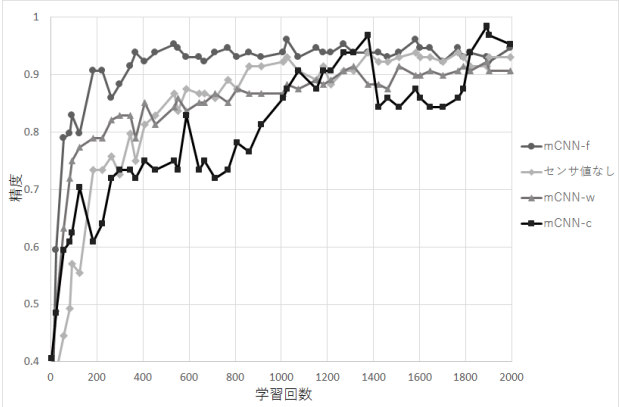

図 3: 精度の比較

4.3 実験結果

各モデルを比較した結果を図 3 に示す．各モデルの最高精度は mCNN-c，mCNN-f，ビーコンなしの順に高い．最も早く 95%以上の精度を出したのは mCNN-f である．mCNN-w は常にセンサ値なしのモデルを下回る結果である．

4.4 考察

mCNN-w は推論を行う際，重みが定数として計算されるため，ここで付けた重みはセンサ値の特徴を表現することはできない．一方で mCNN-f はセンサ値の入力層と連結との間に全結合層を挟むことによって，その特徴を表現し，その後の全結合層もうまくパラメータを調整出来たと考えられる．すなわち，本手法によって学習が容易になったと捉えることができる．これは識別するクラスが増えた際に，より効果が示されるのではないかと考えられる．

5 おわりに

本稿ではセンサ値を用いたマルチモーダル CNN のモデルを提案した．そして実環境で得られたデータセットを用いて比較検証を行い，提案したモデルの一つ，mCNN-f の優位性を示した．今後はより複雑なデータセットなどで，モデルの実用性について検証を続けたい．

参考文献

[1] Ngiam, J., Khosla, A., Kim, M., Nam, et al.: Multimodal deep learning, in Proceedings of the 28th international conference on machine learning (ICML), pp. 689-696 (2011)

[2] Lin, M., Zhengdong, L., Lifeng, S., et al.: Multimodal Convolutional Neural Networks for Matching Image and Sentence, in International Conference on Computer Vision (ICCV) 2015

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.