Do Neural Network Cross-Modal Mappings Really Bridge Modalities?

Guillem Collell
Department of Computer Science
KU Leuven
`gcollell@kuleuven.be`

Marie-Francine Moens
Department of Computer Science
KU Leuven
`sien.moens@cs.kuleuven.be`

Abstract

Feed-forward networks are widely used in cross-modal applications to bridge modalities by mapping distributed vectors of one modality to the other, or to a shared space. The predicted vectors are then used to perform e.g., retrieval or labeling. Thus, the success of the whole system relies on the ability of the mapping to make the neighborhood structure (i.e., the pairwise similarities) of the predicted vectors akin to that of the target vectors. However, whether this is achieved has not been investigated yet. Here, we propose a new similarity measure and two ad hoc experiments to shed light on this issue. In three cross-modal benchmarks we learn a large number of language-to-vision and vision-to-language neural network mappings (up to five layers) using a rich diversity of image and text features and loss functions. Our results reveal that, surprisingly, the neighborhood structure of the predicted vectors consistently resembles more that of the input vectors than that of the target vectors. In a second experiment, we further show that untrained nets do not significantly disrupt the neighborhood (i.e., semantic) structure of the input vectors.

1 Introduction

Neural network mappings are widely used to bridge modalities or spaces in cross-modal retrieval (Qiao et al., 2017; Wang et al., 2016; Zhang et al., 2016), zero-shot learning (Lazaridou et al., 2015b, 2014; Socher et al., 2013) in building multimodal representations (Collell et al., 2017) or in word translation (Lazaridou et al., 2015a), to name a few. Typically, a neural network is firstly trained

to predict the distributed vectors of one modality (or space) from the other. At test time, some operation such as retrieval or labeling is performed based on the nearest neighbors of the predicted (mapped) vectors. For instance, in zero-shot image classification, image features are mapped to the text space and the label of the nearest neighbor word is assigned. Thus, the success of such systems relies entirely on the ability of the map to make the predicted vectors similar to the target vectors in terms of semantic or neighborhood structure.[1] However, whether neural nets achieve this goal in general has not been investigated yet. In fact, recent work evidences that considerable information about the input modality propagates into the predicted modality (Collell et al., 2017; Lazaridou et al., 2015b; Frome et al., 2013).

To shed light on these questions, we first introduce the (to the best of our knowledge) first existing measure to quantify *similarity between the neighborhood structures of two sets of vectors*. Second, we perform extensive experiments in three benchmarks where we learn image-to-text and text-to-image neural net mappings using a rich variety of state-of-the-art text and image features and loss functions. Our results reveal that, contrary to expectation, the semantic structure of the mapped vectors consistently resembles more that of the input vectors than that of the target vectors of interest. In a second experiment, by using six concept similarity tasks we show that the semantic structure of the input vectors is preserved after mapping them with an untrained network, further evidencing that feed-forward nets naturally preserve semantic information about the input. Overall, we uncover and rise awareness of a largely

---

[1] We indistinctly use the terms *semantic structure*, *neighborhood structure* and *similarity structure*. They refer to all pairwise similarities of a set of N vectors, for some similarity measure (e.g., Euclidean or cosine).
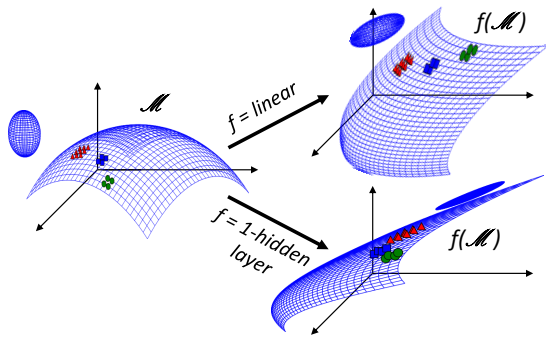
Figure 1: Effect of applying a mapping f to a (disconnected) manifold $\mathcal{M}$ with three hypothetical classes (■, ▲ and ●).

ignored phenomenon relevant to a wide range of cross-modal / cross-space applications such as retrieval, zero-shot learning or image annotation.

Ultimately, this paper aims at: (1) Encouraging the development of better architectures to bridge modalities / spaces; (2) Advocating for the use of semantic-based criteria to evaluate the quality of predicted vectors such as the neighborhood-based measure proposed here, instead of purely geometric measures such as mean squared error (MSE).

2 Related Work and Motivation

Neural network and linear mappings are popular tools to bridge modalities in *cross-modal retrieval* systems. Lazaridou et al. (2015b) leverage a text-to-image linear mapping to retrieve images given text queries. Weston et al. (2011) map label and image features into a shared space with a linear mapping to perform *image annotation*. Alternatively, Frome et al. (2013), Lazaridou et al. (2014) and Socher et al. (2013) perform *zero-shot* image classification with an image-to-text neural network mapping. Instead of mapping to latent features, Collell et al. (2018) use a 2-layer feedforward network to map word embeddings directly to image pixels in order to *visualize* spatial arrangements of objects. Neural networks are also popular in other cross-space applications such as *cross-lingual* tasks. Lazaridou et al. (2015a) learn a linear map from language A to language B and then translate new words by returning the nearest neighbor of the mapped vector in the B space.

In the context of zero-shot learning, shortcomings of cross-space neural mappings have also been identified. For instance, "hubness" (Radovanović et al., 2010) and "pollu-

tion" (Lazaridou et al., 2015a) relate to the high-dimensionality of the feature spaces and to overfitting respectively. Crucially, we do not assume that our cross-modal problem has any class labels, and we study the similarity between input and mapped vectors and between output and mapped vectors.

Recent work evidences that the predicted vectors of cross-modal neural net mappings are still largely informative about the input vectors. Lazaridou et al. (2015b) qualitatively observe that abstract textual concepts are grounded with the visual input modality. Counterintuitively, Collell et al. (2017) find that the vectors "imagined" from a language-to-vision neural map, outperform the original visual vectors in concept similarity tasks. The paper argued that the reconstructed visual vectors become grounded with language because the map preserves topological properties of the input. Here, we go one step further and show that the mapped vectors often resemble the input vectors more than the target vectors in semantic terms, which goes against the goal of a cross-modal map.

Well-known theoretical work shows that networks with as few as one hidden layer are able to approximate any function (Hornik et al., 1989). However, this result does not reveal much neither about test performance nor about the semantic structure of the mapped vectors. Instead, the phenomenon described is more closely tied to other properties of neural networks. In particular, continuity guarantees that topological properties of the input, such as connectedness, are preserved (Armstrong, 2013). Furthermore, continuity in a topology induced by a metric also ensures that points that are close together are mapped close together. As a toy example, Fig. 1 illustrates the distortion of a manifold after being mapped by a neural net.[2]

In a noiseless world with fully statistically dependent modalities, the vectors of one modality could be perfectly predicted from those of the other. However, in real-world problems this is unrealistic given the noise of the features and the fact that modalities encode complementary information (Collell and Moens, 2016). Such unpredictability combined with continuity and topology-preserving properties of neural nets propel the phenomenon identified, namely mapped vectors resembling more the input than the target vectors, in nearest neighbors terms.

---

[2]Parameters of these mappings were generated at random.

3 Proposed Approach

To bridge modalities $\mathcal{X}$ and $\mathcal{Y}$, we consider two popular cross-modal mappings $f : \mathcal{X} \to \mathcal{Y}$.

(i) **Linear** mapping (*lin*):

$$f(x) = W_0 x + b_0$$

with $W_0 \in \mathbb{R}^{d_y \times d_x}$, $b_0 \in \mathbb{R}^{d_y}$, where d_x and d_y are the input and output dimensions respectively.

(ii) Feed-forward **neural network** (*nn*):

$$f(x) = W_1 \sigma(W_0 x + b_0) + b_1$$

with $W_1 \in \mathbb{R}^{d_y \times d_h}$, $W_0 \in \mathbb{R}^{d_h \times d_x}$, $b_0 \in \mathbb{R}^{d_h}$, $b_1 \in \mathbb{R}^{d_y}$ where d_h is the number of hidden units and $\sigma()$ the non-linearity (e.g., tanh or sigmoid). Although single hidden layer networks are already universal approximators (Hornik et al., 1989), we explored whether deeper nets with **3 and 5 hidden layers** could improve the fit (see Supplement).

Loss: Our primary choice is the *MSE*: $\frac{1}{2}\|f(x) - y\|^2$, where y is the target vector. We also tested other losses such as the *cosine*: $1 - \cos(f(x), y)$ and the *max-margin*: $\max\{0, \gamma + \|f(x) - y\| - \|f(\tilde{x}) - y\|\}$, where $\tilde{x}$ belongs to a different class than (x, y), and γ is the margin. As in Lazaridou et al. (2015a) and Weston et al. (2011), we choose the first $\tilde{x}$ that violates the constraint. Notice that losses that do not require class labels such as *MSE* are suitable for a wider, more general set of tasks than discriminative losses (e.g., cross-entropy). In fact, cross-modal retrieval tasks often do not exhibit any class labels. Additionally, our research question concerns the cross-space mapping problem in isolation (independently of class labels).

Let us denote a set of N input and output vectors by $X \in \mathbb{R}^{N \times d_x}$ and $Y \in \mathbb{R}^{N \times d_y}$ respectively. Each input vector x_i is paired to the output vector y_i of the same index ($i = 1, \cdots, N$). Let us henceforth denote the mapped input vectors by $f(X) \in \mathbb{R}^{N \times d_y}$. In order to explore the similarity between $f(X)$ and X, and between $f(X)$ and Y, we propose two *ad hoc* settings below.

3.1 Neighborhood Structure of Mapped Vectors (Experiment 1)

To measure the similarity between the neighborhood structure of two sets of *paired* vectors V and Z, we propose the *mean nearest neighbor overlap* measure (*mNNO^K*(V, Z)). We define the *nearest neighbor overlap* **NNO^K**(v_i, z_i) as the *number of K nearest neighbors that two paired vectors v_i, z_i share in their respective spaces*. E.g., if the 3 (= K) nearest neighbors of v_{cat} in V are $\{v_{dog}, v_{tiger}, v_{lion}\}$ and those of z_{cat} in Z are $\{z_{mouse}, z_{tiger}, z_{lion}\}$, the **NNO^3**$(v_{cat}, z_{cat})$ is 2.

Definition 1 *Let $V = \{v_i\}_{i=1}^N$ and $Z = \{z_i\}_{i=1}^N$ be two sets of N paired vectors. We define:*

$$\textit{mNNO}^K(V, Z) = \frac{1}{KN} \sum_{i=1}^N NNO^K(v_i, z_i) \quad (1)$$

with **NNO^K**$(v_i, z_i) = |NN^K(v_i) \cap NN^K(z_i)|$, *where $NN^K(v_i)$ and $NN^K(z_i)$ are the indexes of the K nearest neighbors of v_i and z_i, respectively.*

The normalizing constant K simply scales $mNNO^K(V, Z)$ between 0 and 1, making it independent of the choice of K. Thus, a $mNNO^K(V, Z) = 0.7$ means that the vectors in V and Z share, on average, 70% of their nearest neighbors. Notice that *mNNO* implicitly performs retrieval for some similarity measure (e.g., Euclidean or cosine), and quantifies how semantically similar two sets of paired vectors are.

3.2 Mapping with Untrained Networks (Experiment 2)

To complement the setting above (Sect. 3.1), it is instructive to consider the limit case of an untrained network. Concept similarity tasks provide a suitable setting to study the semantic structure of distributed representations (Pennington et al., 2014). That is, semantically similar concepts should ideally be close together. In particular, our interest is in comparing X with its projection $f(X)$ through a mapping with random parameters, to understand the extent to which the mapping may disrupt or preserve the semantic structure of X.

4 Experimental Setup

4.1 Experiment 1

4.1.1 Datasets

To test the generality of our claims, we select a rich diversity of cross-modal tasks involving texts at three levels: *word* level (ImageNet), *sentence* level (IAPR TC-12), and *document* level (Wiki).
ImageNet (Russakovsky et al., 2015). Consists of ∼14M images, covering ∼22K WordNet synsets

(or meanings). Following Collell et al. (2017), we take the most relevant word for each synset and keep only synsets with more than 50 images. This yields 9,251 different words (or instances).

IAPR TC-12 (Grubinger et al., 2006). Contains 20K images (18K train / 2K test) annotated with 255 labels. Each image is accompanied with a short description of one to three sentences.

Wikipedia (Pereira et al., 2014). Has 2,866 samples (2,173 train / 693 test). Each sample is a section of a Wikipedia article paired with one image.

4.1.2 Hyperparameters and Implementation

See the Supplement (Sect. 1) for details.

4.1.3 Image and Text Features

To ensure that results are independent of the choice of image and text features, we use 5 (2 image + 3 text) features of varied dimensionality (64-d, 128-d, 300-d, 2,048-d) and two directions, text-to-image ($T \rightarrow I$) and image-to-text ($I \rightarrow T$). We make our extracted features publicly available.[3]

Text. In *ImageNet* we use 300-dimensional GloVe[4] (Pennington et al., 2014) and 300-d word2vec (Mikolov et al., 2013) word embeddings. In *IAPR TC-12* and *Wiki*, we employ state-of-the-art bidirectional gated recurrent unit (bi-GRU) features (Cho et al., 2014) that we learn with a classification task (see Sect. 2 of Supplement).

Image. For *ImageNet*, we use the publicly available[5] VGG-128 (Chatfield et al., 2014) and ResNet (He et al., 2015) visual features from Collell et al. (2017), where we obtained 128-dimensional VGG-128 and 2,048-d ResNet features from the last layer (before the softmax) of the forward pass of each image. The final representation for a word is the average feature vector (centroid) of all available images for this word. In *IAPR TC-12* and *Wiki*, features for individual images are obtained similarly from the last layer of a ResNet and a VGG-128 model.

4.2 Experiment 2

4.2.1 Datasets

We include six benchmarks, comprising three types of concept similarity: **(i) Semantic similarity**: *SemSim* (Silberer and Lapata, 2014), *Simlex999* (Hill et al., 2015) and *SimVerb-3500* (Gerz et al., 2016); **(ii) Relatedness**: *MEN* (Bruni et al.,

2014) and *WordSim-353* (Finkelstein et al., 2001); **(iii) Visual similarity**: *VisSim* (Silberer and Lapata, 2014) which includes the same word pairs as *SemSim*, rated for visual similarity instead of semantic. All six test sets contain human ratings of similarity for word pairs, e.g., ('cat','dog').

4.2.2 Hyperparameters and Implementation

The parameters in W_0, W_1 are drawn from a random uniform distribution $[-1, 1]$ and b_0, b_1 are set to zero. We use a tanh activation $\sigma()$.[6] The output dimension d_y is set to 2,048 for all embeddings.

4.2.3 Image and Text Features

Textual and visual features are the same as described in Sect. 4.1.3 for the *ImageNet* dataset.

4.2.4 Similarity Predictions

We compute the prediction of similarity between two vectors z_1, z_2 with both the cosine $\frac{z_1 z_2}{\|z_1\|\|z_2\|}$ and the Euclidean similarity $\frac{1}{1+\|z_1-z_2\|}$.[7]

4.2.5 Performance Metrics

As is common practice, we evaluate the predictions of similarity of the embeddings (Sect. 4.2.4) against the human similarity ratings with the *Spearman correlation* ρ. We report the average of 10 sets of randomly generated parameters.

5 Results and Discussion

We test statistical significance with a two-sided Wilcoxon rank sum test adjusted with Bonferroni. The null hypothesis is that a compared pair is equal. In Tab. 1, * indicates that $mNNO(X, f(X))$ differs from $mNNO(Y, f(X))$ ($p < 0.001$) on the same mapping, embedding and direction. In Tab. 2, * indicates that performance of mapped and input vectors differs ($p < 0.05$) in the 10 runs.

5.1 Experiment 1

Results below are with cosine neighbors and $K = 10$. Euclidean neighbors yield similar results and are thus left to the Supplement. Similarly, results in ImageNet with GloVe embeddings are shown below and word2vec results in the Supplement. The choice of $K = \{5, 10, 30\}$ had no visible effect on results. Results with *3- and 5-layer* nets did not show big differences with the results below (see Supplement). The *cosine* and *max-margin* losses

---

[3]http://liir.cs.kuleuven.be/software.html
[4]http://nlp.stanford.edu/projects/glove
[5]http://liir.cs.kuleuven.be/software.html

[6]We find that sigmoid and ReLu yield similar results.
[7]Notice that papers generally use only cosine similarity (Lazaridou et al., 2015b; Pennington et al., 2014).
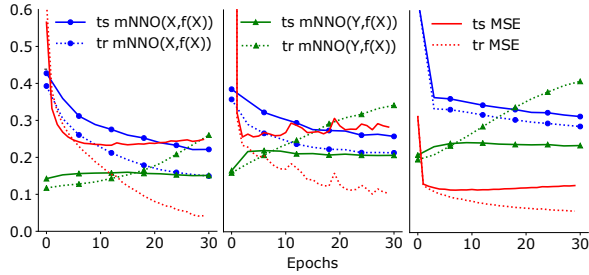
Figure 2: Learning a **nn** model in **Wiki** (left), **IAPR TC-12** (middle) and **ImageNet** (right).

performed slightly worse than *MSE* (see Supplement). Although Lazaridou et al. (2015a) and Weston et al. (2011) find that *max-margin* performs the best in their tasks, we do not find our result entirely surprising given that max-margin focuses on inter-class differences while we look also at intra-class neighbors (in fact, we do not require classes).

Tab. 1 shows our core finding, namely that the semantic structure of $f(X)$ resembles more that of X than that of Y, for both *lin* and *nn* maps.

			ResNet		VGG-128	
			$X, f(X)$	$Y, f(X)$	$X, f(X)$	$Y, f(X)$
ImageNet	$I \rightarrow T$	lin	**0.681***	0.262	**0.723***	0.236
		nn	**0.622***	0.273	**0.682***	0.246
	$T \rightarrow I$	lin	**0.379***	0.241	**0.339***	0.229
		nn	**0.354***	0.27	**0.326***	0.256
IAPR TC-12	$I \rightarrow T$	lin	**0.358***	0.214	**0.382***	0.163
		nn	**0.336***	0.219	**0.331***	0.18
	$T \rightarrow I$	lin	**0.48***	0.2	**0.419***	0.167
		nn	**0.413***	0.225	**0.372***	0.182
Wikipedia	$I \rightarrow T$	lin	**0.235***	0.156	**0.235***	0.143
		nn	**0.269***	0.161	**0.282***	0.148
	$T \rightarrow I$	lin	**0.574***	0.156	**0.6***	0.148
		nn	**0.521***	0.156	**0.511***	0.151

Table 1: Test mean nearest neighbor overlap. Boldface indicates the largest score at each $mNNO^{10}(X, f(X))$ and $mNNO^{10}(Y, f(X))$ pair, which are abbreviated by $X, f(X)$ and $Y, f(X)$.

Fig. 2 is particularly revealing. If we would only look at *train* performance (and allow train MSE to reach 0) then $f(X) = Y$ and clearly train $mNNO(f(X), Y) = 1$ while $mNNO(f(X), X)$ can only be smaller than 1. However, the interest is always on *test* samples, and (near-)perfect *test* prediction is unrealistic. Notice in fact in Fig. 2 that even if we look at *train* fit, MSE needs to be close to 0 for $mNNO(f(X), Y)$ to be

reasonably large. In all the combinations from Tab. 1, the *test mNNO*$(f(X), Y)$ never surpasses *test mNNO*$(f(X), X)$ for any number of epochs, even with an oracle (not shown).

5.2 Experiment 2

Tab. 2 shows that untrained linear (f_{lin}) and neural net (f_{nn}) mappings preserve the semantic structure of the input X, complementing thus the findings of Experiment 1. Experiment 1 concerns learning, while, by "ablating" the learning part and randomizing weights, Experiment 2 is revealing about the natural tendency of neural nets to preserve semantic information about the input, regardless of the choice of the target vectors and loss function.

	WS-353		Men		SemSim	
	Cos	Eucl	Cos	Eucl	Cos	Eucl
f_{nn}(GloVe)	**0.632**	**0.634***	0.795	**0.791***	0.75*	**0.744***
f_{lin}(GloVe)	0.63	0.606	0.798	0.781	0.763	0.712
GloVe	**0.632**	0.601	**0.801**	0.782	**0.768**	0.716
f_{nn}(ResNet)	0.402	0.408*	0.556	**0.554***	0.512	0.513
f_{lin}(ResNet)	**0.425**	0.449	0.566	0.534	0.533	0.514
ResNet	0.423	**0.457**	**0.567**	0.535	**0.534**	**0.516**

	VisSim		SimLex		SimVerb	
	Cos	Eucl	Cos	Eucl	Cos	Eucl
f_{nn}(GloVe)	0.594*	**0.59***	0.369	**0.363***	0.313	**0.301***
f_{lin}(GloVe)	0.602*	0.576	0.369	0.341	**0.326**	0.23
GloVe	**0.606**	0.58	**0.371**	0.34	0.32	0.235
f_{nn}(ResNet)	0.527*	**0.526***	0.405	**0.406**	0.178	0.169
f_{lin}(ResNet)	0.541	0.498	**0.409**	0.404	0.198	0.182
ResNet	**0.543**	0.501	**0.409**	0.403	**0.211**	**0.199**

Table 2: Spearman correlations between human ratings and the similarities (cosine or Euclidean) predicted from the embeddings. Boldface denotes best performance per input embedding type.

6 Conclusions

Overall, we uncovered a phenomenon neglected so far, namely that neural net cross-modal mappings can produce mapped vectors more akin to the input vectors than the target vectors, in terms of semantic structure. Such finding has been possible thanks to the proposed measure that explicitly quantifies similarity between the neighborhood structure of two sets of vectors. While other measures such as mean squared error can be misleading, our measure provides a more realistic estimate of the semantic similarity between predicted and target vectors. In fact, it is the semantic structure (or pairwise similarities) what ultimately matters in cross-modal applications.

Acknowledgments

References

Mark Anthony Armstrong. 2013. *Basic topology*. Springer Science & Business Media.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *JAIR* 49(1-47).

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Guillem Collell and Marie-Francine Moens. 2016. Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations. In *COLING*. ACL, pages 2807–2817.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *AAAI*. AAAI.

Guillem Collell, Teddy Zhang, and Marie-Francine Moens. 2017. Imagined Visual Representations as Multimodal Embeddings. In *AAAI*. AAAI, pages 4378–4384.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *WWW*. ACM, pages 406–414.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*. pages 2121–2129.

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* .

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*. volume 5, page 10.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* .

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4):665–695.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks* 2(5):359–366.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL*. pages 1403–1414.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015a. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*. volume 1, pages 270–280.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015b. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. pages 3111–3119.

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

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI* 36(3):521–535.

Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2017. Visually aligned word embeddings for improving zero-shot learning. *arXiv preprint arXiv:1707.05427* .

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. On the existence of obstinate results in vector space models. In *SIGIR*. ACM, pages 186–193.

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. *IJCV* 115(3):211–252.

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

---

[8]http://www.chistera.eu/projects/muster

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*. pages 935–943.

Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* .

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*. volume 11, pages 2764–2770.

Yang Zhang, Boqing Gong, and Mubarak Shah. 2016. Fast zero-shot image tagging. In *CVPR*. IEEE, pages 5985–5994.