Brain-inspired multimodal learning based on neural networks

Chang Liu  
*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Fuchun Sun  
*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Bo Zhang  
*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Follow this and additional works at: [https://tsinghuauniversitypress.researchcommons.org/brain-science-advances](https://tsinghuauniversitypress.researchcommons.org/brain-science-advances)

Part of the Biomedical Engineering and Bioengineering Commons, Nervous System Diseases Commons, Neurology Commons, Neuroscience and Neurobiology Commons, Neurosciences Commons, and the Neurosurgery Commons

**Recommended Citation**  

This Research Article is brought to you for free and open access by Tsinghua University Press: Journals Publishing. It has been accepted for inclusion in Brain Science Advances by an authorized editor of Tsinghua University Press: Journals Publishing.
Brain-inspired multimodal learning based on neural networks

Chang Liu, Fuchun Sun (✉), Bo Zhang

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ARTICLE INFO
Received: 15 July 2018
Revised: 06 August 2018
Accepted: 10 August 2018

© The authors 2018. This article is published with open access at journals.sagepub.com/home/BSA

Creative Commons NonCommercial-NoDerivs CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License (http://www.creativecommons.org/licenses/by-nc-nd/4.0/) which permits non-commercial use, reproduction and distribution of the work as published without adaptation or alteration, without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

KEYWORDS
multimodal learning, brain-inspired learning, deep learning, neural networks

ABSTRACT
Modern computational models have leveraged biological advances in human brain research. This study addresses the problem of multimodal learning with the help of brain-inspired models. Specifically, a unified multimodal learning architecture is proposed based on deep neural networks, which are inspired by the biology of the visual cortex of the human brain. This unified framework is validated by two practical multimodal learning tasks: image captioning, involving visual and natural language signals, and visual-haptic fusion, involving haptic and visual signals. Extensive experiments are conducted under the framework, and competitive results are achieved.

1 Introduction
Research in multimodal learning has accelerated in recent decades, aiming to bridge gaps when learning from different modalities in order to mine their correlations and make better decisions resulting from their fusion. A common research scheme in multimodal learning is to learn the characteristics of each modality and then combine all the characteristics to inform a final decision. Thus, the premise of multimodal learning relies on the design of learning models for each modality, after which a fusion model must be built to arrive at a final conclusion. For example, image captioning is a task that describes a given image with a sentence. It involves two modalities, the sight of a given image, and natural language, which is the descriptive sentence. Another example is the visual and haptic fusion common in the robotics literature. This is a common task where a robot must make decisions based on visual information, such as images/videos taken by a camera, and haptic information, such as tactile readings from sensors equipped at the end of robot hands.

Address correspondence to Fuchun Sun, E-mail: fcsun@mail.tsinghua.edu.cn
Coping with both modalities gives the robot the ability to make better decisions and improve their function, such as their fine-level grasp of objects. In both of these scenarios, tasks involve two modalities, and thus both belong to the domain of multimodal learning. It is then crucial to design learning algorithms for each modality as well as their fusion.

Brain-inspired learning algorithms have been a leading trend in the learning and comprehension of different modalities. Improved knowledge of the human brain is essential in improving artificial intelligence, and multimodal learning is key. As the understanding of the biology of the human brain improves, models and algorithms centering on multimodal learning have been proposed based on these findings. Such models and algorithms are usually termed brain-inspired learning models. For example, the HMAX model [1] is a biologically inspired vision model based on the simple and complex cells in the visual cortex [2]. Similar vision models also based on such a layered architecture can be found in modern deep neural networks, such as convolutional neural networks [3].

Effectively incorporating brain-inspired learning algorithms for each modality with multimodal learning, which emphasizes the coordination of different modalities, offers interesting research directions. This paper aims at solving multimodal learning problems by leveraging brain-inspired models, primarily deep neural networks. The illustration of the proposed model can be found in Fig. 1, which will be detailed in Section 4. The main contributions of this paper are as follows:

1. Define multimodal learning and provide a unified solution architecture.
2. Based on the unified learning framework, test and validate the model on several single-modality learning problems, then explore multiple modality applications.
3. Further validate the framework on multimodal learning problems. Specifically, we explore the feasibility of the proposed model on image captioning and visual-haptic fusion. The former involves visual and natural language modalities, and the latter involves visual and haptic modalities.

The paper is organized as follows:

Section 2 reviews the literature of brain-inspired neural networks, as well as the applications of multimodal learning. Section 3 presents the basic mathematical models of neural networks, which forms the basis of subsequent chapters. Section 4 presents the unified architecture of multimodal learning based on neural networks. Sections 5

![Figure 1](https://mc03.manuscriptcentral.com/brainsa)

The proposed unified multimodal learning scheme based on brain inspired models. The core of the model is recurrent neural networks, which contains the multimodal inputs at each time step. The multimodal input is a combined feature of each modality. For each modality, the representation is the feature extraction using some neural networks. The final decision is based on the last hidden state of the recurrent neural networks, which contains the information of all the modalities at all time steps.
and 6 focus on two different multimodal learning tasks, image captioning and visual-haptic fusion, respectively, with the purpose of validating the proposed learning algorithms. Section 7 concludes the paper.

## 2 Related works

### 2.1 Visual cortex and models

Research on the visual cortex of the human brain has advanced computer vision research in multiple ways. This section presents inspiring research on the human visual cortex found in modern multimodal learning settings.

**Figure 2** is an illustration of the visual cortex and the HMAX model based on it. The term visual cortex refers to the primary visual cortex (also known as striate cortex or V1) and extrastriate visual cortical areas such as V2, V3, V4, and V5 [4, 5], as shown in Fig. 2(a). These areas form a layered structure, where visual information flows from lower to higher receptive areas. Inspired by this layered architecture in the visual cortex, the HMAX model, a visual recognition model, was proposed in 1999 by Maximilian Riesenhuber and Tomaso Poggio, and is shown in Fig. 2(c). The model mimics simple and complex cells (shown in Fig. 2(b)) in the visual cortex, where the computation units form a layered structure with each layer either simple or complex. With such an architecture, the HMAX model successfully captures the characteristics of biological cells in the visual cortex and performs well in real-world recognition applications. Following this work, many variants of the HMAX model were developed. These aimed to improve the performance of HMAX by introducing additional biological mechanisms or constraints. For example, Hu et al. proposed a Sparse-HMAX model that implemented more layers and sparse coding mechanisms in order to mimic the sparse firing phenomena of visual cortex cells [6]. Dura-Bernal et al. [7] introduced feedback into HMAX, which has been found in the visual cortex [8–11]. More recent computation models based on the biology
of the visual cortex of the human brain are neural networks, which are detailed in the following section.

### 2.3 Machine learning and deep neural networks

Machine learning has shown its effectiveness in many fields, and it has been applied in practical scenarios, such as the application of random forests in Microsoft Kinect [12], in advertising click-through rate prediction, and logistic regression [13]. Traditional machine learning methods are often limited by the original form of natural data [14]. These machine learning techniques generally require the design of appropriate features for specific problems and require significant expert knowledge, system engineering, etc. Cognitively, people do not need to design different feature representations for different tasks, since throughout life, they learn to deal with different tasks from experience. Deep learning is such a “representation learning” method, which is a learning model of more complex multilevel representations by combining simple nonlinear modules [14]. Compared with traditional machine learning methods, deep learning has the advantage of not requiring expert knowledge in a certain field, nor does it need to design features for specific problems. Deep neural networks, as an important research area in deep learning systems, can be considered as a means to deal with modal information by learning hierarchical concepts, with each concept related to other relatively simple concepts. The relationship between the models has been previously described [15]. Common deep neural networks inspired by human brain biology include convolutional neural networks, recurrent neural networks, and deep belief networks.

### 2.4 Multimodal learning

Multimodal learning tasks do not have a standard definition, but can be generally described as observing different phenomena and acquiring data and information from different types of sensors, and then experimenting in different conditions. The information obtained when observing phenomena is termed a “modality” [16]. Multimodal learning refers to feature extraction and integration analysis beyond single-mode data, as well as further decision processes based on these features and analysis [17]. In the context of this paper, multimodal learning refers to multimodal sensor data obtained by analyzing and characterizing the multiple modal sensors of a robot while interacting with its environment, as well as the process of merging these data to make decisions.

Multimodal learning has wider application than single-mode research, and thus tends to have more practical applications. First, the study of learning can supplement data from single-mode research, which to some extent assists decision-making in original single-mode research. Second, some tasks are by nature multimodal tasks. Often, raw data and related decision-making processes involve more than one mode of data, such as tasks involving image understanding. Such a problem essentially involves visual image or video data, as well as natural language data. Since multimodal learning tasks usually involve time-series data, which are different from static modal information such as images, the context of the modal information must be determined in time to further assist final decision making. In summary, multimodal learning research benefits from more modal and temporal information, and some problems are multimodal problems by nature, offering important research directions and practical applications.

### 3 Prerequisites

Prior to describing the proposed unified multimodal learning architecture, this section presents
the prerequisites of the model, providing a basis for understanding the multimodal learning algorithms.

3.1 M-P neuron and multilayer perceptron

Neurons are the most basic unit of deep neural networks, which are derived from neuronal models in neuroscience and brain science. By modelling the working mechanism of neurons in the human brain, McCulloch et al. first proposed a concept of artificial neurons [18], the “M-P” neuron model, as shown in Fig. 3. The M-P model takes in a set of inputs \( x_1, x_2, \ldots, x_n \) and outputs the non-linear activation of the linear summation of the inputs, as shown in Eq. (1).

\[
\begin{align*}
    a &= \sum_i (w_i x_i + b) \\
    y &= f(a)
\end{align*}
\]

where \( f \) is the nonlinear activation function, which is usually one of the following activate functions: sigmoid, tanh, ReLU, or LeakyReLU, as shown in Fig. 4.

The multilayer perceptron (MLP) is a hierarchical feedforward artificial neural network, with each unit node of each layer forming a neuron. Take the simplest three-layer multilayer perceptron as an example, which includes an input layer, a hidden layer, and an output layer. More complex multilayer perceptrons essentially add a hidden layer to this three-layer perceptron structure model, so the mathematical expressions are generic. The structure of the three-layer multilayer perceptron is shown in Fig. 5.

Given sample set \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \), where \( (x_i, y_i) \) represents the \( i \)-th sample of the data set, the output of the MLP model is

\[
\begin{align*}
    h_j &= f\left(\sum_{i=1}^{d} w_{ij}^{(1)} x_i\right) \\
    a_k &= f\left(\sum_{j=1}^{q} w_{kj}^{(2)} h_j\right)
\end{align*}
\]

where \( w_{ij}^{(1)} \) denotes the weights from the \( i \)-th node of the input layer to the \( j \)-th node of the hidden layer, and \( w_{kj}^{(2)} \) denotes the weights from the \( j \)-th node of the hidden layer to the \( k \)-th node of the output layer.

3.2 Convolutional neural networks

Convolutional neural networks (CNNs) can effectively model more complex structural information by introducing a neural connection that is not fully connected, as is the case for the MLP. In particular, CNNs typically include convolution, pooling, and fully connected operations. The convolution operation of a CNN gives it the characteristics of local connectivity and parameter sharing. Local connectivity refers to the convolution of a kernel and an input feature map. The convolution kernel only operates on a part of the feature map, instead of the full map as would be done with an MLP. With weight sharing, the parameters of the convolution kernels remain the same. With local connectivity and weight sharing, a CNN can effectively reduce the number of model parameters compared to an MLP. An illustration of CNNs is shown in Fig. 5.

3.3 Recurrent neural networks

Recurrent neural networks (RNNs) are another special artificial neural network that contain self-connected recursive connections between neurons, as shown in Fig. 6. Compared to MLPs and CNNs, RNNs can effectively model data in
time by introducing recursive connections. In multimodal learning research, dynamic information with a temporal component is very common, and includes auditory, visual, video, and natural language information.

4 Unified learning architecture

To model multimodal learning, it is essential to define the initial problem. Assume that a multimodal learning task contains $N$ modal raw data, denoted as $m_1, m_2, \ldots, m_N$. Then the learning...
task must consider $N$ modalities comprehensively to make decisions. The multimodal learning problem can then be defined in three parts. First, $N$ sets of characterization functions $R_1, R_2, \ldots, R_N$ must be found such that each $R_i$ characterizes the original data of the $i$-th modality as Feature $x_i$. Second, the decision function $F$ must be found, which maps the feature representations $x_1, x_2, \ldots, x_N$ of all modalities into the hidden layer space $H$, denoted as the hidden layer fusion representation $h$. Then, decision method $C$ performs a final decision task according to the hidden layer fusion representation $h$. The above procedure is formulated in Eq. (3).

$$x_i = R_i(m_i), i = 1, 2, \ldots, N;$$
$$h = F(x_1, x_2, \ldots, x_N);$$
$$\theta = \text{argmax}_\theta p(C|h)$$

Based on the above definition of the multimodal learning problem, this paper proposes a multimodal time-series data modeling method based on deep neural networks. The input to this model is the raw data for $N$ modalities and the output is the final decision $C$. This model adopts a general solution to the problem of multimodal learning. With multimodal time-series data modeling, static modal data (such as images) and temporally dynamic data (such as video, natural language data, and haptic data) can be used. Some modal data contain information, and some modes contain information at a single time.

With dynamic and static information in multimodal contexts, the core of the proposed method is a fusion model based on a recurrent neural network, which can effectively model the dynamic data in time, as shown in Fig. 1. In the figure, the upper corners represent time dimension information $t$, and the lower corners represent modal information, such as Modality$^t_1$, which represents the second modality of data at time node $t$. The input to the recurrent neural network unit is the output of the fusion network (Fusion Net), and the input of the converged network is the output of each modal neural network NeuralNet$^t_i$. Therefore, the feature extraction function of each mode is the neural network designed NeuralNet$^t_i$ for the model.

In the following sections, we focus on two specific multimodal learning problems in order to validate the proposed learning model. One is image captioning, which involves visual and natural language modalities, and the other is a visual-haptic fusion task, which involves haptic and visual modalities.

## 5 Image captioning

In image captioning, an image is input to a computer, which is required to generate a human-level description of the given image. In multimodal learning settings, image captioning tasks require the modeling of visual and natural language modalities. Following the procedure of the proposed unified multimodal learning architecture, image captioning can be divided into three parts. The first represents the visual modality, or the image to be described. The second represents the natural language modality, or the sentence (the words) to generate. The third connects these two modalities and builds the final model.

The proposed image captioning model is presented in Fig. 7. It extracts an object feature from an input image with a CNN, sorts objects according to their significance, maps them to the hidden layer space, and sends them to the coding stage of an RNN. The target sequence is word information that is mapped to the same hidden layer space. The mapping from the source sequence to the target sequence uses an RNN with LSTM recursive units. Word generation is based on the generated word and the latest hidden layer information from the encoding stage; then a word is selected from the dictionary according to its softmax probability.
Theoretically, any object detection technique can be leveraged to extract object features, as long as it can achieve a good object coverage rate. In practice, we leverage a state-of-the-art object detection algorithm, R-FCN [19]. The significance of each object is determined by the detection score and the ratio of the area the object takes in the image. The recursive unit LSTM follows the implementation in Ref. [20]. To validate the proposed model, we test our model on the Flickr8k image captioning dataset [21]. The results are shown in Table 1, where MMT-rnd, MMT-desc, and MMT-asc refer to objects arranged in random, descending, and ascending order, respectively. From Table 1, we learn that the proposed image captioning model performs better than previous neural network-based methods consistently, validating the viability of the proposed unified multimodal learning architecture for image captioning. Figure 8 shows additional visualization results of the model.

### 6 Visual-haptic fusion

In visual-haptic fusion, a robot must make decisions based on visual and haptic input signals. Generally, decision-making is usually more reliable using multiple modalities rather than one. For example, for object recognition tasks performed in some environments, due to complete or partial occlusion, a robot may not recognize an object through a single visual modality. The robot may visually obtain a position estimate of the object and attempt to grasp it. Then, the grasping posture is further adjusted by tactile feedback, and information on the type of the object may be finally obtained by using

![Figure 7](https://mc03.manuscriptcentral.com/brainsa) - The proposed image captioning model on the basis of the unified multimodal learning architecture.

**Table 1** Image captioning results on Flickr8k Datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLSTM [22]</td>
<td>64.7</td>
<td>45.9</td>
<td>31.8</td>
<td>21.6</td>
</tr>
<tr>
<td>Google NIC [23]</td>
<td>63.0</td>
<td>41.0</td>
<td>27.0</td>
<td>18.0</td>
</tr>
<tr>
<td>DeepVS [24]</td>
<td>57.9</td>
<td>38.3</td>
<td>24.5</td>
<td>16.0</td>
</tr>
<tr>
<td>Explain [25]</td>
<td>57.8</td>
<td>27.5</td>
<td>23.1</td>
<td>-</td>
</tr>
<tr>
<td>m-RNN [26]</td>
<td>56.5</td>
<td>38.6</td>
<td>25.6</td>
<td>17.0</td>
</tr>
<tr>
<td>MMT-rnd</td>
<td>62.7</td>
<td>44.4</td>
<td>30.7</td>
<td>20.9</td>
</tr>
<tr>
<td>MMT-desc</td>
<td>63.8</td>
<td>45.3</td>
<td>31.2</td>
<td>21.1</td>
</tr>
<tr>
<td>MMT-asc</td>
<td>65.7</td>
<td>47.6</td>
<td>33.1</td>
<td>22.6</td>
</tr>
</tbody>
</table>
the sensing and visual modal information along with the tactile modality. To determine whether to grasp an opaque object such as a plastic bottle, the quality of the liquid in the vessel cannot be determined solely by vision, so it is impossible for the robot to grasp the vessel through its existing grasping experience. Here, the robot can be assisted by tactile sensation, and the tactile feedback used to determine the final grasping force. In other cases, sufficient performance cannot be achieved by relying on a single modality. Sensing information from a single modality can be beset by problems such as excessive error in some environments. Visual sensors are sensitive to light sources, and captured visual sensing information may suffer from issues related to deformation, illumination, or scale change. Likewise, tactile sensing is sensitive to the material, and different types of objects may be similar in material, while the materials on the same type of object may vary greatly. Since single-mode sensor data cannot accurately describe object attributes, processing information from haptic and visual modes can be used to provide a final description of object attributes.

Given this, under the unified multimodal learning framework, a visual-haptic fusion model is proposed, as shown in Fig. 9. The core of the model is an RNN, with an LSTM unit. At each time step, the network takes in the fusion representation of the visual and haptic signals, represented by the CNN feature and raw signal, respectively. The fusion network is a simple MLP, yet it proves to be sufficient for the hidden representation of the two modalities. To validate the proposed model, we tested the model on the PHAC-2 visual and haptic dataset [26]. Proposed by Chu et al., PHAC-2 captures visual and haptic signals during the grasping procedure of a Willow Garage PR2 robot hand. The hand is equipped with a SynTouch bio-haptic sensor, and four movements were recorded: (1) squeeze, (2) hold, (3) slow-slide, and (4) fast-slide. Sixty household objects were used during the experiments, as shown in Fig. 10. Each object was described by one or more of 24 objectives, and the
task was to predict the existence of each objective for each object of each grasping sample. The prediction result is shown in Table 2, which shows that the proposed model performed competitively.

**Table 2** Prediction results on PHAC-2.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haptic-CNN [27]</td>
<td>78.2%</td>
</tr>
<tr>
<td>Visual-CNN [27]</td>
<td>77.2%</td>
</tr>
<tr>
<td>Multimodal [27]</td>
<td>85.9%</td>
</tr>
<tr>
<td>Visual (ours)</td>
<td>79.3%</td>
</tr>
<tr>
<td>Haptic (ours)</td>
<td>76.5%</td>
</tr>
<tr>
<td>V-H Fusion (ours)</td>
<td>86.3%</td>
</tr>
</tbody>
</table>

### 7 Conclusion

This study tackles the problem of multimodal learning using brain-inspired models, specifically leveraging modern deep neural networks to mine the characteristics of each modality. A unified multimodal learning framework is proposed, and the framework is validated through two practical tasks, image captioning and visual-haptic fusion. Experimental results show the feasibility and validity of the model. Future work includes testing the model in more complex experiment settings.

### Conflict of interests

All contributing authors have no conflict of interests.

### Financial support

This paper is jointly supported by National Natural Science Foundation of China (Grant Nos. 61621136008, 61327809, 61210013, 91420302, and 91520201).

### References


Chang Liu received the B.S. degree from the Department of Computer Science and Technology at Harbin Institute of Technology in July 2013. He received his Doctor of Philosophy in Computer Science and Technology at Tsinghua University, Beijing, July 2018. His research focuses on understanding the contents and meanings of multimodal media inputs, like computer vision inputs in image or video form.

E-mail: liuchang8am@gmail.com
**Fuchun Sun** received the B.S. and M.S. degrees from Naval Aeronautical Engineering Academy, Yantai, China, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, where he is currently a full professor. His current research interests include intelligent control, networked control system and management, neural networks, fuzzy systems, nonlinear systems, and robotics. He has published two books and over 100 papers which have appeared in various journals and conference proceedings. E-mail: fcsun@mail.tsinghua.edu.cn

**Bo Zhang** received the degree from the Department of Automatic Control, Tsinghua University, in 1958. He is currently a professor with the Department of Computer Science and Technology, Tsinghua University. He has published over 150 papers and four monographs. His current research interests include artificial intelligence, pattern recognition, neural networks, and intelligent control. He is a fellow of the Chinese Academy of Sciences.