Learning Multimodal Word Representation via Dynamic Fusion Methods

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, Shanghai, China
{shaonan.wang,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract

Multimodal models have been proven to outperform text-based models on learning semantic word representations. Almost all previous multimodal models typically treat the representations from different modalities equally. However, it is obvious that information from different modalities contributes differently to the meaning of words. This motivates us to build a multimodal model that can dynamically fuse the semantic representations from different modalities according to different types of words. To that end, we propose three novel dynamic fusion methods to assign importance weights to each modality, in which weights are learned under the weak supervision of word association pairs. The extensive experiments have demonstrated that the proposed methods outperform strong unimodal baselines and state-of-the-art multimodal models.

Introduction

Representing the meaning of a word is a prerequisite to solve many natural language problems, such as calculating semantic relations between different words, finding the most relevant images of a word and so on. In recent years, computational semantic models that represent word meanings from patterns of word co-occurrence in corpora have received a lot of research interests (Turney and Pantel 2010; Mikolov et al. 2013; Clark 2015; Wang, Zhang, and Zong 2017a). However, compared to human semantic representation, these purely text-based models are severely impoverished for lacking perceptual information attached to the physical world. This observation has led to the development of multimodal word representation models that utilize both linguistic (e.g., text) and perceptual information (e.g., images, audios). Such models can learn better semantic word representations than text-based models, as evidenced by a range of evaluations (Andrews, Vigliocco, and Vinson 2009; Bruni, Tran, and Baroni 2014).

Learning good multimodal word representations relies not only on the quality of the word representations from linguistic and perceptual inputs, but also the ability to productively combine these representations. However, the existing multimodal models generally treat the word representations from different modalities equally. This is inconsistent with

the fact that meaning of concrete words like `horse` and `computer` are mostly learned from perceptual experiences of seeing, touching and listening. In contrast, more abstract words, such as `hope` and `lovely`, are encoded mostly in linguistic modality rather than perceptual modality, which has been found in cognitive psychology (Wang et al. 2010; Binder et al. 2016) and computational experiments (Hill, Reichart, and Korhonen 2014; Hill and Korhonen 2014).

All these factors motivate us to build a multimodal model that can dynamically fuse information from linguistic and perceptual modalities according to different types of words. We can optimize the importance weights of different modalities for a word if the word has the gold representation. As no gold word representation exists in reality, we resort to word pairs which share the same meaning, so that they can guide each other. In this paper we utilize word association pairs[1], which are generated by subjects firstly reading a cue word and then writing down the first word(s) that come to mind. Some examples are `wealthy` and `rich`, `jigsaw` and `puzzle`, `larger` and `bigger`. We assume that these association word pairs can lead us to learn the importance weights for different modalities. For instance, representations of abstract words `larger` and `bigger` are composed by linguistic and perceptual vectors, and the linguistic vectors are more important in representing abstract word meaning (i.e., the two words share more similarity in linguistic modality). To achieve the goal of making these two association words obtain similar representations, the model will assign more weights to their linguistic vectors.

In light of these considerations, we propose three novel dynamic fusion methods to improve multimodal word representations. The three methods utilize a modality-specific gate, category-specific gate and sample-specific gate respectively, to learn different weights of linguistic and perceptual representations for each input modality, each supersense category and each word sample respectively. Furthermore, we perform extensive analysis to shed light on the principle of the proposed dynamic fusion methods. To summarize, our main contributions are two-fold:

- We present a novel dynamic fusion method for multi-

---

[1]We have also tried other resources, such as synonyms from WordNet. However, these datasets are noisy and perform slightly worse, thus we only report results of word associations.

modal representations, which utilizes a small set of word association pairs to learn different weights of different modalities for semantic word representations. The core idea is to introduce weak supervision to learn a generic fusion rule. Results on six standard benchmarks demonstrate that our method significantly improves the quality of baseline multimodal representations.

- Quantitative analysis shows that the proposed models can successfully assign different weights to linguistic and perceptual representations, and the learned weights show clear difference between concrete and abstract words. This offers initial support for the idea that humans differently encode concrete words and abstract words, and it also indicates that the proposed model can assist in exploring human semantic representation.

Background and Related Work

Cognitive Grounding

Dual coding theory (Hiscock 1974) posits that concrete words are represented in the brain in terms of a visual and linguistic code, whereas abstract words are encoded only in the linguistic modality. This theory has been initially validated by a number of neuroimaging studies (Wang et al. 2010; Andrew et al. 2017).

In summary of previous studies, Wang et al. (2010) conduct a meta-analysis for differences in human neural representation of abstract and concrete words. Their results show that abstract words elicit greater activity in linguistic-related brain area while concrete words elicit greater activity in perceptual-related brain area. With the help of computational models, Andrew et al. (2017) decode functional Magnetic Resonance Imaging (fMRI) activity patterns associated with concrete and abstract words. They observe that both linguistic and visual representations can significantly decode most concrete nouns, while the abstract nouns can only be decoded by linguistic representations.

To sum up, these studies hold that both linguistic and perceptual information affect human representations of concrete words, while only linguistic modality plays a large role in representing meaning of abstract words. In this respect, our method employs a representation process analogous to that of humans, in which linguistic and perceptual modalities contribute differently to concrete and abstract words.

Multimodal Models

There is by now a large literature of multimodal representation models, and the existing models can be generally classified into two groups:

1) **Joint training models** that build multimodal representations with raw inputs of both linguistic and perceptual resources.

A class of models extends Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) to jointly learn topic distributions from words and perceptual units (Fellbaum 1998; Andrews, Vigliocco, and Vinson 2009; Silberer and Lapata 2012; Roller and Schulte im Walde 2013). The 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 a word 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 linguistic representations by adding a max-margin objective function to minimize the distance between linguistic vectors and visual vectors.

The joint training methods implicitly propagate perceptual information to word representations and at the same time learn multimodal representations. However, these methods utilize raw text corpus in which words associated with perceptual information account for a small portion. This weakens the effect of introducing perceptual information, and consequently leads to only limited improvement of linguistic vectors.

2) **Separate training models** that independently learn linguistic and perceptual representations and integrate them afterwards.

The simplest approach is concatenation, which fuses linguistic and visual vectors by concatenating them. It has been proven to be effective in learning multimodal models (Bruni, Tran, and Baroni 2014; Hill, Reichart, and Korhonen 2014; Collell, Zhang, and Moens 2017). Variations of this method employ transformation and dimension reduction on the concatenation result, including application of singular value decomposition (SVD) (Bruni, Tran, and Baroni 2014) or canonical correlation analysis (CCA) (Hill, Reichart, and Korhonen 2014). In addition, Silberer and Lapata (2014) and Silberer et al. (2017) use stacked autoencoder to learn multimodal representations by embedding linguistic and visual inputs into a common space with the objective function of reconstructing the individual inputs. However, the above methods can only generate multimodal representations of those words that have perceptual information, thus reducing multimodal vocabulary drastically.

An empirically superior model addresses this problem by firstly predicting missing perceptual information. This includes Hill et al. (2014) who utilize the ridge regression method to learn a mapping matrix from linguistic modality to visual modality, and Collell et al. (2017) who employ a feed-forward neural network to learn the mapping relation between linguistic vectors and visual vectors. Applying the mapping function on linguistic representations, they obtain the predicted visual vectors for all words in linguistic vocabulary. Then they calculate multimodal representations by concatenating linguistic and predicted visual vectors. Furthermore, they find that irrelevant visual information is discarded in process of associating language to vision, which makes the predicted visual vectors outperform original visual vectors on various semantic similarity experiments.

According to this classification, our method falls into the second group. However, the fact that representations from different modality contribute differently to word meanings is ignored by existing models. This paper aims to solve this

problem by assigning different importance weights for linguistic and perceptual representations according to different type of words, which can be seen as a weighted combination model.

In multimodal representation models, the effectiveness of weighted combination is first emphasized by Bruni et al. (2014), in which weights are super-parameters and the same for all words. Furthermore, Kiela et al. (2014) propose the Dispersion method to distinguish abstract words from concrete words, based on the observation that diversity of a word's images negatively correlates with its concreteness (in which diversity is the average cosine distance between all the visual representations of a word). Then they give zero weights to the perceptual representations of abstract words in building multimodal word representations. However, this method ignores the concreteness of each word, and can not handle words without images due to relying on visual information.

Proposed Method

The problem of learning multimodal representations of a word can be formulated as $M_i = \mathcal{G}(L_i, P_i)$, where $\mathcal{G}$ is the fusion function which combines the ith word's linguistic representations L_i with its (predicted) perceptual representations P_i. In this section we describe the details of our proposed method (Figure 1): (1) build the linguistic and perceptual representations. Following most previous work, we employ visual vectors as the perceptual representations, which contain a much smaller vocabulary than linguistic vectors. (2) Learn a mapping from the linguistic to visual space. In this way, we get the predicted visual vectors for all words in linguistic vocabulary. (3) Generate multimodal representations by combining linguistic and predicted visual representations with dynamic fusion method. (4) Train the proposed model with max-margin objective function.

Obtaining Liguistic and Visual Representations

We employ the Glove vectors as our linguistic representations, which are trained by global word co-occurrence statistics. For visual representations, we employ image collections from ImageNet (Russakovsky et al. 2015), in which each image is attached to a word and each word corresponds to multiple images. To generate visual vectors for each word, we use the forward pass of a pre-trained CNN model 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.

Learning to Propagate Language to Vision

As introduced in the previous section, the words with corresponding visual images are only a small subset of the linguistic vocabulary. To obtain the visual vector for each word, we need a text-to-vision mapping function that transforms the linguistic vectors into visual ones. In this section, we introduce how to design the mapping function.

Suppose that $L \in \mathbb{R}^{m_l \times n_l}$ be the linguistic representations containing m_l words, $V \in \mathbb{R}^{m_v \times n_v}$ be the visual representations of $m_v (\ll m_l)$ words, where n_l and n_v are di-
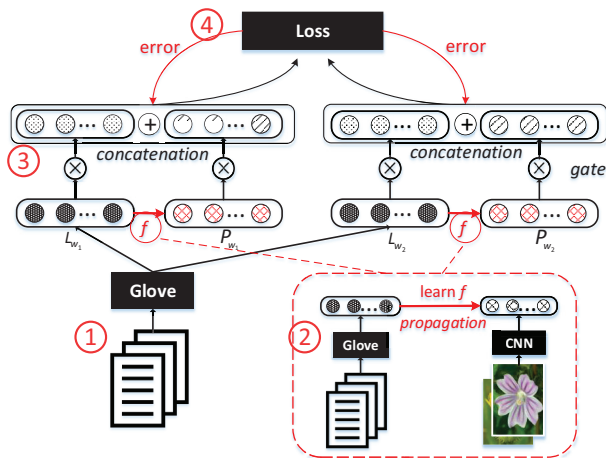

Figure 1: Overview of our model, where the four numbers correspond to four steps of our method. L_{w_1} and L_{w_2} are representations of one word association pair. P_{w_1} and P_{w_2} are the predicted visual representations from outputs of text-to-vision mapping f. The multimodal representation of a word is weighted concatenation of its linguistic vectors and its predicted vectors, in which weights are learnable parameters.

mensions of the linguistic and visual representations respectively. The linguistic representations of the m_v visual words are denoted as $L_v \in \mathbb{R}^{m_v \times n_l}$. Our goal is to learn a mapping from linguistic to visual space. To achieve this, we utilize ridge regression method which learns n_v regression coefficients $A_j \in \mathbb{R}^{n_l \times 1}$ that maps each linguistic representation L_v into a particular feature (the jth dimension of the vector) of visual representations V_j. The objective for learning A_j is then to minimize:

$$||L_v A_j - V_j||_2^2 + \lambda ||A_j||_2^2, \qquad (1)$$

where λ is the regularization parameter. Finally, all n_v coefficients of A_j are applied together to map the n_l-dimensional linguistic vectors to get the n_v-dimensional predicted visual representations $P = LA \in \mathbb{R}^{m_l \times n_v}$.

Generating Multimodal Representations

To build better multimodal representations, we explore three different *gates* (Figure 2) to learn the importance weights of textual and predicted visual representations respectively:

(1) Modality-specific gate Analysis on the inner properties of linguistic and visual vectors shows that the two vectors capture some of the same properties (Collell and Moens 2016; Wang et al. 2018), which are redundant in representing the meaning of a word. Based on this observation, we design a modality-specific gate to give a weight value or a weight vector g_L for linguistic modality and g_P for visual modality respectively.

(2) Categoty-specific gate Psychological researches (Handjaras et al. 2016) prove that human semantic representation shows clear difference between different word
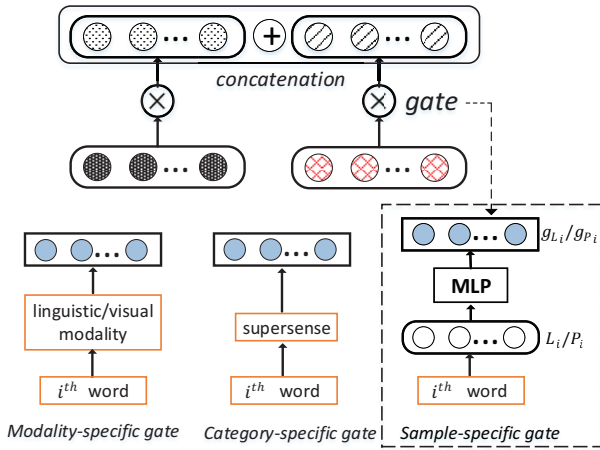
Figure 2: Illustration of three different *gates*, in which each gate can be a value gate or a vector gate. Take sample-specific gate (in the dashed box) as an example, it calculates importance weights g_{L_i} for linguistic representations and g_{P_i} for visual representations according to specific word i (in which the importance weight is either a vector parameter for vector gate or a value parameter for value gate). Similarly, the category-specific gate and modality-specific gate calculate importance weights according to modality and word supersense respectively.

categories. For instance, the category of living things on average receives higher saliency ratings on visual properties than on emotion category. To model the above observation, we design a category-specific gate to give two weight values or weight vectors (i.e., g_{L_m} for linguistic modality and g_{P_m} for visual modality) for each word supersense[2] m.

(3) Sample-specific gate   Considering that the meaning of each word has different dependencies on linguistic and visual information, we propose the sample-specific gate to assign two weight values or weight vectors (one for each modality) for each word. The weight parameters are calculated by the following feed-forward neural networks:

$$
\begin{aligned}
g_{L_i} &= tanh(W_L L_i + b_P) \\
g_{P_i} &= tanh(W_P P_i + b_P),
\end{aligned}
\tag{2}
$$

where g_{L_i} and g_{P_i} are the value gate or vector gate of the ith word's linguistic representation L_i and visual representation P_i respectively. For the value gate, W_L and W_P are vector parameters with size of $d \times 1$, and b_L and b_P are value parameters. For the vector gate, the parameters W_L and W_P are matrices with size of $d \times d$, and b_L and b_P are vectors with size of $d \times 1$.

In the above dynamic fusion methods, the value gate controls the importance weights of different input representa-

---

[2]The supersense refers to 41 WordNet's supersenses (e.g., *animal, body, food, emotion, motion*), in which we tag a word with its most frequent supersense in the sense-annotated corpora: https://github.com/UKPLab/acl2016-supersense-embeddings

tions 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 linguistic and visual representations with their corresponding gates, and concatenate the results to get the multimodal representations:

$$
M_i = \begin{cases}
[g_L \odot L_i \; ; \; g_P \odot P_i] & for \text{ M-gate} \\
[g_{L_m} \odot L_i \; ; \; g_{P_m} \odot P_i] & for \text{ C-gate} \\
[g_{L_i} \odot L_i \; ; \; g_{P_i} \odot P_i] & for \text{ S-gate}
\end{cases}
\tag{3}
$$

Where M_i is the multimodal representation of the ith word, operator $[v_1; v_2]$ denotes concatenation of vector v_1 and v_2, and $\odot$ denotes element-wise multiplication. In the C-gate model, m represents the category of the ith word.

Training Multimodal Models

The training data consists of a set W of word association pairs (w_1, w_2). To learn the model parameters of different gates p_{gates} (i.e., g_L and g_P for M-gate; g_{L_m} and g_{P_m} for C-gate; W_L, b_L, W_P and b_P for S-gate), we minimize a max-margin objective function as follows:

$$
\begin{aligned}
\sum_{(w_1, w_2) \in W} &(\max(0, 1 - M_{w_1} \cdot M_{w_2} + M_{w_1} \cdot M_{n_1}) \\
&+ \max(0, 1 - M_{w_1} \cdot M_{w_2} + M_{w_2} \cdot M_{n_2}))
\end{aligned}
\tag{4}
$$

where M_i denotes the multimodal representation of word i which can be calculated by equation (3), and n_1 and n_2 are randomly selected negative examples. The intuition for this objective is that we want the two association words to be more similar to each other than the negative examples.

Experimental Setup

Datasets

We use 300-dimensional GloVe vectors[3] which are trained on the Common Crawl corpus consisting of 840B tokens and a vocabulary of 2.2M words. 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. We use a pre-trained model of VGG-net[4] to embed visual information, resulting in 8048 vectors of 128 dimensions.

The training dataset are selected from about 20,000 word association pairs, and each word pair is generated by at least one subject[5]. We calculate the association score for each word pair (cue word + target word) as: *the number of person who generated the word pair divide the total number of person who is presented with the cue word*. To select high-quality association pairs, we delete those whose score is lower than 0.2 or with words that are not in the Glove

---

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

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

[5]The dataset is collected by (De Deyne, Perfors, and Navarro 2016) and can be found at: https://simondedeyne.me/data.

vocabulary. For better generalization ability, we delete word pairs that contain words in the testing datasets, which results in 1,494 word pairs. We use the remaining word association pairs as the development dataset (word pairs together with their association scores).

Model Settings

Our models are implemented with Theano (Bergstra et al. 2010) and Lasagne (Dieleman et al. 2015), and optimized with Adagrad (Duchi, Hazan, and Singer 2011). We test the initial learning rate over {0.05, 0.01, 0.5, 0.1}, set batch size to 25, and train the model for 5 epochs. We set the initial parameters in three gates to 1.0 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. Note that we do not update word embeddings because 1) words in the training dataset are not in the testing dataset, and 2) more importantly we aim to learn generic composition rules. The data and code for training and evaluation can be found at: https://github.com/wangshaonan/dynamicFusion

Experiments

Evaluation Tasks

We test the baseline and proposed models on 6 standard evaluation benchmarks, covering two different tasks: (i) Semantic relatedness: Men-3000 (Bruni, Tran, and Baroni 2014) and Wordrel-252 (Agirre et al. 2009); (ii) Semantic similarity: Simlex-999 (Hill, Reichart, and Korhonen 2015), 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 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.

Baseline Multimodal Models

For fair comparison, we re-implement several representative systems with our own linguistic and visual vectors. The **Concatenation (CONC) model** (Kiela and Bottou 2014) is simple concatenation of normalized linguistic and visual vectors. The **Ridge** (Hill, Reichart, and Korhonen 2014) and **Mapping** (Collell, Zhang, and Moens 2017) models first learn a mapping matrix from linguistic modality to visual modality using the ridge regression method and feedforward neural network respectively. After applying the mapping function on the linguistic representations, they obtain the predicted visual vectors for all words in linguistic vocabulary. Then they concatenate the normalized linguistic and predicted visual vectors to get multimodal representations. All above models are implemented with sklearn. Model hyper-parameters are tuned by 5-fold cross validation (20% of data for testing and 80% for training) with evaluation metric of mean square error[6]. The **Dispersion model**

computes multimodal representations by using weighted concatenation of linguistic and visual representations, in which weights are 1 or 0 for two modalities according to whether it is concrete or abstract words. Same as Kiela et al. (2014), we set the threshold (which distinguish concrete and abstract words) as the median image dispersion[7], and give zero weights to the visual representations for abstract words before concatenation.

Results and Discussion

As shown in Table 1, our proposed multimodal models clearly outperform baseline unimodal and multimodal models (in group 2 and 3). We use Wilcoxon signed-rank test to check if significant difference exists between two models. Results show that our multimodal models with vector gates perform significantly better ($p < 0.05$) than all baseline models, while the multimodal models with value gates do no show significant difference over Ridge model.

Overall performance Our multimodal models with vector gate (i.e., M-gate-vec, C-gate-vec, S-gate-vec) improve Ridge in VIS and ZS region. In other words, our models achieve better performance on words with (mostly concrete words) or without visual information (more abstract words). This suggests that the dynamic fusion methods can dynamically fuse different modality inputs. The good results in ZS region also indicate that our models have good generalization capacity. Therefore, our multimodal representations with vector gates clearly accomplish one of their foremost goals, namely to improve the multimodal representations for all types of words.

Unimodal baselines On the linguistic side, we additionally test Skip-gram model. Comparing unimodal models (in group 2), we can see that Glove outperforms Skip-gram on four datasets while Skip-gram takes superiority on the other two datasets, indicating that these two text-based models may encode different types of information. The CNN model, which learns representations from visual modality, gets worse performance than Glove and Skip-gram.

Multimodal baselines The CONC model that combines Glove vectors and CNN 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 vectors and predicted visual vectors, improve over Glove on five out of six datasets in both ALL and VIS regions. This indicates that the predicted visual vectors contains richer information than purely visual representations and are more helpful in building multimodal models. In ZS region (zero-shot region shows the result of word pairs without visual vectors), multimodal models of Mapping and Ridge only significantly outperform Glove on the *SEMSIM* dataset.

Our multimodal models Among our proposed models, the multimodal models with vector gate are clearly better than the ones with value gate (i.e., M-gate-val, C-gate-val,

---

[6]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.

[7]We calculate image dispersion with the toolkit: https://github.com/douwekiela/mmfeat

Table 1: Spearman correlations between model predictions and human ratings on six evaluation datasets. The bold scores are the best results per column and #inst. denotes the number of word pairs. For each test, ALL correspond to the whole testing set, VIS (visual) to those word pairs for which we have both visual and linguistic vectors, and ZS (zero-shot) denotes word pairs for which we hove only linguistic vectors. We divide all models into four groups: (1) previous multimodal models in which results are reprinted from (Collell, Zhang, and Moens 2017). (2) Uni-modal models with linguistic and visual inputs respectively. (3) Our re-implementation of baseline multimodal models. (4) Our proposed multimodal models with the dynamic fusion method, where M,C,S denote modality-specific, category-specific and sample-specific respectively. gate-val and gate-vec denote value gate and vector gate respectively.

	MEN			SIMLEX			SEMSIM			SIMVERB			WORDSIM			WORDREL		
	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS
Kiela & Bottou 2014	-	0.72	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Silberer & Lapata	-	-	-	-	-	-	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
Skip-gram	0.763	0.755	0.734	0.442	0.435	0.412	0.714	0.728	0.735	0.363	0.308	0.368	0.775	0.726	0.715	0.609	0.833	0.537
Glove	0.802	0.799	0.753	0.408	0.371	0.397	0.744	0.751	0.808	0.283	0.320	0.280	0.798	0.688	0.792	0.682	0.759	0.631
CNN	-	0.566	-	-	0.406	-	-	0.502	-	-	0.235	-	-	0.526	-	-	0.422	-
CONC	-	0.786	-	-	0.442	-	-	0.709	-	-	**0.437**	-	-	0.665	-	-	0.666	-
Mapping	0.806	0.815	0.729	0.408	0.407	0.374	0.769	0.771	**0.840**	0.282	0.358	0.277	0.781	0.696	0.790	0.650	0.751	0.581
Ridge	0.806	0.816	0.733	0.418	0.405	0.394	0.764	0.766	0.832	0.287	0.329	0.283	0.786	0.689	0.785	0.660	0.765	0.588
Dispersion	-	0.700	-	-	0.415	-	-	0.650	-	-	0.222	-	-	0.544	-	-	0.630	-
M-gate-val	0.808	0.815	0.741	0.417	0.400	0.396	0.765	0.768	0.834	0.288	0.324	0.284	0.792	0.697	0.791	0.672	0.762	0.603
C-gate-val	0.809	0.815	0.742	0.411	0.397	0.395	0.766	0.770	0.820	0.292	0.339	0.290	0.807	0.706	**0.801**	0.682	0.732	0.638
S-gate-val	0.813	0.808	0.766	0.418	0.389	0.408	0.762	0.764	0.833	0.289	0.352	0.286	0.801	0.687	0.795	0.682	0.763	0.630
M-gate-vec	**0.836**	0.820	**0.824**	0.466	0.430	0.465	**0.774**	**0.778**	0.826	0.347	0.371	0.348	**0.821**	**0.729**	0.782	**0.729**	0.796	**0.700**
C-gate-vec	0.820	0.803	0.813	0.463	0.435	0.452	0.766	0.771	0.801	0.359	0.314	0.363	0.803	0.716	0.777	0.695	0.749	0.672
S-gate-vec	0.829	**0.821**	0.812	**0.493**	**0.466**	**0.496**	0.740	0.745	0.793	**0.364**	0.347	**0.369**	0.804	0.723	0.752	0.717	**0.837**	0.672
#inst.	3000	1065	1935	999	261	738	7576	1789	5787	3500	41	3459	203	45	158	252	28	224

S-gate-val). This indicates that combining representations from different modalities is more complex than weighted concatenation, and thus needing deep fusion methods that can selectively combine the inside elements of different representations. Another observation is that the multimodal model with category-specific vector gate is not as effective as other two models with vector gates. This is possibly due to that the tagging process of word supersense introduces some errors.

Model Analysis

Effects of training data size To investigate the effects of training data size, we conduct experiments with less training data. As can be seen in Figure 3, decreasing the number of training data clearly harms the performance of models with vector gate, but leads to no obvious difference for the models with value gate. Additionally, we observe that the models with vector gate can obtain a quite good result with 40% (600 training pairs) of data, indicating that our dynamic fusion methods can be successfully trained with a small training set.

Effects of different gates To inspect whether the proposed models meet our expectation, i.e., assigning different weights to linguistic and visual representations for concrete words and abstract words respectively, we conduct a quantitative analysis using a set of concrete and abstract words. Specifically, we utilize the University of South Florida
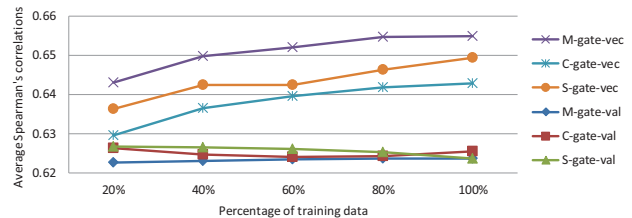

Figure 3: Effects of training data size on the model performance, which are evaluated by averaged Spearman's correlations on all evaluation datasets.

dataset (USF) [8], which includes concreteness ratings for over 6,000 words collected from thousands of participants[9]. To extract a set of abstract and concrete words, we first select words those appear in both USF dataset and the linguistic vocabulary, and order these words according to their concreteness ratings. Next we sample at random from the first and fourth quartiles, in which we get 796 concrete words and 788 abstract words. Then we examine the weights assigned to different modalities on abstract and concrete words respectively. In the following, we separately describe the results of our proposed dynamic fusion methods with different gating mechanism.

---

[8]The dataset can be download at: http://web.usf.edu/FreeAssociation/

[9]Examples of word and its concreteness are: *(tree, 7), (eye, 6.28), (wind, 5.4), (dark, 4.68), (work, 3.88), (effort, 2.22), (hope, 1.18)*.

M-gate-val obtains a weight value of 1.089 for linguistic modality, and 0.911 for visual modality.

C-gate-val learns one weight value for each supersense category in linguistic and visual modality respectively. The five categories with highest weight ratio of linguistic to visual are `Attribute`, `Location`, `Cognition`, `Quantity`, and `State`. The five categories with lowest weight ratio of linguistic to visual are `Animal`, `Object`, `Motion`, `Shape`, and `Plant`[10].

S-gate-val calculates one weight value for each word in linguistic and visual modality respectively. We then compute the average weight ratio of linguistic to visual modality respectively on the set of abstract and concrete words. As a result, we get 1.965:1 for concrete words, and 2.203:1 for abstract words. Moreover, the five words with the highest ratio of linguistic to visual are: `much`, `seem`, `curious`, `sense`, `mind`, whereas the five words with lowest ratio are: `wharf`, `walkway`, `married`, `beverage`, `tower`. In addition, we test the Spearman correlation between word concreteness and weight ratio of linguistic to visual modality for all these words, which results a correlation score of 0.614.

M-gate-vec assigns one vector for each modality. To inspect the importance weight of linguistic and visual modality, we calculate the l_2-norm of the two vectors. Finally, we get a value of 1.186 for linguistic modality and 0.045 for visual modality.

C-gate-vec learns one vector for each supersense category in linguistic and visual modality respectively. We then calculate the l_2-norm of these vectors and the weight ratio of linguistic to visual modality in each category[11]. The five categories with highest ratio are `Attribute`, `Cognition`, `State`, `Social`, and `Change`. The five categories with lowest ratio are `Shape`, `Object`, `Creation`, `Motion`, and `Plant`[12].

S-gate-vec calculates one vector for each word in linguistic and visual modality respectively. We then compute the averaged l_2-norm weight ratio of linguistic to visual modality on the set of abstract and concrete words respectively. As a result, we get 2.975:1 for concrete words, and 3.714:1 for abstract words. Moreover, the five words with the highest ratio are: `really`, `think`, `seriously`, `reason`, `believe`, and the five words with lowest ratio are: `volcano`, `palace`, `salad`, `shackle`, `tomato`. Furthermore, we test the Spearman correlation between word concreteness and weight ratio of linguistic to visual modality for all these words, in which we get a correlation score of 0.458.

From the above results, we observe that (1) the proposed models can successfully assign different weights to linguistic and visual modalities, and the learned weights show clear difference between concrete and abstract words. (2) For models with modality-specific gates (M-gate-val, M-gate-vec), representations of linguistic modality always achieve higher weights, which indicates that linguistic vectors are more important in building multimodal representations on the whole. (3) As for models with category-specific gates (C-gate-val, C-gate-vec), the categories which contain mostly abstract words achieve higher weight ratio of linguistic to visual modality, which means that the linguistic modality is more important to abstract words. (4) In models with sample-specific gates (S-gate-val, S-gate-vec), abstract words achieve higher weight ratio of linguistic to visual modality. Moreover, the learned weight-ratio shows high correlation with word concreteness, indicating that the proposed model can assist in related psycholinguistic experiments.

Conclusion and Future Work

Motivated by the fact that different semantic word representations require information from different modality inputs, in this paper we propose three simple but effective fusion methods for learning multimodal word representations. Experimental evaluations show that our proposed models achieve substantial gains in accuracy on all six benchmarks. Qualitative analyses further evidence that the proposed methods can dynamically fuse representations from different modalities according to different types of words.

Future work includes exploring better semantic word representations by combining information from other modality inputs. Moreover, the visual representations can be enhanced by utilizing more fine-grain semantic understanding of a image, which can be achieved with operations like image segmentation. We believe that one of the promising directions is learning from human semantic representation to build a more cognitive-inspired computational model (Wang, Zhang, and Zong 2017b).

Acknowledgements

References

Agirre, E.; Alfonseca, E.; Hall, K.; Kravalova, J.; Paşca, M.; and Soroa, A. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL*, 19–27.

Andrew, J. A.; Douwe, K.; Stephen, C.; and Massimo, P. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *TACL* 5:17–30.

---

[10]Example words in these categories: Attribute (*age, power, strength*), Location (*area, west, south*), Cognition (*thought, believed, known*), Quantity (*miles, meters, number*), State (*condition, problems, health*), Animal(*cow, frogs, birds*), Object (*river, valley, lake*), Motion (*follow, travel, enter*), Shape (*concave, angle, lines*), and Plant (*fruit, flowers, vines*).

[11]The higher ratio means that the category needs more linguistic information which corresponds to abstract words

[12]Example words in these categories: Social (*succeed, encourage, join*), Change (*lengthen, simplify, diminishing*), Creation (*publish, design, fix*), Motion (*drop, swim, roll*).

Andrews, M.; Vigliocco, G.; and Vinson, D. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review* 116(3):463.

Bergstra, J.; Breuleux, O.; Bastien, F.; Lamblin, P.; Pascanu, R.; Desjardins, G.; Turian, J.; Warde-Farley, D.; and Bengio, Y. 2010. Theano: A cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf*, 1–7.

Binder, J. R.; Conant, L. L.; Humphries, C. J.; Fernandino, L.; Simons, S. B.; Aguilar, M.; and Desai, R. H. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology* 33(3-4):130–174.

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

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)* 49(2014):1–47.

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

Collell, G., and Moens, M.-F. 2016. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *COLING*, 2807–2817.

Collell, G.; Zhang, T.; and Moens, M.-F. 2017. Imagined visual representations as multimodal embeddings. In *AAAI*, 4378–4384.

De Deyne, S.; Perfors, A.; and Navarro, D. J. 2016. Predicting human similarity judgments with distributional models: The value of word associations. *In COLING* 1861–1870.

Dieleman, S.; Schlüter, J.; Raffel, C.; Olson, E.; Sønderby, S. K.; Nouri, D.; Maturana, D.; Thoma, M.; Battenberg, E.; Kelly, J.; et al. 2015. Lasagne: First release. *Zenodo: Geneva, Switzerland.*

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul):2121–2159.

Fellbaum, C. 1998. *WordNet.* Wiley Online Library.

Gerz, D.; Vulić, I.; Hill, F.; Reichart, R.; and Korhonen, A. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Handjaras, G.; Ricciardi, E.; Leo, A.; Lenci, A.; Cecchetti, L.; Cosottini, M.; Marotta, G.; and Pietrini, P. 2016. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage* 135:232–242.

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

Hill, F.; Reichart, R.; and Korhonen, A. 2014. Multi-modal models for concrete and abstract concept meaning. *TACL* 2:285–296.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4).

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

Kiela, D., and Bottou, L. 2014. Learning image embeddings using convolutional neural networks for improved multimodal semantics. In *EMNLP*, 36–45.

Kiela, D.; Hill, F.; Korhonen, A.; and Clark, S. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL*, 835–841.

Lazaridou, A.; Pham, N. T.; and Baroni, M. 2015. Combining language and vision with a multimodal skip-gram model. In *NAACL*, 153–163.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Silberer, C., and Lapata, M. 2012. Grounded models of semantic representation. In *EMNLP*, 1423–1433.

Silberer, C., and Lapata, M. 2014. Learning grounded meaning representations with autoencoders. In *ACL*, 721–732.

Silberer, C.; Ferrari, V.; and Lapata, M. 2017. Visually grounded meaning representations. *IEEE transactions on pattern analysis and machine intelligence* 39(11):2284–2297.

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

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

Wang, S.; Zhang, J.; Lin, N.; and Zong, C. 2018. Investigating inner properties of multimodal representation and semantic compositionality with brain-based componential semantics. *To appear in AAAI*.

Wang, S.; Zhang, J.; and Zong, C. 2017a. Exploiting word internal structures for generic chinese sentence representation. In *EMNLP*, 298–303.

Wang, S.; Zhang, J.; and Zong, C. 2017b. Learning sentence representation with guidance of human attention. In *IJCAI*, 4137–4143.