The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

マルチモーダル深層学習の神経表現と算術計算に対する有効性の検討

Effective Neural Representation for Arithmetic Tasks Induced by Multi-modal Deep Learning Models

野田　寛眞[*1]
Noda Kamma

宗田　卓史[*1, 2]
Soda Takafumi

山下　祐一[*1]
Yamashita Yuichi

[*1] 国立精神・神経医療研究センター
National Center of Neurology and Psychiatry

[*2] 東京医科歯科大学
Tokyo Medical and Dental University

The associations of multimodal information are essential for human cognitive functions. Recently, multimodal learning has received a lot of attention in the field of machine learning. Investigating the impact of multimodal learning on its representation could facilitate our understanding of multimodal associative learning in humans. Here, a multimodal deep learning model was used as a computational model of multimodal associations in the brain. Representations of numerical information, such as handwritten numbers and images of geometric figures, were learnt and compared using single-modal and multimodal learning. Multimodal learning models acquired better latent representations. Furthermore, the models trained using multimodal method performed better on the downstream arithmetic task. These results showed that changes in latent representations acquired during multimodal associative learning were directly related to cognitive function. This supports the potential of multimodal learning research to provide new insights into the understanding of higher cognitive functions in humans, including mathematical abilities.

1. はじめに

視覚，聴覚，触覚など様々なモダリティ情報を連携し処理する能力は，言語や意味記憶をはじめとしたヒトの高次認知機能に不可欠である[Büchel 98, Kraut 02]。一方，その失調は，意味性認知症，計算障害，相貌失認など，様々な神経心理学的症状につながると考えられている。例えば，共感覚は，多感覚統合の変調の一例であり，文字にある色が見えるなど，ある感覚刺激が他の感覚を同時に呼び起こすクロスモーダルな現象である。しかしながら，モダリティ間にわたる情報処理と，高次認知機能や神経心理学的症状との間に，如何なるメカニズムが存在するのかは，十分に解明されていない。

マルチモーダルな情報処理は，機械学習の領域でも研究されてきた。一例として，言語，画像，音などの複数の情報モダリティを利用することで，モデルの性能が向上することが報告されている[Shi 19]。その中でも深層学習モデルは，神経回路に着想を得ており，ヒトの脳との類似性が示唆されている[Yang 20]。そのため，深層学習モデルのマルチモーダル学習の理解は，ヒトの記憶，意味処理，計算能力などの高次認知機能や，神経心理学的症状の原理解明に貢献すると考えられる。

深層学習モデルの優れた性能には，マルチモーダル学習によって獲得された情報表現が鍵となっている可能性が指摘されている[Guo 19]。そこで，本研究では，多様なモダリティの情報を処理する脳の計算モデルとしてマルチモーダル深層学習モデルを用いることで，マルチモーダル学習が情報表現に与える影響，並びに，認知課題の成績に与える影響を明らかにすることを目的とした。実験では，深層学習モデルは，数感覚に関連した情報を，手書きの数字や幾何学図形の画像などの複数のモダリティを用いて学習した。学習の結果獲得された潜在表現の特性を解析し，算術計算タスクのパフォーマンスに与える影響を調べた。

連絡先：山下祐一，国立精神・神経医療研究センター，東京都小平市小川東町 4-1-1, yamay@ncnp.go.jp

2. 方法

2.1 実験設定

本実験では，マルチモーダル深層学習モデルによる想起的生成と連想的生成の 2 つを利用した[Noda 22]。想起的生成は，ある単一モダリティ内で入力データをそのまま再構成する処理である。一方，連想的生成は，入力と同じ情報を有するが異なるモダリティのデータの生成を行うクロスモーダルな処理である。例えば，連想的生成では，シンボルとして表現された“3”という入力情報に対して“3”個のオブジェクトが含まれた画像を生成する。本研究では，想起的生成と連想的生成を同時に行うことを，ヒトの認知過程におけるマルチモーダルな情報処理と見做し，この 2 つの方略を同時に行う学習と，想起的生成のみを行う学習が，潜在表現の獲得に与える影響を比較した。

本実験のために，オブジェクト個数による感覚的情報と，数に関するシンボル的情報という異なる情報モダリティを含む OSCN-CMNIST と呼ばれるデータセットを構築した（図 1）。数の概念は，「6 つの正方形の絵」というオブジェクト画像や，「6 というシンボル」で表現される場合がある。OSCN（object-shape-color-number dataset）は，オブジェクトの配置，オブジェクトの色，オブジェクトの形，オブジェクトの数の 4 つの要素で構成された画像のデータセットであり，各画像は，オブジェクトの個数を用いて数の概念を表現している。一方，CMNIST（Colored Modified National Institute of Standards and Technology database）

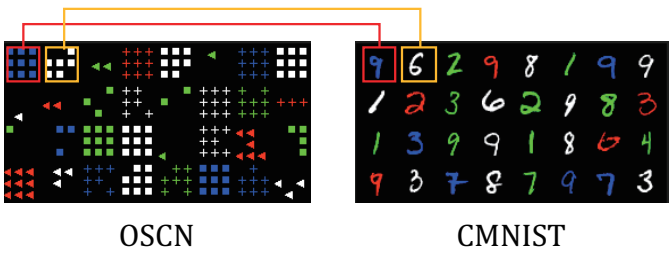

図 1. OSCN-CMNIST データセット

は，手書き数字画像のデータセットである MNIST に色をつけたものである。このデータセットでは，象徴的なアラビア数字を用いて数の概念を表現している。OSCN-CMNIST は，OSCN と CMNIST の両画像が，同一の数字と色を表現するよう画像がペアになる一方で，形状やレイアウトはランダムに対応づくように構成されている。

2.2 深層学習モデル

想起的生成と連想的生成を行う深層学習モデルとして，Mixture-of-experts multimodal variational autoencoder（MMVAE）を用いた [Shi 19]。MMVAE は，マルチモーダル学習のために提案された深層学習モデルであり，高い生成力を有する。加えて，学習したモダリティの共通表現が獲得されるため，解釈が容易である。

マルチモーダルモデルA^{Multi}では，同一の情報を有するがモダリティの異なる 2 つのデータのペア(x^{M_1}, x^{M_2})が入力となる。エンコーダE_iは潜在変数z^{M_i}を生成する$(i = 1, 2)$。一方，各デコーダ$D_j (j = 1, 2)$は，各潜在変数 z^{M_i}に対して出力$\hat{x}_j^{M_i}$を生成する。このように，MMVAE は，想起的生成（再構成）と連想的生成（クロスモーダル生成）を同時に行う。潜在変数がどのエンコーダによって推論されたかに関わらず，推論された潜在変数を用いて同様のデータを生成するようにモデルが学習されるため，複数のモダリティ間の潜在変数は重なり合うように学習される。その結果，モダリティ間に共有した情報は共通の部分空間で表現され，モダリティ間に独立した情報は異なる部分空間で表現されることが期待される。一方，シングルモーダルモデルA_1^{Single}，A_2^{Single}の場合，想起的生成（再構成）のみを行う。この場合，各モデルには，単一のモダリティ情報のみが入力として割り当てられるため，独立して学習される。そのため，モダリティ間の関係を学習することは不可能であり，複数のモダリティの潜在表現が自発的に重なり合うことは期待されない。

3. 結果

3.1 生成画像

まず，OSCN-CMNIST を用いて学習したモデルを用いて，想起的生成によって作成された画像と，連想的生成によって作成された画像を検討した。連想的生成の場合，若干の誤りが含まれたものの，出力された画像は明瞭かつ正確であり，どちらの学習条件でも学習に成功したと考えられた。

3.2 潜在表現

学習方法の差異による潜在表現の違いを分析するために，まず OSCN のテストデータを MMVAE のエンコーダに入力し，OSCN の潜在変数（20 次元）を取得した。この潜在変数を t-SNE によって圧縮し，数字ラベルごとに色付けした上で 2 次元空間にプロットした（図 2）。マルチモーダル学習では，数字ラベルごとにクラスタを形成しており，数字ラベルが混在したクラスタは減少した。そのため，数字ラベルによる潜在空間の分離可能性は，シングルモーダル学習と比較し，マルチモーダル学習が高かった。加えて，マルチモーダル学習の場合，数字ラベルの大小に比例するように，各数値ラベルは潜在空間上に配置されていた。これは，このモデルが，数の大きさの概念を潜在表現として獲得していることを示唆している。

同様の手法により，CMNIST のデータを入力としたときの潜在表現も検討した（図 2）。CMNIST のみを用いるシングルモー

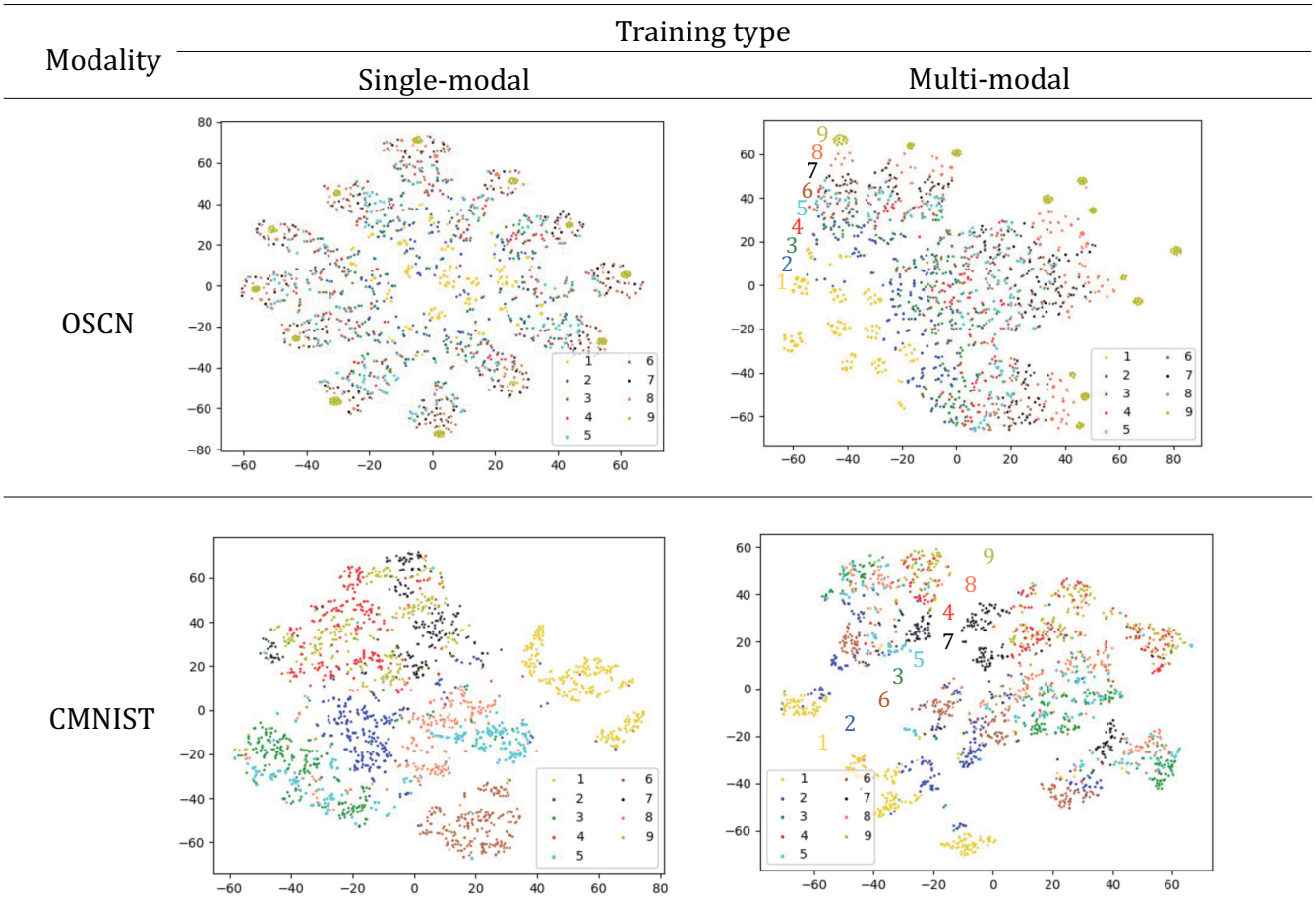

図 2. 潜在表現の可視化

ダル学習では，潜在空間における順序構造は観察されなかった。これは，CMNIST を用いたシングルモーダル学習の場合，入力データは大きさの情報を有さないためだと考えられる。一方，マルチモーダル学習では，オブジェクト個数による感覚的情報がCMNISTに含まれないにも関わらず，対応する数字ラベルの大きさに比例した潜在空間の構造が観察された。これは，MMVAEが，OSCNとCMNISTという2つの情報モダリティを関連付けた結果だと考えられる。

同様の手順を用いて，色ラベルや形ラベルについても潜在表現の可視化を行なった。OSCN に関して，シングルモーダル学習とマルチモーダル学習の両者においてクラスタ化された潜在表現の存在が伺えた。一方で，マルチモーダルモデルでは，色に基づくクラスタリングが行われた上で，形状に基づく分割が為されるという階層的な潜在空間の構造が示唆された。CMNIST について，色ラベルに基づくクラスタ構造はシングルモーダル学習では観察されなかったが，マルチモーダル学習では観察された。これは，マルチモーダル学習の場合，対になった OSCN と CMNIST の色を共通の潜在表現として獲得したためだと考えられる。

3.3 算術計算課題

潜在表現の検討により，クラスタ構造や順序構造など数感覚に重要となる潜在表現の獲得に対して，マルチモーダル学習が有効であることが分かった。そこで，この潜在表現が，数感覚を要する下流認知タスクである算術計算に有用であるかどうかを分析した。学習した潜在表現が，入力データの量的特性を正しく反映している場合，その潜在表現を用いて加算や減算を行うことができると考えられる。例えば，“1”，“9”，“8”の OSCN 画像をエンコーダ E_{OSCN} に与えることで，それぞれの潜在変数 $z_{OSCN}(1), z_{OSCN}(9), z_{OSCN}(8)$が得られる。その後，これらの潜在変数を用いて加算・減算を行い，その結果をデコーダ D_{OSCN} に与え画像を生成する。図3は，D_{OSCN}に $z(1)+z(9)-z(8)$を与えた場合の出力である。このように，いくつかの誤りは存在するものの，“2”を示す正しいオブジェクト画像を出力できた。

4. 議論

本研究では，数に関するシンボル的情報とオブジェクト的情報を利用し，マルチモーダル学習が潜在表現の獲得に与える影響を検討した。その結果，マルチモーダル学習では，色ラベルや数字ラベルなど，2つのデータセット間の共通性を反映した，優れた潜在表現を獲得できた。これは，深層学習モデルにおけるマルチモーダル学習の有用性を支持するこれまでの研究と一貫している[Guo 19, Suzuki 16]。さらに，OSCN と CMNIST を組み合わせたマルチモーダル学習は，順序構造やクラスタ構造など，数感覚に関連した潜在表現の獲得にも有益であることがわかった。これは，CMNIST が有するシンボル的情報と，OSCN が有するオブジェクト画像の感覚情報との連合に起因すると考えられる。さらに，マルチモーダルモデルによって獲得された潜在表現は，算術計算課題においても有益であった。このように，マルチモーダル学習は，情報の潜在表現の学習と，関連する下流タスクの成績に重要であった。

今後は，大規模な実世界データを用いた実験や，大規模モデルを用いた研究が必要である。より複雑な認知スキルの再現に成功すれば，深層学習モデルにおける計算メカニズムと，ヒトの脳における生物学的現象との対応関係が明らかになる可能性がある。これは，ヒトの脳や高次認知の理解につながるだけでなく，マルチモーダル処理に関連する疾患の病態解明に貢献する可能性がある。

$$z(1)+z(9)-z(8)$$

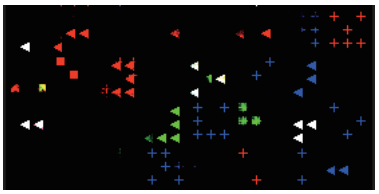

図 3. 算術課題の結果

参考文献

[Büchel 98] Büchel, C., Price, C., & Friston, K.: A multimodal language region in the ventral visual pathway, *Nature*, Vol. 394, No. 6690, pp. 274–277 (1998)

[Guo 19] Guo, W., Wang, J. and Wang, S.: Deep multimodal representation learning: a survey, *IEEE Access*, Vol. 7, pp. 63373–63394 (2019)

[Kraut 02] Kraut, M. A., Kremen, S., Moo, L. R., Segal, J. B., Calhoun, V. and Hart Jr., J.: Object activation in semantic memory from visual multimodal feature input, *Journal of Cognitive Neuroscience*, Vol. 14, No. 1, pp. 37–47 (2002)

[Noda 22] Noda, K., Soda, T. and Yamashita, Y.: Emergence of Number Sense in Deep Multi-modal Neural Networks, *PsyArXiv*, (2022)

[Shi 19] Shi, Y., Siddharth, N., Paige, B. and Torr, P. H. S.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. *In Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Curran Associates Inc, 32 (2019)

[Suzuki 16] Suzuki, M., Nakayama, K. and Matsuo, Y.: Joint Multimodal Learning with Deep Generative Models, arXiv:1611.01891 (2016)

[Yang 20] Yang, G. R. and Wang, X. J.: Artificial neural networks for neuroscientists: a primer. *Neuron*, Vol. 107, No. 6, pp. 1048–1070 (2020)