Learning grounded word meaning representations on similarity graphs

Mariella Dimiccoli
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
mdimiccoli@iri.upc.edu

Herwig Wendt
CNRS, IRIT
Univ. of Toulouse, France
herwig.wendt@irit.fr

Pau Batlle *
California Institute of Technology
Pasadena, California

Abstract

This paper introduces a novel approach to learn visually grounded meaning representations of words as low-dimensional node embeddings on an underlying graph hierarchy. The lower level of the hierarchy models modality-specific word representations through dedicated but communicating graphs, while the higher level puts these representations together on a single graph to learn a representation jointly from both modalities. The topology of each graph models similarity relations among words, and is estimated jointly with the graph embedding. The assumption underlying this model is that words sharing similar meaning correspond to communities in an underlying similarity graph in a low-dimensional space. We named this model Hierarchical Multi-Modal Similarity Graph Embedding (HM-SGE). Experimental results validate the ability of HM-SGE to simulate human similarity judgements and concept categorization, outperforming the state of the art. [1]

1 Introduction

During the last decade, there has been an increasing interest in deriving semantic representations from text corpora in terms of word vector space for both words in context and words in isolation (Mikolov et al., 2013; Ling et al., 2015; McCann et al., 2017; Devlin et al., 2018; Peters et al., 2018). However, semantic representations are tied to sensory experience, at least for concrete nouns (Anderson et al., 2017), and deriving them by relying solely on text leads to lack of grounding on the extra-language modality. For this reason, several works have addressed the problem of grounding perceptual information in the form of visual information approximated by feature norms elicited from humans (Andrews et al., 2009; Silberer and Lapata, 2012), or extracted automatically from images (Kiela and Bottou, 2014; Lazaridou et al., 2015; Silberer et al., 2016), or a combination of them (Roller and Im Walde, 2013). Furthermore, several integration mechanisms for the linguistic and perceptual modalities have been proposed. They include methods employing transformation and dimension reduction on the concatenation of unimodal representations (Bruni et al., 2014; Hill and Korhonen, 2014); generative probabilistic models (Andrews et al., 2009; Roller and Im Walde, 2013; Feng and Lapata, 2010); deep learning methods such as autoencoders and recursive neural networks (Silberer et al., 2016; Socher et al., 2013). While the approaches above take as input previously extracted unimodal representations, (Lazaridou et al., 2015; Hill and Korhonen, 2014; Zablocki et al., 2018) learn directly from raw data by integrating both modalities non-linearly in a low-dimensional space within a skip-gram model framework. These approaches typically lead to marginal improvements since only a small part of the vocabulary is covered by a corresponding visual information. Recent work has focused on generating pre-trainable generic representations for visual-linguistic tasks by using an instance-level contextualized approach (Su et al., 2020; Lu et al., 2019; Sun et al., 2019). Such models have proved to be effective for several natural language applications such as visual question answering and visual commonsense reasoning, but their ability to simulate human behaviour phenomena has never been assessed so far.

In this paper, we propose a novel graph-based model for learning word representations across vision and language modalities that can simulate human behaviour in many NLP tasks. Our assumption is that words sharing similar meaning correspond to communities in an underlying hierarchy of graphs in low-dimensional spaces. The lower level of the hierarchy models modality-specific word represen-

* Work done during an internship at the IRI (CSIC-UPC).

[1] Code available: https://github.com/mdimiccoli/HM-SGE/.

Work partially funded by projects MINECO/ERDF RyC, PID2019-110977GA-I00, MDM-2016-0656.

tations through modality-specific but communicating graphs, while the higher level puts these representations together on a single graph to learn a single representation jointly from both modalities. At each level, the graph topology models similarity relations among words, and it is estimated jointly with the graph embedding. To the best of our knowledge this is the first model that uses a graph-based approach to learn grounded word meaning representations. Technically, our method is a joint feature learning and clustering approach. Moreover, it is compatible with both associative and domain-general learning theories in experimental psychology, following which the learning of a visual (linguistic) output is triggered and mediated by a linguistic (visual) input (Reijmers et al., 2007), and these mediated representations are then further encoded in higher-level cortices (Rogers et al., 2004). This work has applications in cognitive science, prominently in simulations of human behavior involving deep dyslexia, semantic priming and similarity judgments, among others.

The contributions of this work are as follows: 1) We propose a novel technical approach to learn grounded semantic representations as low-dimensional embeddings on a hierarchy of similarity graphs, 2) we achieve state-of-the-art performance with respect to several baselines for simulating human behaviour in the tasks of similarity rating and categorization, 3) we demonstrate the ability of our model to perform inductive inference, 4) we validate the proposed approach through an extensive ablation study, and provide insights about the learnt representations through qualitative results and visualizations, 5) our model is compatible with associative and domain-general learning theories in experimental psychology.

2 Related work

Distributional Semantic Models (DSMs) are based on the distributional hypothesis (Harris, 2013), following which words that appear in similar linguistic contexts are likely to have related meanings. DSMs associate each word with a vector (a.k.a. word embedding) that encodes information about its co-occurrence with other words in corpora (Mikolov et al., 2013; Ling et al., 2015). Recently, instance-level contextualized word embeddings (Devlin et al., 2018) have emerged as a natural extension of type-level non-contextualized DSMs and have demonstrated their effectiveness with respect to their counterpart in a wide variety of common NLP tasks (McCann et al., 2017; Devlin et al., 2018; Peters et al., 2018).

However, humans learn the verbal description of objects by hearing words while looking at /listening to/interacting with objects. Therefore, in recent years there has been an increasing interest in developing linguistic models augmented with perceptual information. These are commonly called *grounded semantic spaces*. Following the classification introduced by (Collell et al., 2017), we can distinguish two integration strategies: 1) *a posteriori combination*, where each modality is learnt separately and they are integrated afterwards, and 2) *simultaneous learning*, where a single representation is learnt from raw input data enriched with both modalities.

A posteriori combination. Several works aimed at projecting directly vision and language into a common space. Among them, (Bruni et al., 2014) concatenate and project two independently constructed textual and visual spaces onto a lower-dimensional space using Singular Value Decomposition (SVD). Other approaches along the same line build on extensions of topic models as Latent Dirichlet Allocation (LDA), where topic distributions are learnt from the observed variables (words and other perceptual units) (Andrews et al., 2009; Roller and Im Walde, 2013; Feng and Lapata, 2010). In (Kiela and Bottou, 2014) an empirical improvement is obtained by using state-of-the-art convolutional neural networks to extract visual features, and the skip-gram model for textual features, that are simply concatenated. In (Silberer et al., 2016) a stacked auto-encoder framework is used to learn a representation by means of an unsupervised criterion (the minimization of the reconstruction error of the attribute-based representation) and then fine-tuned with a semi-supervised criterion (object classification of the input). In the approach proposed by (Wang et al., 2018) the weights of the unimodal feature concatenation are learnable parameters that allow to dynamically fuse representations from different modalities according to different types of words. Other works focus on learning bimodal representations that are task-specific, with the goal of reasoning about one modality given the other (Lazaridou et al., 2014; Socher et al., 2013). For example, the aim of image retrieval is to find a mapping between two modalities to tackle an image based task such as zero-shot learning (Frome et al., 2013) or caption generation/retrieval and

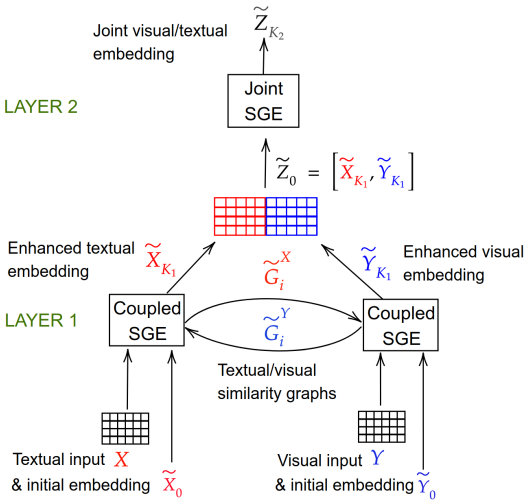


Figure 1: HM-SGE model overview. The first layer learns unimodal representations conditioned to the other modality via its corresponding similarity graph. The concatenation of the enhanced representations is then used to learn jointly a common representation.

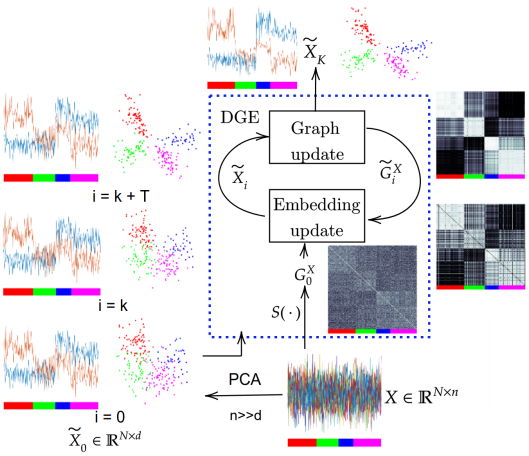


Figure 2: DGE model of which our HM-SGE modules share the high-level structure. A similarity graph is estimated from the initial high dimensional embeddings that are then projected in a low-dimensional space. DGE alternates a graph embedding step with a graph update step.

caption-based image retrieval (Socher et al., 2014). These models typically use a training criterion and an architecture suited to the task at hand.

Simultaneous learning. Little work has explored the possibility of learning multimodal representations directly from raw input data, i.e, images and corpora, building on the skip-gram framework. (Hill and Korhonen, 2014) treat perceptual input as a word linguistic context and has proved to be effective in propagating visual knowledge into abstract words. (Lazaridou et al., 2015) modify the skip-gram objective function to predict both visual and linguistic features and is especially good in zero-shot image classification. (Zablocki et al., 2018) contributed to this research line by leveraging the visual surroundings of objects to fulfill the distributional hypothesis for the visual modality. This class of approaches typically leads only to a small empirical improvement of linguistic vectors since words from the raw text corpus associated with images (and hence perceptual information) cover only a small portion of the training dataset. In the last few years, increasing efforts are being devoted to deriving generic pre-trainable representations for visual-linguistic tasks based on transformers (Sun et al., 2019; Lu et al., 2019; Su et al., 2020).

Graph-based word meaning representations. Despite the success of graph-based models for sentence meaning (Koller et al., 2019), their use for encoding word meaning representation has been little explored so far. Recently, Graph Convolutional Network (GCN) based approaches have been used to improve text-based word meaning representation by incorporating syntactic information (Vashishth et al., 2019; Tran et al., 2020; Ryabinin et al., 2020). To the best of our knowledge, graph based models have never been exploited for grounded word meaning representations.

3 Proposed approach

Model assumptions and psychological foundations. We assume that each modality, linguistic (X) and visual (Y), can be well represented by an unknown underlying graph in a low-dimensional space, in which nodes correspond to words and weighted edges to respective word similarity. The associative nature of human memory, a widely accepted theory in experimental psychology (Anderson and Bower, 2014), suggests that beyond being complementary, there exists a certain degree of correlation between these modalities (Reijmers et al., 2007). Therefore, in a first step, our model aims at exploiting these correlations to enhance each unimodal representation individually. This simulates the fact that when humans face, for instance, the task of visual similarity rating of a pair of images, their mental search necessarily includes both components, the visual and the semantic one.

In a second step, these enhanced unimodal representations are mapped into a common embedding space by inducing semantic representations that

integrate both modalities. This agrees with neuroscientific findings showing that it is unlikely that humans have separate representations for different aspects of word meaning (Rogers et al., 2004). Overall, the proposed two layer structure, where the first layer has two modality-specific branches that provide input to a second single-generic branch, is compatible with theories of both associative memory (Reijmers et al., 2007) and domain-general learning (Rogers et al., 2004).

Algorithm 1 Hierarchical Multimodal SGE

N − *number of words*
Input : $X \in R^{N\times n}$, $Y \in R^{N\times m}$ *V and T feature matrices*
Output : $\tilde{X} \in R^{N\times d} (d \ll n)$, $\tilde{Y} \in R^{N\times p} (p \ll m)$,
$\tilde{Z} \in R^{N\times d+p}$ *graph embedded feature matrices*

/* **Initialize embeddings and graphs** */
$\tilde{X} = \text{PCA}_d(X) \in R^{N\times d}, \tilde{Y} = \text{PCA}_p(Y) \in R^{N\times p}$
$\tilde{\text{G}}_0^X = S_l(\tilde{X}) \in R^{N\times N}, \tilde{\text{G}}_0^Y = S_l(\tilde{Y}) \in R^{N\times N}$
$\text{G}_0^X = S_l(X) \in R^{N\times N}, \text{G}_0^Y = S_l(Y) \in R^{N\times N}$

/* **Layer 1: Coupled SGE loop** */
for $i \leftarrow 1$ **to** K_1 **do**
— graph structure update: semantic prior
$\tilde{\text{G}}_i^X \leftarrow g_\mu(\tilde{\text{G}}_i^X, \text{kmeans}(\tilde{X}_i; N_C)), \tilde{\text{G}}_i^Y \leftarrow g_\mu(\tilde{\text{G}}_i^Y, \text{kmeans}(\tilde{X}_i; N_C))$
estimate communities and encode similarity in graph
— graph embedding update
for $(A, B) \in \{(X, Y), (Y, X)\}$
$\tilde{A}_i = {}_{\tilde{A}}(1-\alpha^A)\mathcal{L}(S_l(\tilde{A}), \tilde{\text{G}}_{i-1}^A) + \alpha^A \mathcal{L}(S_l(\tilde{A}), \text{G}_0^A) + \beta^A \mathcal{L}(S_l(\tilde{A}), \tilde{\text{G}}_{i-1}^B)$
update embedded features given current graphs
end

/* **Initialize embedding and graph** */
$\tilde{Z}_0 = (\tilde{X}, \tilde{Y}), \text{G}_0^Z = \text{G}^Z = S_l(\hat{Z}) \in R^{N\times N}$
/* **Layer 2: Joint SGE loop** */
for $i \leftarrow 1$ **to** K_2 **do**
— graph structure update: semantic prior
$\tilde{\text{G}}_i^Z \leftarrow g_\mu(\tilde{\text{G}}_i^Z, \text{kmeans}(\tilde{Z}_i; N_C))$
estimate communities and encode similarity in graph
— graph embedding update
$\tilde{Z}_i = {}_{\tilde{Z}}\ (1-\alpha^Z)\mathcal{L}_1(S_l(\tilde{Z}), \hat{\text{G}}_{i-1}^Z) + \alpha^Z \mathcal{L}_2(S_l(\tilde{Z}), \text{G}_0^Z)$
update features given current graph
end

Architecture overview. Fig. 1 illustrates the proposed model. It takes as input linguistic (X) and visual (Y) vector representations. In the first layer, the initial embedding of both modalities are enhanced individually by relying on the other modality. In the last layer, the conditional embeddings are concatenated and jointly optimized. For each modality, we first build a fully-connected similarity graph G_0^X from the initial features $X \in \mathbb{R}^{N\times n}$, where N is the number of samples and n is the feature dimension. G_0^X subsequently serves to regularize the process of jointly learning the underlying graph and an embedding $\tilde{X} \in \mathbb{R}^{N\times d}$ of dimension $d \ll n$, which is achieved by alternating two steps: at iteration i, 1) update node embeddings $\tilde{X}_i$ by taking into account the current graph estimate $\tilde{\text{G}}_{i-1}^X$ (reflected by edge weights) and that of the other modality, and 2) update the graph estimate $\tilde{\text{G}}_i^X$ (fully connected similarity graph of $\tilde{X}_i$) by taking into account semantic similarity priors.

Conceptually, the alternating of a graph embedding step and a graph structure update step of our SGE is similar to the recently introduced Dynamic Graph Embedding (DGE) (Dimiccoli and Wendt, 2020) (see Fig.2), where a low-dimensional embedding is learnt from image sequences for the downstream task of temporal segmentation. However, we changed the way each of theses steps is formulated. Firstly, to learn unimodal (visual or textual) representations that take into account semantic communities in both graphs (textual and visual), we allow each modality to share its graph with the other modality. This is achieved by modifying the embedding loss during the embedding step. Secondly, since our data are not temporally linked, we do not model temporal constraints in the clustering step, but just semantic similarity among words. Finally, we propose a two layers hierarchical architecture, that is tailored to the learning of visually grounded meaning representations.

Coupled similarity graph embedding update. Formally, the embedding update for $\tilde{X}$ (and analogously for $\tilde{Y}$) is computed as:

$$\tilde{X}_i = \arg\min_{\tilde{X}} (1-\alpha^X)\mathcal{L}(S_l(\tilde{X}), \tilde{\text{G}}_{i-1}^X) + \alpha^X \mathcal{L}(S_l(\tilde{X}), \text{G}_0) + \beta^X \mathcal{L}(S_l(\tilde{X}), \tilde{\text{G}}_{i-1}^Y), \quad (1)$$

where $\mathcal{L}$ is a cross-entropy loss function that includes normalization of its arguments and S_l stands for a cosine-distance based pairwise similarity function with exponential kernel of bandwidth l. The first term in (1) controls the fit of the representation $\tilde{X}$ with the learnt graph $\tilde{\text{G}}$ in low-dimensional embedding space, while the second term ensures that it keeps also aspects of the initial graph G_0^X; $\alpha \in [0, 1]$ controls the relative weight of the terms; the hyperparameters (β^X, β^Y) tune the respective weights of the graphs of the other modalities in the unimodal representations.

Similarity graph update. To obtain an update for the graph at the i-th iteration, assuming that $\tilde{X}_i$ is given, the model starts from an initial estimate as $\tilde{\text{G}}_i^X = S_l(\tilde{X}_i)$ and makes use of the

model assumptions to modify $\tilde{\mathrm{G}}_i^X$. In particular, the semantic prior assumes that the most similar nodes form communities in the graph and leads to decreasing the edge weights between nodes of $\tilde{\mathrm{G}}_i^X$ that do not belong to the same community. Practically, this can be implemented by estimating communities using clustering, e.g. k-means with N_C classes, and multiplying the edge weights for node pairs belonging to different communities by a factor $\mu \in (0, 1)$; we denote this operation as $\tilde{\mathrm{G}}_i^X \leftarrow g_\mu(\tilde{\mathrm{G}}_i^X, \mathrm{kmeans}(\tilde{X}_i; N_C))$.

Joint similarity graph embedding update. After K_1 iterations, the learnt representations are concatenated, $\tilde{Z}_0 = (\tilde{X}, \tilde{Y})$, and input to the second layer of our model. It learns the joint representation $\hat{Z}$ that integrates both modalities as node embeddings on an underlying graph encoding visually grounded word meaning. The last term of Eq. (1) is omitted at this stage. A detailed description of our framework is given in Algorithm 1.

4 Experimental results

4.1 Experimental setting

Visual and textual representations. As visual and textual input feature vectors we used the attribute based representations proposed by (Silberer et al., 2016). Specifically, the visual modality is encoded via 414-dimensional vectors of attributes obtained automatically from images by training a SVM-based classifier for each attribute on the VISA dataset (Silberer et al., 2016). More specifically, we used initial meaning representations of words for the McRae nouns (McRae et al., 2005) covered by the VISA dataset (Silberer et al., 2016), that consists exclusively of concrete nouns.

The textual modality was encoded in two different ways: through textual attributes and via word embeddings. Textual attributes were extracted by running Strudel (Baroni et al., 2010) on the WaCkypedia corpus (Baroni et al., 2009), and by retaining only the ten attributes with highest log-likelihood ratio scores for each target word. The union of the selected attributes leads to 2,362 dimensional textual vectors. Word embeddings were obtained by training the skip-gram model (Mikolov et al., 2013) on the WaCkypedia corpus (Baroni et al., 2009), resulting in 500-dimensional embedding vectors. The attribute-based representations in both modalities were scaled to the $[-1, 1]$ range.

Hyperparameter settings. We use a common embedding dimension $d = p = 15$ for both modalities. The SGE hyperparameters are set to $(\alpha^X, \mu^X, N_C^X) = (0.1, 0.95, 25)$ and $(\alpha^Y, \mu^Y, N_C^Y) = (0.3, 0.7, 5)$ for the first layer of our model. For the second layer of our model, we set $(\alpha^Z, \mu^Z, N_C^Z) = (0.05, 0.7, 20)$ and $(0.1, 0.7, 6)$ when using textual or skip-gram as input, respectively. The number of SGE iterations are set to $K_1 = 4$ or 5 (first layer) and $K_2 = 2$ or 5 (second layer) when textual attributes or skip-gram representations are used for the textual modality, respectively. These hyperparameters have been optimized for each SGE individually using grid search; the cross-coupling parameters were also optimized separately and set to $(\beta^X, \beta^Y) = (0.01, 0.1)$.

Performance measures. Similarly to previous work (Silberer et al., 2016), we evaluate our model on two different semantic tasks, namely word similarity rating and categorization. Specifically, we measure how well our model predictions of word similarity correlate with human semantic and visual similarity ratings using Spearman's correlation. Our similarity ratings are calculated as the cosine similarity of learnt representation vectors. As human similarity ratings, we used those published in (Silberer et al., 2016). The semantic categories are induced by following a clustering-based approach, namely the Chinese Whispers algorithm (Biemann, 2006), and the quality of the clusters produced was evaluated using the F-score measure introduced in the SemEval 2007 task (Agirre and Soroa, 2007).

Computation. The complexity of our approach is $\mathcal{O}(N^2)$. Experiments were conducted on a 2018 Dell Precision T7920 workstation with 64GB RAM and a single NVIDIA Titan XP GPU.

4.2 Comparative results.

In Tab. 1, we compare our HM-SGE to the state of the art. Results are presented for two different sets of input features: attribute vectors for the visual modality (vAttrib) combined with either attribute vectors (tAttrib) or skip-gram encoding (skip-gram) for the textual modality (top and center part of Tab. 1, respectively). The bottom part of Tab. 1 presents comparisons with models for raw data (Lazaridou et al., 2015) and (Bruni et al., 2014) and pre-trained VL-BERT model (Su et al., 2020) [2] for which we derived the type-level representations.

[2] https://github.com/jackroos/VL-BERT

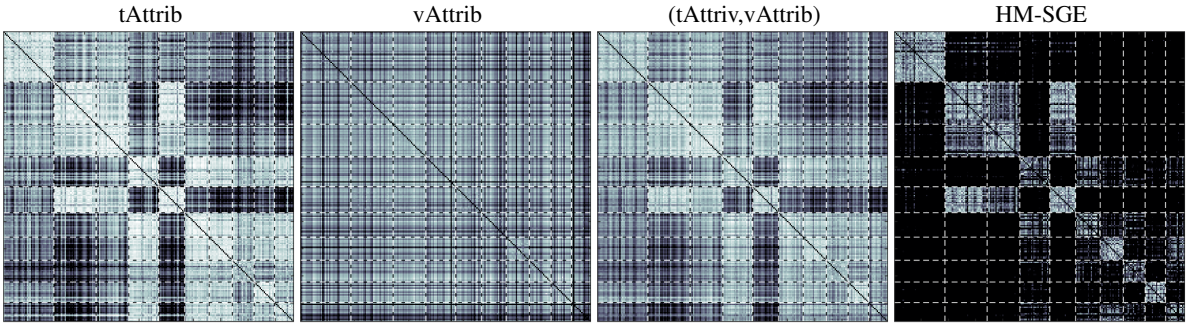


Figure 3: Similarity matrices of (a) textual attributes, (b) visual attributes, (c) visual and textual attributes (concatenation), (d) HM-SGE representations for the 10 categories *food, animal, bird, tools, mammal, weapon, instrument, transportation, clothing, device* (boundaries indicated by dashed lines). Image intensity values are individually clipped to cover the upper 3/8 of observed similarity values.

Models	Semantic Similarity			Visual Similarity			Categorization		
	T	V	T+V	T	V	T+V	T	V	T+V
HM-SGE (tAttrib, vAttrib)	**0.74**	**0.68**	**0.77**	**0.59**	**0.60**	0.64	**0.43**	**0.39**	**0.45**
SAE (tAttrib, vAttrib)	0.67	0.61	0.72	0.55	**0.60**	**0.65**	0.36	0.35	0.43
GCN (tAttrib, vAttrib)	—	—	0.74	—	—	0.59	—	—	0.42
SVD (tAttrib, vAttrib)	—	—	0.70	—	—	0.59	—	—	0.39
CCA (tAttrib, vAttrib)	—	—	0.58	—	—	0.56	—	—	0.37
CONC (tAttrib, vAttrib)	0.63	0.62	0.71	0.49	0.57	0.60	0.35	0.37	0.33
CTL (tAttrib, vAttrib)	0.67	0.64	0.69	0.54	0.57	0.58	0.38	0.29	0.30
HM-SGE (skip-gram, vAttrib)	**0.79**	**0.68**	**0.78**	**0.62**	**0.60**	0.65	**0.45**	**0.40**	0.47
SAE (skip-gram, vAttrib)	0.74	0.61	0.77	0.59	**0.60**	**0.66**	0.44	0.35	**0.48**
GCN (skip-gram, vAttrib)	—	—	0.76	—	—	0.60	—	—	0.42
SVD (skip-gram, vAttrib)	—	—	0.75	—	—	0.63	—	—	0.43
CCA (skip-gram, vAttrib)	—	—	0.59	—	—	0.57	—	—	0.35
CONC (skip-gram, vAttrib)	0.71	0.62	0.75	0.56	0.57	0.62	0.37	0.37	0.45
CTL (skip-gram, vAttrib)	0.74	0.67	0.71	0.60	0.57	0.60	0.44	0.29	0.43
VL-BERT	—	—	0.66	—	—	0.56	—	—	0.36
Lazaridou et al.	0.70	0.62	0.70	0.55	0.57	0.61	0.37	0.37	0.39
Bruni et al.	—	—	0.50	—	—	0.44	—	—	0.34

Table 1: Comparative results in terms of Spearman's correlations between model predictions and human similarity ratings, and categorization on two sets of input features (tAttrib,vAttrib) and (skip-gram,vAttrib). Here T, V, T+V denote textual, visual and textual&visual. The bold scores are the best results per semantic task. Since the only stochastic part of our algorithm is the use of several random centroid seeds for kmeans, it can effectively be considered deterministic and results vary extremely little.

Each column of Tab. 1 corresponds to a unimodal (textual (T) or visual (V)) or joint (T+V) representation.

As baseline methods trained with the same attribute based input, we considered: SAE (Silberer et al., 2016), SVD (Bruni et al., 2014), CCA (Hill and Korhonen, 2014) and cross-transfer learning (CTL) (Both et al., 2017) models trained on the same attribute-based input as in (Silberer et al., 2016); SVD and CCA models first concatenate normalized textual and visual vectors and then conduct SVD or CCA; CONC stands for concatenation of normalized textual and visual vectors (Kiela and Bottou, 2014). CTL transfer information from the common space to unimodal representations by using either a mask (estimated by correlation, or multilinear regression) or a function generating artificial features (linear, or a neural net) estimated from multimodal examples. We implemented the mask approach and reported the better of the results for correlation or multilinear regression.

Moreover, we provide a novel graph-based baseline that learns word embeddings via a two-layer Graph Convolutional Networks (GCN) trained to classify words. The graph structure (edges) was created by using visual features, and node embeddings were initialized by textual features.

The unimodal (output by layer 1) representations of our HM-SGE always achieve state-of-the-art results, largely outperforming the SAE method (up

Models	Semantic Similarity			Visual Similarity			Categorization		
	T	V	T+V	T	V	T+V	T	V	T+V
HM-SGE (tAttrib, vAttrib)	—	—	**0.77**	—	—	**0.64**	—	—	**0.45**
Layer2 only (tAttrib, vAttrib)	—	—	0.76	—	—	0.63	—	—	0.44
Layer1 only (tAttrib, vAttrib)	**0.74**	**0.68**	0.76	**0.59**	**0.60**	0.63	**0.43**	**0.39**	0.44
tAttrib + vAttrib	0.63	0.62	0.71	0.49	0.57	0.60	0.35	0.37	0.33
HM-SGE (skip-gram, vAttrib)	—	—	**0.78**	—	—	**0.65**	—	—	**0.47**
Layer2 only (skip-gram, vAttrib)	—	—	0.77	—	—	0.64	—	—	0.46
Layer1 only (skip-gram, vAttrib)	**0.79**	**0.68**	**0.78**	**0.62**	**0.60**	0.64	**0.45**	**0.40**	0.46
skip-gram + vAttrib	0.71	0.62	0.75	0.56	0.57	0.62	0.37	0.37	0.45

Table 2: Ablation study. T, V, T+V denote textual, visual and textual&visual. The best results are in bold.

Word pairs 1-4	Word pairs 5-8	Word pairs 9-12
lettuce-spinach	cello-violin	cloak-robe
clarinet-trombone	leopard-tiger	cabbage-spinach
cabbage-lettuce	raspberry-strawberry	pants-shirt
airplane-helicopter	chapel-church	blouse-dress

Word clusters
catfish, cod, crab, eel, guppy, mackerel, minnow, octopus perch, salmon, sardine, squid, trout, tuna
ambulance, bus, car, jeep, limousine, taxi, trailer, train truck, van
bike, buggy, cart, dunebuggy, motorcycle, scooter, tractor tricycle, unicycle, wagon
ant, beetle, butterfly, caterpillar, cockroach, grasshopper hornet, housefly, moth, spider, wasp, worm
apartment, barn, brick, bridge, building, bungalow, cabin cathedral, chapel, church, cottage, fence, gate, house, hut inn, pier, shack, shed, skyscraper

Table 3: Left: Word pairs with highest semantic and visual similarity according to HM-SGE model. Pairs are ranked from highest to lowest similarity. Right: Examples of clusters produced by CW using semantic representations obtained with our HM-SGE.

to $+7\%$, $+4\%$ on average). The unified representations of our HM-SGE (output by layer 2 following layer 1) achieves state-of-the-art results in most cases: for semantic similarity rating and categorization when using textual attributes, and for semantic similarity only when using the skip-gram model for the textual modality (up to $+5\%$, $+3\%$ on average). In the other cases (visual similarity ratings; categorization for skip-gram model) the unified representations also achieve results comparable to the best performing method (SAE), up to 1% difference. Overall, we improved reported performance measures with respect to the SOTA, by up to 7% on average for the 18 cases, in particular by 5%(semantic similarity), 1% (visual similarity) and 3% (categorization).

Our model also outperforms embeddings obtained using the pretrained VL-BERT by a large extent. This is not surprising considering that recent studies (Mickus et al., 2021; Rogers et al., 2020) have raised concerns about the coherence of BERT (text-based) embedding space. Further, as one would expect, tAttrib dominates vAttrib when modeling semantic similarity and categorization, while vAttrib dominates tAttrib for visual similarity ratings only. More importantly, the joint use of tAttrib and vAttrib improves all evaluation metrics, hence corroborating the fact that the model has learnt to leverage on their redundancy and complementarity. Joint representations also improve performance when based on skip-gram encoding, except for semantic similarity, which is strongly dominated by the skip-gram features. Examples of our model output are given in Tab. 3, showing word pairs with highest similarity rating (left) and examples of word clusters (categories, right).

4.3 Model validation and illustration

Ablation study. To validate the proposed HM-SGE model, in Tab. 2 we report results obtained with HM-SGE (top rows), and with HM-SGE upon removal of one of its layers: removal of the two coupled SGE (second rows, Layer2), removal of final SGE (third rows, Layer1), no HM-SGE (bottom rows); for rows 2 to 4, joint representations (T+V) are obtained upon concatenation of the individual visual and textual representations. It can be seen that Layer1 as well as Layer2 alone yield significant performance improvements when compared with the initial representations (T, V and T+V). This validates the independent capabilities of the individual components of our model to learn meaningful word representations. Yet, best performance for the joint representation (T+V) are obtained only upon combination of the two layers, demonstrating the importance of both layers in our model.

Qualitative results. An illustration corresponding to the rows 1 and 4 of Tab. 2 is provided in Fig. 3, which plots the affinity matrices for textual attributes (T), visual attributes (V), concatenation thereof (T+V) and HM-SGE for the 10 categories *food, animal, bird, tools, mammal, weapon, instrument, transportation, clothing, device*; category

Concept	Textual NN	Visual NN	HM-SGE NN
cabin	hut, tent	house, cottage, hut, bungalow	hut, cottage, house, shack, bungalow
ox	cow	bull, pony, cow, calf, camel, pig, sheep, lamb	bull, cow, pony, sheep
sardine	tuna	trout	tuna, salmon, trout
bagpipe	accordion	clamp, accordion, tuba, faucet	accordion, tuba
hamster	chipmunk, squirrel	rat, squirrel	squirrel, chipmunk, rat, groundhog
spoon	bowl	ladle, whistle, hammer	ladle

Table 4: Nearest neighbors (NN, words with similarity larger than 0.9 times that of best pair), for textual, visual and MM-SGE vectors.

boundaries are indicated by dashed lines, and image intensity values are clipped to cover the upper 3/8 of observed similarity values for each representation individually. It is observed that T yields visually better results than V, and the concatenation T+V inherits similarity from both attributes but is dominated by T. The observed large off-diagonal similarity values for T, V and T+V lead to expect that categorization results based on these attributes are poor. The affinity obtained by HM-SGE is more structured, with large values essentially coinciding with within-category affinities (the 10 diagonal blocks) and a few off-diagonal blocks. Interestingly, these off-diagonal blocks correspond with the arguably meaningful two groups of categories *animal, bird, mammal* (rows-columns 2,3,5) and *tools, weapon* (rows-columns 4,6).

Finally, Tab. 4 exemplifies such results and provides a different view by showing the nearest neighbors (NN) for six words in terms of similarity computed on textual attributes, visual attributes, and our HM-SGE attributes; all neighbors with similarity values of at least 90% that of the closest neighbor are given. The examples illustrate that the unified word meaning representations learnt by HM-SGE lead to NN that are not a simple union or intersection of visual and textual NN, but that HM-SGE is capable of removing pairs that make less sense (e.g. *tent* for *cabin*; *calf,camel,pig* for *ox*; *clamp,faucet* for *bagpipe*; *bowl,whistle,hammer* for *spoon*) and can identify and add new meaningful pairs (e.g. *salmon* for *sardine*; *groundhog* for *hamster*).

Inductive inference. It is possible to use our model to perform inductive inference when one of the two modalities for a concept, say modality A, is missing. To this end, it suffices to replace in Eq. (1) the corresponding row and column of the matrix $\tilde{\mathrm{G}}^{A}_{i-1}$ with those of $\tilde{\mathrm{G}}^{B}_{i-1}$ for the other modality B. For example, when only the visual component $\tilde{X}$ for the concept *bluejay* is given, the textual representation $\tilde{Y}$ learnt by HM-SGE outputs the nearest neighbors *robin, stork, falcon, finch*, which are all birds; to give another example, from the visual attribute for *shelves* HM-SGE predicts textual nearest neighbors *dresser, cabinet, cupboard, bureau, desk, closet*. Analogously, HM-SGE predicts visual nearest neighbors *shelves, cabinet* and *dagger, knife, spear* for *cupboard* and *sword*, respectively, when only the textual attribute is given for these two concepts.

5 Conclusion

This paper has proposed a novel approach, named HM-SGE, to learn grounded word meaning representations as low-dimensional node embeddings on a hierarchy of graphs. The first layer of the hierarchy encodes unimodal representations conditioned to the other modality, and the second layer integrates these enhanced unimodal representations into a single one. The proposed HM-SGE approach is compatible with theories of associative memory (Reijmers et al., 2007) and of domain-general learning (Rogers et al., 2004). Comparative results on word similarity simulation and word categorization show that our model outperforms baselines and related models trained on the same attribute-based input. Our evaluation reveals that HM-SGE is particularly good at learning enhanced unimodal representations that simulate how the response of our brain to semantic tasks involving a single modality is always triggered by other modalities. Moreover, it succeeds in encoding these unimodal representations into a meaningful unified representation, compatible with the point of view of domain-general learning theory. The ablation study thoughtfully validates the proposed hierarchical architecture. Beside quantitative results, we give several insights on the learnt grounded semantic space through visualization of nearest neighbors, clusters, and most similar pairs. These additional results corroborate the quality of the learnt multimodal representations. Furthermore, the proposed approach is able to perform inductive inference for concepts for which only one modality is available.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, pages 7–12.

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. *Transactions of the Association for Computational Linguistics*, 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.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive science*, 34(2):222–254.

Chris Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pages 73–80. Association for Computational Linguistics.

Fabian Both, Steffen Thoma, and Achim Rettinger. 2017. Cross-modal knowledge transfer: Improving the word embedding of apple by looking at oranges. In *Proceedings of the Knowledge Capture Conference*, pages 1–8.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mariella Dimiccoli and Herwig Wendt. 2020. Learning event representations for the temporal segmentation of image sequences by dynamic graph embedding. *IEEE Transactions on Image Processing*.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Adv. neural information processing systemsmm*, pages 2121–2129.

Zellig S Harris. 2013. *Papers in structural and transformational linguistics*. Springer.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 255–265.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*, pages 36–45.

Alexander Koller, Stephan Oepen, and Weiwei Sun. 2019. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *31st Conference on Neural Information Processing Systems (NeurIPS)*.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2021. What do you mean, bert? assessing bert as a distributional semantics model. *Proceedings of the Society for Computation in Linguistics: Vol. 3, Article 34*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Adv. neural information processing systemsmm*, 26:3111–3119.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Leon G Reijmers, Brian L Perkins, Naoki Matsuo, and Mark Mayford. 2007. Localization of a stable neural correlate of associative memory. *Science*, 317(5842):1230–1233.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Timothy T Rogers, Matthew A Lambon Ralph, Peter Garrard, Sasha Bozeat, James L McClelland, John R Hodges, and Karalyn Patterson. 2004. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, 111(1):205.

Stephen Roller and Sabine Schulte Im Walde. 2013. A multimodal lda model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157.

Max Ryabinin, Sergei Popov, Liudmila Prokhorenkova, and Elena Voita. 2020. Embedding words in non-vector space with unsupervised graph learning. *arXiv preprint arXiv:2010.02598*.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2016. Visually grounded meaning representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2284–2297.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Adv. neural information processing systemsmm*, 26:935–943.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Thy Thy Tran, Makoto Miwa, and Sophia Ananiadou. 2020. Syntactically-informed word representations from graph neural network. *Neurocomputing*, 413:431–443.

Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, and Partha Talukdar. 2019. Incorporating syntactic and semantic information in word embeddings using graph convolutional networks.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018. Learning multimodal word representation via dynamic fusion methods. *arXiv preprint arXiv:1801.00532*.

Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2018. Learning multi-modal word representation grounded in visual context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.