Associative Multichannel Autoencoder for Multimodal Word Representation

Shaonan Wang[1,2], **Jiajun Zhang**[1,2], **Chengqing Zong**[1,2,3]
[1] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{shaonan.wang,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract

In this paper we address the problem of learning multimodal word representations by integrating textual, visual and auditory inputs. Inspired by the re-constructive and associative nature of human memory, we propose a novel *associative multichannel autoencoder* (AMA). Our model first learns the associations between textual and perceptual modalities, so as to predict the missing perceptual information of concepts. Then the textual and predicted perceptual representations are fused through reconstructing their original and associated embeddings. Using a gating mechanism our model assigns different weights to each modality according to the different concepts. Results on six benchmark concept similarity tests show that the proposed method significantly outperforms strong unimodal baselines and state-of-the-art multimodal models.

1 Introduction

Representing the meaning of a word is a prerequisite to solve many linguistic and non-linguistic problems, such as retrieving words with the same meaning, finding the most relevant images or sounds of a word and so on. In recent years we have seen a surge of interest in building computational models that represent word meanings from patterns of word co-occurrence in corpora (Turney and Pantel, 2010; Mikolov et al., 2013; Pennington et al., 2014; Clark, 2015; Wang et al., 2018b). However, word meaning is also tied to the physical world. Many behavioral studies suggest that human semantic representation is grounded in the external environment and sensorimotor experience (Landau et al., 1998; Barsalou, 2008). This has led to the development of multimodal representation models that utilize both textual and perceptual information (e.g., images, sounds).

As evidenced by a range of evaluations (Andrews et al., 2009; Bruni et al., 2014; Silberer et al., 2016), multimodal models can learn better semantic word representations (a.k.a. embeddings) than text-based models. However, most existing models still have a number of drawbacks. First, they ignore the associations between modalities, and thus lack the ability of information transferring between modalities. Consequently they cannot handle words without perceptual information. Second, they integrate textual and perceptual representations with simple concatenation, which is insufficient to effectively fuse information from various modalities. Third, they typically treat the representations from different modalities equally. This is inconsistent with many psychological findings that information from different modalities contributes differently to the meaning of words (Paivio, 1990; Anderson et al., 2017).

In this work, we introduce the associative multichannel autoencoder (AMA), a novel multimodal word representation model that addresses all the above issues. Our model is built upon the stacked autoencoder (Bengio et al., 2007) to learn semantic representations by integrating textual and perceptual inputs. Inspired by the re-constructive and associative nature of human memory, we propose two associative memory modules as extensions. One is to learn associations between modalities (e.g., associations between textual and visual features), so as to reconstruct corresponding perceptual information of concepts. The other is to learn associations between related concepts, by reconstructing embeddings of both target words and their associated words. Furthermore, we propose a gating mechanism to learn the importance weights of different modalities to each word.

To summarize, our main contributions in this work are two-fold:

- We present a novel associative multichannel autoencoder for multimodal word representation, which is capable of utilizing associations between different modalities and related

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 115–124
Brussels, Belgium, October 31 - November 4, 2018.

concepts, and assigning different importance weights to each modality according to different words. Results on six standard benchmarks demonstrate that our methods outperform strong unimodal baselines and state-of-the-art multimodal models.

- Our model successfully integrates cognitive insights of the re-constructive and associative nature of semantic memory in humans, suggesting that rich information contained in human cognitive processing can be used to enhance NLP models. Furthermore, our results shed light on the fundamental questions of how to learn semantic representations, such as the plausibility of reconstructing perceptual information, associating related concepts and grounding word symbols to external environment.

2 Background and Related Work

2.1 Cognitive Grounding

A large body of research evidences that human semantic memory is inherently re-constructive and associative (Collins and Loftus, 1975; Anderson and Bower, 2014). That is, memories are not exact static copies of reality, but are rather reconstructed from their stimuli and associated concepts each time they are retrieved. For example, when we see a *dog*, not only the concept itself, but also the corresponding perceptual information and associated words will be jointly activated and reconstructed. Moreover, various theories state that the different sources of information contribute differently to the semantic representation of a concept (Wang et al., 2010; Ralph et al., 2017). For instance, Dual Coding Theory (Hiscock, 1974) posits that concrete words are represented in the brain in terms of a perceptual and linguistic code, whereas abstract words are encoded only in the linguistic modality.

In these respects, our method employs a retrieval and representation process analogous to that of humans, in which the retrieval of perceptual information and associated words is triggered and mediated by a linguistic input. The learned cross-modality mapping and reconstruction of associated words are inspired by the human mental model of associations between different modalities and related concepts. Moreover, word meaning is tied to both linguistic and physical environment, and relies differently on each modality inputs (Wang et al., 2018a). These are also captured by our multimodal representation model.

2.2 Multimodal Models

The existing multimodal representation models can be generally classified into two groups: 1) *Jointly training models* build multimodal representations with raw inputs of textual and perceptual resources. 2) *Separate training models* independently learn textual and perceptual representations and integrate them afterwards.

2.2.1 Jointly training models

A class of models extends Latent Dirichlet Allocation (Blei et al., 2003) to jointly learn topic distributions from words and perceptual units (Andrews et al., 2009; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013). Recently introduced work is an extension of the Skip-gram model (Mikolov et al., 2013). For instance, Hill and Korhonen (2014) propose a corpus fusion method that inserts the perceptual features of concepts in the training corpus, which is then used to train the Skip-gram model. Lazaridou et al. (2015) propose MMSkip model, which injects visual information in the process of learning textual representations by adding a max-margin objective to minimize the distance between textual and visual vectors. Kiela and Clark (2015) adopt the MMSkip to learn multimodal vectors with auditory perceptual inputs.

These methods can implicitly propagate perceptual information to word representations and at the same time learn multimodal representations. However, they utilize raw text corpus in which words having perceptual information account for a small portion. This weakens the effect of introducing perceptual information and consequently leads to the slight improvement of textual vectors.

2.2.2 Separate training models

The simplest approach is concatenation which fuses textual and visual vectors by concatenating them. It has been proven to be effective in learning multimodal representations (Bruni et al., 2014; Hill et al., 2014; Collell et al., 2017). Variations of this method employ transformation and dimension reduction on the concatenation result, including application of singular value decomposition (SVD) (Bruni et al., 2014) or canonical correlation analysis (CCA) (Hill et al., 2014). There is also work using deep learning methods to project different modality inputs into a common space, including restricted Boltzman machines (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012), autoencoders (Silberer and Lapata, 2014; Silberer et al., 2016), and recursive neural networks (Socher et al., 2013). However, the above methods can only generate multimodal vectors of those words that have perceptual information, thus reducing multimodal vocabulary drastically.

An empirically superior model addresses this problem by predicting missing perceptual information firstly. This includes Hill et al. (2014) who utilize the ridge regression method to learn a mapping matrix from textual modality to visual modality, and Collell et al. (2017) who employ a feed-forward neural network to learn the mapping relation between textual vectors and visual vectors. Applying the mapping function on textual representations, they obtain the predicted visual vectors for all words in textual vocabulary. Then they calculate multimodal representations by concatenating textual and predicted visual vectors. However, the above methods learn separate mapping functions and fusion models, which are somewhat inelegant. In this paper we employ a neural-network mapping function to integrate these two processes into a unified multimodal models.

According to this classification, our method falls into the second group. However, existing models ignore either the associative relations among modalities, associative relations among relative words, or the different contributions of each modality. This paper aims to integrate more perceptual information and the human-like associative memory into a unified multimodal model to learn better word representations.

3 Associative Multichannel Autoencoder

We first provide a brief description of the basic multichannel autoencoder for learning multimodal word representations (Figure 1). Then we extend the model with two associative memory modules and a gating mechanism (Figure 2) in the next sections.

3.1 Basic Mutichannel Autoencoder

An autoencoder is an unsupervised neural network which is trained to reconstruct a given input from its latent representation (Bengio, 2009). In this work, we propose a variant of autoencoder called multichannel autoencoder, which maps multimodal inputs into a common space.

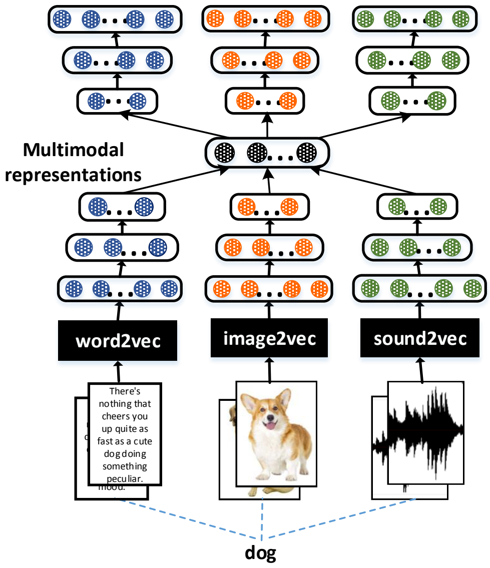
Multimodal representations
word2vec
image2vec
sound2vec
There's nothing that cheers you up quite as fast as a cute dog doing something peculiar.
mood.
dog

Figure 1: Architecture of the multichannel autoencoder with inputs of textual, visual and auditory sources.

Our model extends the unimodal and bimodal autoencoder (Ngiam et al., 2011; Silberer and Lapata, 2014) to induce semantic representations integrating textual, visual and auditory information. As shown in Figure 1, our model first transforms input textual vector x_t, visual vector x_v and auditory vector x_a to hidden representations:

$$\begin{aligned} h_t &= g(W_t x_t + b_t) \\ h_v &= g(W_v x_v + b_v) \\ h_a &= g(W_a x_a + b_a). \end{aligned} \tag{1}$$

Then the hidden representations are concatenated together and mapped to a common space:

$$h_m = g(W_m[h_t; h_v; h_a] + b_m). \tag{2}$$

The model is trained to reconstruct the hidden representations of the three modalities from the multimodal representation h_m:

$$[\hat{h}_t; \hat{h}_v; \hat{h}_a] = g(W'_m h_m + b_{\hat{m}}), \tag{3}$$

and finally to reconstruct the original embeddings of textual, visual and auditory inputs:

$$\begin{aligned} \hat{x}_t &= g(W'_t \hat{h}_t + b_{\hat{t}}) \\ \hat{x}_v &= g(W'_v \hat{h}_v + b_{\hat{v}}) \\ \hat{x}_a &= g(W'_a \hat{h}_a + b_{\hat{a}}), \end{aligned} \tag{4}$$

where $\hat{x}_t$, $\hat{x}_v$, $\hat{x}_a$ are the reconstruction of input vectors x_t, x_v, x_a, and $\hat{h}_t$, $\hat{h}_v$, $\hat{h}_a$

are the reconstruction of hidden representations h_t, h_v, h_a. The learning parameters $\{W_t, W_v, W_a, W'_t, W'_v, W'_a, W_m, W'_m\}$ are weight matrices, $\{b_t, b_v, b_a, b_{\hat{t}}, b_{\hat{v}}, b_{\hat{a}}, b_m, b_{\hat{m}}\}$ are bias vectors. Here $[\cdot\,;\,\cdot]$ denotes the vector concatenation, and g denotes the non-linear function which we use $tanh(\cdot)$.

Training a single-layer autoencoder corresponds to optimizing the learning parameters to minimize the overall loss between inputs and their reconstructions. Following (Vincent et al., 2010), we use squared loss:

$$\min_{\theta_1} \sum_{i=1}^{n} (||x_t^i - \hat{x}_t^i||^2 + ||x_v^i - \hat{x}_v^i||^2 + ||x_a^i - \hat{x}_a^i||^2), \tag{5}$$

where i denotes the i^{th} word, and the model parameters are $\theta_1 = \{W_t, W_v, W_a, W_m, W'_t, W'_v, W'_a, W'_m, b_t, b_v, b_a, b_m, b_{\hat{t}}, b_{\hat{v}}, b_{\hat{a}}, b_{\hat{m}}\}$.

Autoencoders can be stacked to create deep networks. To enhance the quality of semantic representations, we employ a stacked multichannel autoencoder, which is composed of multiple hidden layers that are stacked together.

3.2 Integrating Modality Associations

In reality, the words that have corresponding images or sounds are only a small subset of the textual vocabulary. To obtain the perceptual vectors for each word, we need associations between modalities (i.e., text-to-vision and text-to-audition mapping functions), that transform the textual vectors into visual and auditory ones. Previous methods learn separate mapping functions and fusion models, which are somewhat inelegant. Here we employ a neural-network mapping function to incorporate this modality association module into multimodal models.

Take text-to-vision mapping as an example. Suppose that $T \in \mathbb{R}^{m_t \times n_t}$ is the textual representation containing m_t words, $V \in \mathbb{R}^{m_v \times n_v}$ is the visual representation containing m_v ($\ll m_t$) words, where n_t and n_v are dimensions of the textual and visual representations respectively. The textual and visual representations of the i^{th} concept are denoted as T_i and V_i respectively. Our goal is to learn a mapping function $f : g(W_p T + b_p)$ from textual to visual space such that the prediction $f(T_i)$ is similar to the actual visual vector V_i. The set of visual representations along with their corresponding textual representations are used to learn the mapping function. To train the model, we employ a square loss:

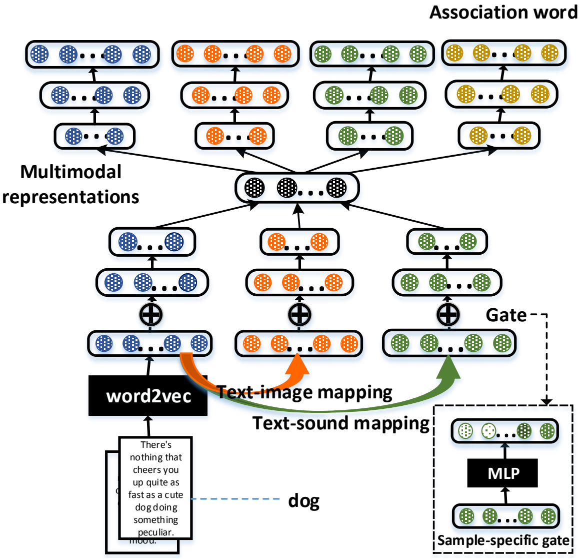
Association word
Multimodal representations
Gate
word2vec
Text-image mapping
Text-sound mapping
MLP
Sample-specific gate
There's nothing that cheers you up quite as fast as a cute dog doing something peculiar.
dog

Figure 2: Architecture of the proposed associative multichannel autoencoder.

$$\min_{\theta_2} \sum_{i=1}^{m_v} ||f(T_i) - V_i||^2, \tag{6}$$

where the training parameters are $\theta_2 = \{W_p, b_p\}$. We adopt the same method to learn the text-to-audition mapping function.

3.3 Integrating Word Associations

Word associations are a proxy for an aspect of human semantic memory that is not sufficiently captured by the usual training objectives of multimodal models. Therefore we assume that incorporating the objective of word associations helps to learn better semantic representations. To achieve this, we propose to reconstruct the vector of associated word from the corresponding multimodal semantic representation. Specifically, in the decoding process we change the equation (3) to:

$$[\hat{h}_t, \hat{h}_v, \hat{h}_a, \hat{h}_{asc}] = g(W'_m h_m + b_{\hat{m}}), \tag{7}$$

and equation (4) to:

$$\begin{aligned} \hat{x}_t &= g(W'_t \hat{h}_t + b_{\hat{t}}) \\ \hat{x}_v &= g(W'_v \hat{h}_v + b_{\hat{v}}) \\ \hat{x}_a &= g(W'_a \hat{h}_a + b_{\hat{a}}) \\ \hat{x}_{asc} &= g(W_{asc} \hat{h}_{asc} + b_{asc}). \end{aligned} \tag{8}$$

To train the model, we add an additional objective function, which is the mean square error between the embeddings of the associated word y and their re-constructive embeddings $\hat{x}_{asc}$:

$$\min_{\theta_3} \sum_{i=1}^{n} ||y^i - \hat{x}^i_{asc}||^2, \tag{9}$$

where y^i and x^i are the embeddings of a pair of associated words. Here, y is the concatenation of three unimodal vectors $[y_t; y_v; y_a]$. The parameters of word association module are $\theta_3 = \{W_t, W_v, W_a, W_m, \hat{W}_m, W_{asc}, b_t, b_v, b_a, b_m, b_{\hat{m}}, b_{asc}\}$. This additional criterion drives the learning towards a semantic representation capable of reconstructing its associated representation.

3.4 Integrating a Gating Mechanism

Considering that the meaning of each word has different dependencies on textual and perceptual information, we propose the sample-specific gate to assign different weights to each modality according to different words. The weight parameters are calculated by the following feed-forward neural networks:

$$\begin{aligned} g_t &= g(W_{gt} x_t + b_{gt}) \\ g_v &= g(W_{gv} x_v + b_{gv}) \\ g_a &= g(W_{ga} x_a + b_{ga}), \end{aligned} \tag{10}$$

where g_t, g_v and g_a are value or vector gate of textual, visual and auditory representations respectively. For the value gate, W_{gt}, W_{gv} and W_{ga} are vectors, and b_{gt}, b_{gv} and b_{ga} are value parameters. For the vector gate, the parameters W_{gt}, W_{gv} and W_{ga} are matrices, b_{gt}, b_{gv} and b_{ga} are vectors. The value gate controls the importance weights of different input representations as a whole, whereas the vector gate can adjust the importance weights of each dimension of input representations.

Finally, we compute element-wise multiplication of the textual, visual and auditory representations with their corresponding gates:

$$\begin{aligned} x_{gt} &= x_t \odot g_t \\ x_{gv} &= x_v \odot g_v \\ x_{ga} &= x_a \odot g_a. \end{aligned} \tag{11}$$

The x_{gt}, x_{gv} and x_{ga} can be seen as the weighted textual, visual and auditory representations. The parameters of our gating mechanism is trained together with that of the proposed model.

3.5 Model Training

To train the AMA model, we use overall objective function of equation $(5) + (6) + (9)$. In the training phase, model inputs are textual vectors, the corresponding visual vectors, auditory vectors, and association words (Figure 2). In the testing phase, we only need textual inputs to generate multimodal word representations.

4 Experimental Setup

4.1 Datasets

Textual vectors. We use 300-dimensional GloVe vectors[1] which are trained on the Common Crawl corpus consisting of 840B tokens and a vocabulary of 2.2M words[2].

Visual vectors. Our source of visual vectors are collected from ImageNet (Russakovsky et al., 2015) which covers a total of 21,841 WordNet synsets (Fellbaum, 1998) that have 14,197,122 images. For our experiments, we delete words with fewer than 50 images or words not in the Glove vectors, and sample at most 100 images for each word. To generate a visual vector for each word, we use the forward pass of a pre-trained VGG-net model[3] and extract the hidden representation of the last layer as the feature vector. Then we use averaged feature vectors of the multiple images corresponding to the same word. Finally, we get 8,048 visual vectors of 128 dimensions.

Auditory vectors. For auditory data, we gather audio files from Freesound[4], in which we select words with more than 10 audio files and sample at most 50 sounds for one word. To extract auditory features, we use the VGG-net model which is pre-trained on Audioset[5]. The final auditory vectors are averaged feature vectors of multiple audios of the same word, which contains 9,988 words of 128 dimensions[6].

Word associations. We use the word association data collected by (De Deyne et al., 2016), in which each word pair is generated by at least

[1] http://nlp.stanford.edu/projects/glove

[2] We have tried skip-gram vectors and get the same conclusions.

[3] http://www.vlfeat.org/matconvnet/

[4] http://www.freesound.org/

[5] https://research.google.com/audioset

[6] We build auditory vectors with the released code at: https://github.com/tensorflow/models/tree/master/research/audioset

one subject[7]. This dataset includes mostly words with similar meaning (e.g., occasionally & sometimes, adored & loved, supervisor & boss) and related words (e.g., eruption & volcano, cortex & brain, umbrella & rain). We calculate the association score for each word pair (cue word + target word) as: *the number of person who generated the word pair divided by the total number of people who were presented with the cue word.* For training, we select pairs of associated words above a threshold of 0.15 and delete those that are not in the Glove vocabulary, which results in 7,674 word association data sets[8]. For the development set, we randomly sample 5,000 word association collections together with their association scores.

4.2 Model Settings

Our models are implemented with PyTorch (Paszke et al., 2017), optimized with Adam (Kingma and Ba, 2014). We set the initial learning rate to 0.05, and batch size to 64. We tune the number of layers over 1, 2, 3, the size of multimodal vectors over 100, 200, 300, and the size of each layer in textual channel over 300, 250, 200, 150, 100 and in visual/auditory channel over 128, 120, 90, 60. We train the model for 500 epochs and select the best parameters on the development set. All models are trained for 3 times and the average results are reported in Table 1.

To test the effect of each module, we separately train the following models: multichannel autoencoder with modality association (AMA-M), with modality and word associations (AMA-MW), with modality and word associations plus value/vector gate (AMA-MW-Gval/vec).

For AMA-M model, we initialize the text-to-vision and text-to-audition mapping functions with pre-trained mapping matrices, which are parameters of one-layer feed-forward neural networks. The network uses input of the textual vectors, output of visual or auditory vectors, and is trained with SGD for 100 epochs. We initialize the network biases as zeros and network weights with He-initialisation (He et al., 2015). The best parameters of AMA-M model are 2 hidden layers, with textual channel size of 300, 250 and 150, visual/auditory channel size of 128, 90, 60. For AMA-MW model, we use the best AMA-M model parameters as initialization, and train the model with word association data. The optimal parameter of association channel size is 300, 350, 556 (or 428 for bimodal inputs). For AMA-MW-Gval and AMA-MW-Gvec, we adopt the same training strategy as AMA-MW model. The code for training and evaluation can be found at: `https://github.com/wangshaonan/Associative-multichannel-autoencoder`.

5 Experiments

5.1 Evaluation Tasks

We test the baseline and proposed models on six standard evaluation benchmarks, covering two different tasks: (i) Semantic relatedness: Men-3000 (Bruni et al., 2014) and Wordrel-252 (Agirre et al., 2009); (ii) Semantic similarity: Simlex-999 (Hill et al., 2016), Semsim-7576 (Silberer and Lapata, 2014), Wordsim-203 and Simverb-3500 (Gerz et al., 2016). All test sets contain a list of word pairs along with their subject ratings.

We employ Spearman's correlation method to evaluate the performance of our models. This method calculates the correlation coefficients between model predictions and subject ratings, in which the model prediction is the cosine similarity between semantic representations of two words.

5.2 Baseline Multimodal Models

Most of existing multimodal models only utilize textual and visual modalities. For fair comparison, we re-implement several representative systems with our own textual and visual vectors. The **Concatenation (CONC) model** (Kiela and Bottou, 2014) is simple concatenation of normalized textual and visual vectors. The **Mapping** (Collell et al., 2017) and **Ridge** (Hill et al., 2014) models first learn a mapping matrix from textual to visual modality using feed-forward neural network and ridge regression respectively. After applying the mapping function on the textual vectors, they obtain the predicted visual vectors for all words in textual vocabulary. Then they concatenate the normalized textual and predicted visual vectors to get multimodal word representations. The **SVD** (Bruni et al., 2014) and **CCA** (Hill et al., 2014) models first concatenate normalized textual and visual vectors, and then conduct SVD or CCA transformations on the concatenated vectors.

For multimodal models with textual, visual and

[7] The dataset can be found at: `https://simondedeyne.me/data`.

[8] We have done experiments with Synonyms (which are extracted from WordNet and PPDB corpora), and the results are not as good as using word associations.

Table 1: Spearman's correlations between model predictions and human ratings on six evaluation datasets. Here T, V, A denote textual, visual and auditory. TV denotes bimodal inputs of textual and visual. TVA denotes trimodal inputs of textual, visual and auditory. The bold scores are the best results per column in bimodal models and trimodal models respectively. For each test, ALL corresponds to the whole testing set, V/A to those word pairs for which we have textual&visual vectors in bimodal models or textual&visual&auditory in trimodal models, and ZS (zero-shot) denotes word pairs for which we have only textual vectors. The #inst. denotes the number of word pairs.

	MEN			SIMLEX			SEMSIM			SIMVERB			WORDSIMM			WORDREL		
	ALL	V/A	ZS	ALL	V/A	ZS	ALL	V/A	ZS	ALLL	V/A	ZS	ALL	V/A	ZS	ALL	V/A	ZS
Kiela & Bottou 2014	-	0.72	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-
Silberer & Lapata 2014	-	-	-	-	-	-	0.70	-	-	-	-	-	-	-	-	-	-	-
Lazaridou et al., 2015	0.75	0.76	-	0.40	0.53	-	0.72	0.72	-	-	-	-	-	-	-	-	-	-
Collell et al., 2017	0.811	0.819	0.802	0.410	0.388	0.422	0785	0.791	0.764	0.286	0.371	0.285	0.781	0.698	0.766	0.629	0.797	0.601
Glove-textual (V)	0.802	0.799	0.788	0.408	0.371	0.429	0.744	0.751	0.716	0.283	0.320	0.282	0.798	0.688	0.779	0.682	0.759	0.661
Glove-textual (A)	0.802	0.801	0.830	0.408	0.399	0.456	0.744	0.715	0.762	0.283	0.129	0.397	0.798	0.805	0.785	0.682	0.708	0.652
CNN-visual	-	0.566	-	-	0.406	-	-	0.502	-	-	0.235	-	-	0.526	-	-	0.422	-
Predicted-visual	0.698	0.757	0.656	0.372	0.458	0.347	0.702	0.700	0.709	0.212	0.194	0.211	0.596	0.621	0.557	0.412	0.604	0.384
CNN-auditory	-	0.266	-	-	0.053	-	-	0.159	-	-	0	-	-	0.231	-	-	0.088	-
Predicted-auditory	0.558	0.555	0.597	0.270	0.251	0.296	0.547	0.531	0.559	0.157	0.074	0.227	0.515	0.496	0.544	0.388	0.400	0.372
CONC (TV)	-	0.786	-	-	0.442	-	-	0.709	-	-	0.437	-	-	0.665	-	-	0.666	-
Mapping (TV)	0.806	0.815	0.782	0.408	0.407	0.410	0.769	0.771	0.709	0.282	0.358	0.272	0.781	0.696	0.768	0.650	0.751	0.594
Ridge (TV)	0.806	0.816	0.786	0.418	0.405	0.429	0.764	0.766	0.756	0.287	0.329	0.285	0.786	0.689	0.771	0.660	0.765	0.640
SVD (TV)	0.806	0.816	0.786	0.418	0.405	0.429	0.764	0.766	0.756	0.287	0.330	0.286	0.786	0.689	0.771	0.660	0.764	0.640
CCA (TV)	0.816	0.833	0.798	0.478	**0.507**	0.493	0.656	0.666	0.619	0.333	0.276	0.334	0.757	0.754	0.704	0.626	0.733	0.599
AMA-M (TV)	0.836	0.822	0.834	0.445	0.460	0.471	**0.781**	**0.784**	0.769	0.324	0.403	0.323	0.807	0.754	0.769	0.681	0.814	0.648
AMA-MW (TV)	0.838	0.824	0.822	0.471	0.446	**0.509**	0.757	0.738	0.723	0.343	0.421	0.340	**0.814**	**0.780**	0.737	0.707	0.744	0.659
AMA-MW-Gval (TV)	**0.845**	**0.835**	**0.841**	0.476	0.472	0.506	0.776	0.778	0.767	**0.352**	0.396	**0.352**	0.808	0.758	0.763	**0.726**	0.796	**0.705**
AMA-MW-Gvec (TV)	0.840	0.831	0.835	**0.485**	0.486	0.505	0.766	0.769	**0.778**	0.343	**0.523**	0.342	0.811	0.769	**0.778**	0.694	**0.846**	0.661
CONC (TVA)	-	0.778	-	-	0.451	-	-	0.661	-	-	0.503	-	-	0.687	-	-	0.593	-
Ridge (TVA)	0.805	0.812	0.791	0.417	0.428	0.420	0.764	0.725	0.781	0.286	0.557	0.285	0.785	0.733	0.762	0.659	0.716	0.646
AMA-M (TVA)	0.831	0.814	0.832	0.452	0.488	0.472	**0.778**	**0.741**	**0.793**	0.333	0.531	0.332	0.805	0.751	0.784	0.685	0.703	0.670
AMA-MW (TVA)	0.838	0.826	0.838	0.481	**0.508**	0.508	0.762	0.726	0.777	**0.358**	**0.605**	**0.357**	**0.814**	**0.821**	0.787	**0.734**	**0.819**	**0.711**
AMA-MW-Gval (TVA)	**0.849**	**0.832**	**0.851**	**0.488**	0.500	**0.509**	0.772	0.729	0.790	0.347	0.598	0.347	0.810	0.806	0.782	0.730	0.761	0.710
AMA-MW-Gvec (TVA)	0.843	0.815	0.843	0.477	0.505	0.497	0.767	0.733	0.781	0.346	0.564	0.346	0.812	0.779	**0.788**	0.723	0.729	0.705
#inst.-visual	3000	1065	1935	999	261	738	7546	5757	1789	3500	41	3459	201	45	158	245	28	224
#inst.-auditory	3000	2732	268	999	741	258	7546	2816	4730	3500	1362	2138	201	129	72	245	153	92
#inst.-visual-auditory	3000	964	2036	999	238	761	7546	2322	5224	3500	22	3478	201	30	171	245	25	220

auditory inputs, we implement **CONC** and **Ridge** as baseline models. The trimodal CONC model simply concatenates normalized textual, visual and auditory vectors. The trimodal Ridge model first learns text-to-vision and text-to-audition mapping matrices with ridge regression method. Then it applies the mapping functions on the textual vectors to get the predicted visual and auditory vectors. Finally, the normalized textual, predicted-visual and predicted-auditory vectors are concatenated to get the multimodal representations.

All above baseline models are implemented with Sklearn[9]. Same as the proposed AMA model, the hyper-parameters of baseline models are tuned on the development set using Spearman's correlation method. In Ridge model, the optimal regularization parameter is 0.6. The Mapping model is trained with SGD for maximum 100 epochs with early stopping, and the optimal learning rate is 0.001. The output dimension of SVD and CCA models are 300.

5.3 Results and Discussion

As shown in Table 1, we divide all models into six groups: (1) existing multimodal models (with textual and visual inputs) in which results are reprinted from Collell et al. (2017). (2) Unimodal models with textual, (predicted) visual or (pre-

[9] http://scikit-learn.org/

dicted) auditory inputs. (3) Our re-implementation of baseline bimodal models with textual and visual inputs (TV). (4) Our AMA models with textual and visual inputs. (5) Our implementation of trimodal baseline models with textual, visual and auditory inputs (TVA). (6) Our AMA model with textual, visual and auditory inputs.

Overall performance Our AMA models (in group 4 and 6) clearly outperform their baseline unimodal and multimodal models (in group 2, 3 and 5). We use Wilcoxon signed-rank test to check if significant difference exists between two models. Results show that our multimodal models perform significantly better ($p < 0.05$) than all baseline models.

As shown clearly, our bimodal and trimodal AMA models achieve better performance than baselines in both V/A (visual or auditory, the testing data that have associated visual or auditory vectors) and ZS (zero-shot, the testing data that do not have associated visual or auditory vectors) region. In other words, our models outperform baseline models on words with or without perceptual information. The good results in ZS region also indicate that our models have good generalization capacity.

Unimodal baselines As shown in group 2, the Glove vectors are much better than CNN-visual and CNN-auditory vectors, in which CNN-auditory has the worst performance on capturing concept similarities. Comparing with visual and auditory vectors, the predicted visual and auditory vectors achieve much better performance. This indicates that the predicted vectors contain richer information than purely perceptual representations and are more useful for building semantic representations.

Multimodal baselines For bimodal models (group 3), the CONC model that combines Glove and visual vectors performs worse than Glove on four out of six datasets, suggesting that simple concatenation might be suboptimal. The Mapping and Ridge models, which combine Glove and predicted visual vectors, improve over Glove on five out of six datasets in ALL regions. This reinforces the conclusion that the predicted visual vectors are more useful in building multimodal models. The SVD model gets similar results as Ridge model. The CCA model maps different modality inputs into a common space, achieving better results on some datasets and worse results on the others. The improvement on three benchmark tests shows the potential of mapping multimodal inputs into a common space.

The above results can also be observed in the trimodal CONC and Ridge models (group 5). Overall, the trimodal models, which utilize additional auditory inputs, get slightly worse performance than bimodal models. This is partly caused by the fusion method of concatenation. Note that our proposed AMA models are more effective with trimodal inputs as shown in group 6.

Our multimodal models With either bimodal or trimodal inputs, the proposed AMA-M model outperforms all baseline models by a large margin. Specifically our AMA-M model achieves an relative improvement of 4.1% on average (4.5% with trimodal inputs) over the state-of-the-art Ridge model. This illustrates that our AMA models can productively combine textual and perceptual representations. Moreover, our AMA-MW model, which employs word associations, achieves an average improvement of 1.5% (2.7% with trimodal inputs) over the AMA-M model. That is to say, the representation ability of multimodal models can be clearly improved by learning associative relations between words. Furthermore, the AMA-MW-Gval model improves the AMA-MW model by 1.3% (0.3% with trimodal inputs) on average, illustrating that the gating mechanism (especially the value gate) helps to learn better semantic representations.

In addition, we explore the effect of word association data size. We find that the decrease of association data has no discernible effect on model performance: when using 100%, 80%, 60%, 40%, 20% of the data, the average results are 0.6479, 0.6409, 0.6361, 0.6430, 0.6458 in bimodal model. The same trend is observed in trimodal models.

6 Conclusions and Future Work

We have proposed a cognitively-inspired multimodal model — associative multichannel autoencoder — which utilizes the associations between modalities and related words to learn multimodal word representations. Performance improvement on six benchmark tests shows that our models can efficiently fuse different modality inputs and build better semantic representations.

Ultimately, the present paper sheds light on the fundamental questions of how to learn word meanings, such as the plausibility of reconstructing perceptual information, associating related concepts and grounding word symbols to external environment. We believe that one of the promising future directions is to learn from how humans learn and store semantic word representations to build a more effective computational model.

Acknowledgement

The research work descried in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002103 and also supported by the Natural Science Foundation of China under Grant No. 61333018. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL*, pages 19–27.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *TACL*, 5:17–30.

John R Anderson and Gordon H Bower. 2014. *Human associative memory*. Psychology press.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Yoshua Bengio. 2009. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1):1–127.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

Stephen Clark. 2015. Vector space models of lexical meaning. *Handbook of Contemporary Semantic Theory, The*, pages 493–522.

Guillem Collell, Teddy Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *AAAI*.

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82:407–428.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. *COLING*, pages 1861–1870.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *EMNLP*, pages 255–265.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Merrill Hiscock. 1974. Imagery and verbal processes. *Psyccritiques*, 19(6):487.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45.

Douwe Kiela and Stephen Clark. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *EMNLP*, pages 2461–2470.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Barbara Landau, Linda Smith, and Susan Jones. 1998. Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1):19–24.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *ACL*, pages 153–163.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*, pages 689–696.

Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford University Press.

Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2017. Pytorch.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. 2017. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal lda model integrating textual, cognitive and visual modalities. In *EMNLP*, pages 1146–1157.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2016. Visually grounded meaning representations. *IEEE transactions on pattern analysis and machine intelligence*.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *EMNLP*, pages 1423–1433.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL*, pages 721–732.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408.

Jing Wang, Julie A. Conder, David N. Blitzer, and Svetlana. V. Shinkareva. 2010. Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Human brain mapping*, 31(10):1459–1468.

Shaonan Wang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2018a. Investigating inner properties of multimodal representation and semantic compositionality with brain-based componential semantics. In *AAAI*, pages 5964–5972.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018b. Learning multimodal word representation via dynamic fusion methods. In *AAAI*, pages 5973–5980.