Optimizing Visual Representations in Semantic Multi-modal Models with Dimensionality Reduction, Denoising and Contextual Information

Maximilian Köper(✉), Kim-Anh Nguyen, and Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5B, 70563 Stuttgart, Germany
{maximilian.koeper,kim-anh.nguyen,schulte}@ims.uni-stuttgart.de

Abstract. This paper improves visual representations for multi-modal semantic models, by (i) applying standard dimensionality reduction and denoising techniques, and by (ii) proposing a novel technique *ContextVision* that takes corpus-based textual information into account when enhancing visual embeddings. We explore our contribution in a visual and a multi-modal setup and evaluate on benchmark word similarity and relatedness tasks. Our findings show that NMF, denoising as well as *ContextVision* perform significantly better than the original vectors or SVD-modified vectors.

1 Introduction

Computational models across tasks potentially profit from combining corpus-based, textual information with perceptional information, because word meanings are grounded in the external environment and sensorimotor experience, so they cannot be learned only based on linguistic symbols, cf. the grounding problem (Harnad 1990). Accordingly, various approaches on determining semantic relatedness have been shown to improve by using multi-modal models that enrich textual linguistic representations with information from visual, auditory, or cognitive modalities (Feng and Lapata 2010, Silberer and Lapata 2012, Roller and im Walde 2013, Bruni et al. 2014, Kiela et al. 2014, Kiela and Clark 2015, Lazaridou et al. 2015).

While multi-modal models may be realized as either count or predict approaches, increasing attention is being devoted to the development, improvement and properties of low-dimensional continuous word representations (so-called *embeddings*), following the success of *word2vec* (Mikolov et al. 2013). Similarly, recent advances in computer vision and particularly in the field of deep learning have led to the development of better visual representations. Here, features are extracted from convolutional neural networks (CNNs) (LeCun et al. 1998), that were previously trained on object recognition tasks. For example,

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 292–300, 2018.
https://doi.org/10.1007/978-3-319-73706-5_25

Kiela and Bottou (2014) showed that CNN-based image representations perform superior in semantic relatedness prediction than other visual representations, such as an aggregation of SIFT features (Lowe 1999) into a bag of visual words (Sivic and Zisserman 2003).

Insight into the typically high-dimensional CNN-based representations is sparse, however. It is known that dimension reduction techniques, such as Singular Value Decomposition (SVD), improve performance on word similarity tasks when applied to word representations (Deerwester et al. 1990). In particular, Bullinaria and Levy (2012) observed highly significant improvements after applying SVD to standard corpus vectors. In addition, Nguyen et al. (2016) proposed a method to remove noisy information from word embeddings, resulting in superior performance on a variety of word similarity and relatedness benchmark tests.

In this paper, we provide an in-depth exploration of improving visual representations within a semantic model that predicts semantic similarity and relatedness, by applying dimensionality reduction and denoising. Furthermore, we introduce a novel approach that modifies visual representations in relation to corpus-based textual information. Following the methodology from Kiela et al. (2016), evaluations are carried out across three different CNN architectures, three different image sources and two different evaluation datasets. We assess the performance of the visual modality by itself, and we zoom into a multimodal setup where the visual representations are combined with textual representations. Our findings show that all methods but SVD improve the visual representations. This improvement is especially large on the word relatedness task.

2 Methods

In this section we introduce two dimensionality reduction techniques (Sect. 2.1), a denoising approach (Sect. 2.2) and our new approach *ContextVision* (Sect. 2.3).

2.1 Dimensionality Reduction

Singular Value Decomposition (SVD) (Golub and Van Loan 1996.) is a matrix algebra operation that can be used to reduce matrix dimensionality yielding a new high-dimensional space. SVD is a commonly used technique, also refered to as Latent Semantic Analysis (LSA) when applied to word similarity. ***Non-negative matrix factorization*** (NMF) (Lee and Seung 1999) is a matrix factorisation approach where the reduced matrix contains only non-negative real numbers (Lin 2007). NMF has a wide range of applications, including topic modeling, (soft) clustering and image feature representation (Lee and Seung 1999).

2.2 Denoising

Nguyen et al. (2016) proposed a denoising method **(DEN)** that uses a non-linear, parameterized, feed-forward neural network as a filter on word embeddings to

reduce noise. The method aims to strengthen salient context dimensions and to weaken unnecessary contexts. While Nguyen et al. (2016) increase the dimensionality, we apply the same technique to reduce dimensionality.

2.3 Context-Based Visual Representations

Our novel model *ContextVision* (CV) strengthens visual vector representations by taking into account corpus-based contextual information. Inspired by Lazaridou et al. (2015), our model jointly learns the *linguistic* and *visual* vector representations by combining two modalities (i.e., the linguistic modality and the visual modality). Differently to the multi-modal Skip-gram model by Lazaridou et al. (2015), we focus on improving the visual representation, while Lazaridou et al. aim to improve the linguistic representation, without performing updates on the visual representation, which are fixed in advance.

The linguistic modality uses contextual information and word negative contexts, and in the visual modality the visual vector representations are strengthened by taking the corresponding word vector representations, the contextual information, and the visual negative contexts into account.

We start out with describing the Skip-gram with negative sampling (SGNS) (Levy and Goldberg 2014) which is a variant of the Skip-gram model (Mikolov et al. 2013). Given a plain text corpus, SGNS aims to learn word vector representations in which words that appear in similar contexts are encoded by similar vector representations. Mathematically, SGNS model optimizes the following objective function:

$$\mathrm{J}_{SGNS} = \sum_{w \in V_W} \sum_{c \in V_C} \mathrm{J}_{ling}(w, c) \tag{1}$$

$$\begin{aligned} \mathrm{J}_{ling}(w, c) &= \#(w, c) \log \sigma(w, c) \\ &+ k_l \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-w, c_N)] \end{aligned} \tag{2}$$

where $\mathrm{J}_{ling}(w, c)$ is trained on a plain-text corpus of words $w \in V_W$ and their contexts $c \in V_C$, with V_W and V_C the word and context vocabularies, respectively. The collection of observed words and context pairs is denoted as D; the term $\#(w, c)$ refers to the number of times the pair (w, c) appeared in D; the term $\sigma(x)$ is the sigmoid function; the term k_l is the number of linguistic negative samples and the term c_N is the linguistic sampled context, drawn according to the empirical unigram distribution P. In our model, SGNS is applied to learn the linguistic modality.

In the visual modality, we improve the visual representations through contextual information; therefore the dimensionality of visual representations and linguistic representations needs to be equal in size. We rely on the denoising approach (Nguyen et al. 2016) to reduce the dimensionality of visual representations. The visual vector representations are then enforced by (i) directly increasing the similarity between the visual and the corresponding linguistic vector representations, and by (ii) encouraging the contextual information which co-occurs with

the linguistic information. More specifically, we formulate the objective function of the visual modality, $J_{vision}(v_w, c)$, as follows:

$$\begin{aligned} J_{vision}(v_w, c) = \#(v_w, c)(cos(w, v_w) \\ + \min\{0, \theta - cos(v_w, c) + cos(w, c)\}) \\ + k_v \cdot \mathbb{E}_{c_V \sim P_V}[\log \sigma(-v_w, c_V)] \end{aligned} \tag{3}$$

where $J_{vision}(v_w, c)$ is trained simultaneously with $J_{ling}(w, c)$ on the plain-text corpus of words w and their contexts c. v_w represents the visual information corresponding to the word w; and term θ is the margin; $cos(x, y)$ refers to the cosine similarity between x and y. The terms k_v, $\mathbb{E}_{c_V}$, and P_V are similarly defined as the linguistic modality. Note that if a word w is not associated with the corresponding visual information v_w, then $J_{vision}(v_w, c)$ is set to 0.

In the final step, the objective function which is used to improve the visual vector representations combines Eqs. 1, 2, and 3 by the objective function in Eq. 4:

$$J = \sum_{w \in V_W} \sum_{c \in V_C} (J_{ling}(w, c) + J_{vision}(v_w, c)) \tag{4}$$

3 Experiments

3.1 Experimental Settings

We use an English Wikipedia dump[1] from June 2016 as the corpus resource for training the *ContextVision*, containing approximately 1.9B tokens. We train our model with 300 dimensions, a window size of 5, 15 linguistic negative samples, 1 visual negative sample, and 0.025 as the learning rate. The threshold θ is set to 0.3. For the other methods dimensionality reduction is set to 300[2] dimensions. For the resources of image data, we rely on the publically available visual embeddings taken from Kiela et al. (2016)[3]. The data was obtained from three different image sources, namely Google, Bing, and Flickr. For each image source three state-of-the-art convolutional network architectures for image recognition were applied: AlexNet (Krizhevsky et al. 2012), GoogLeNet (Szegedy et al. 2015) and VGGNet (Simonyan and Zisserman 2014). In each source–CNN combination, the visual representation of a word is simply the centroid of the vectors of all images labeled with the word (mean aggregation). This centroid has 1024 dimensions for GoogLeNet and 4096 dimensions for the remaining two architectures. The size of the visual vocabulary for Google, Bing, and Flickr after computing the centroids is 1578, 1578, and 1582 respectively. For evaluation we relied on two human-annotated datasets, namely the 3000 pairs from MEN (Bruni et al. 2014) and the 999 pairs from SimLex (Hill et al. 2015). MEN focuses on relatedness, and SimLex focuses on similarity.

[1] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2.

[2] We conducted also experiments with 100 and 200 dimensions and obtained similar findings.

[3] http://www.cl.cam.ac.uk/~dk427/cnnexpts.html.

3.2 Visual Representation Setup

Table 1 shows the results for each of the previously introduced methods, as well as the unmodified image representation (Default). It can be seen that NMF, DEN and CV increase performance on all settings except for the combination Google & AlexNet. The performance of SVD is always remarkably similar to its original representations.

Furthermore we computed the average difference for each method across all settings, as shown in Table 2. The performance increased especially on the MEN relatedness task. Here NMF obtains on average a rho correlation of ≈.10 higher than its original representations. Also DEN and CV show a clear improvement, with the latter being most useful for the SimLex task.

To ensure significance we conducted *Steiger*'s test (Steiger 1980) of the difference between two correlations. We compared each ouf the methods against its Default performance.

Table 1. Comparing dimensionality reduction techniques, showing Spearman's ρ on SimLex-999 and MEN. * marks significance over the Default.

		AlexNet		GoogLeNet		VGGNet	
		SimLex	MEN	SimLex	MEN	SimLex	MEN
bing	Default	.324	.560	.314	.513	.312	.545
	SVD	.324	.557	.316	.513	.314	.544
	NMF	.329	**.610***	.341*	**.612***	.330	**.631***
	DEN	.356*	.582*	.342*	.564*	.343*	.599*
	CV	**.364***	.583*	**.358***	.582*	**.357***	.603*
flickr	Default	.271	.434	.244	.366	.262	.422
	SVD	.270	.424	.245	.364	.264	.418
	NMF	.284	.560*	.280*	.556*	.288	**.581***
	DEN	.276	.566*	.273*	.526*	.280	.570*
	CV	**.310***	**.573***	**.287***	**.589***	**.312***	.540*
google	Default	.354	.526	.358	.517	.346	.535
	SVD	**.355**	.527	.359	.518	.348	.536
	NMF	.353	**.596***	**.367**	**.608***	.366	**.609***
	DEN	.343	.559*	.361	.555*	.356	.560*
	CV	.352	.561*	.362	.573*	**.374**	.556*

Table 2. Average gain/loss in ρ across sources and architectures, in comparison to Default.

	SimLex	MEN	Both
SVD	0.11	−0.20	−0.05
NMF	1.71	10.49	6.10
DEN	1.63	7.34	4.48
CV	3.23	8.29	5.76

Out of the 19 settings, NMF obtained significant improvements with *=$p <$ 0.001 in 11 cases. Despite having a lower average gain (Table 2), DEN and CV obtained more significant improvements.

In total we observed most significant improvements on images taken from BING and with the CNN GOOGLENET.

3.3 Multi-modal Setup

In the previous section we explored the performance of the visual representations alone.

We now investigate their performance in a multi-modal setup, combining them with a textual representation. Using the same parameters as in Sect. 3.1 we created word representations relying on an SGNS model (Mikolov et al. 2013). We combined the representations by scoring level fusion (or late fusion). Following Bruni et al. (2014) and Kiela and Clark (2015) we investigate the impact of both modalities by varying a weight threshold (α). Similarity is computed as follows:

$$sim(x, y) = \alpha \cdot ling(x, y) + (1 - \alpha) \cdot vis(x, y) \quad (5)$$

Here $ling(x, y)$ is cosine similarity based on the textual representation only and $vis(x, y)$ for using the visual space.

For the following experiment we focus on ALEXNET, varying the image resource between BING for the SIMLEX task and FLICKR for the MEN task. The results are shown in Fig. 1a for SIMLEX, and in Fig. 1b for MEN.

It can be seen that all representations obtain superior performance on the text-only representation (black dashed line, SIMLEX ρ = .384, MEN ρ = .741). The highest correlation can be obtained using the DEN or VC representations for SIMLEX. Interestingly these two methods obtain best performance when given equal weight to both modalities ($\alpha = 0.5$) while the remaining methods as well as

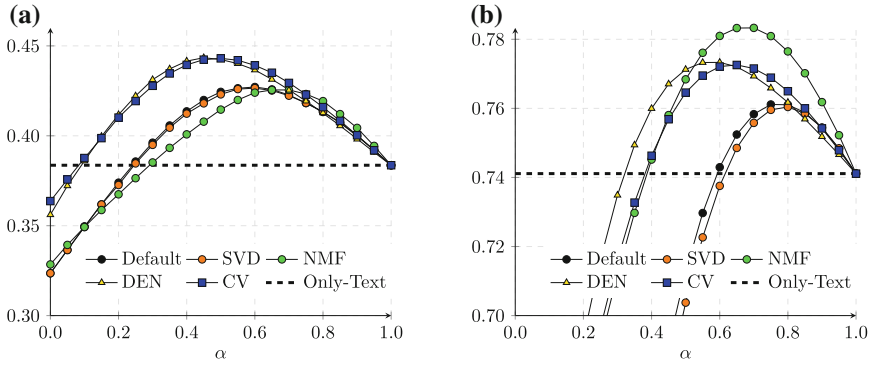

Fig. 1. (a) Comparing multi-modal results on SimLex-999. Image representation from BING using ALEXNET. Y-Axis shows Spearman's ρ. X-axis changes impact of each modality, from only image to the far left to only textual representation. (b) Multi-modal results on MEN. Image representation from FLICKR using ALEXNET.

the unmodified default representations obtain a peak in performance when given more weight to the textual representation. A similar picture emerges regarding the results on MEN, where also NMF obtains superior results (.748).

4 Conclusion

We successfully applied dimensionality reduction as well as denoising techniques, plus a newly proposed method *ContextVision* to enhance visual representations within semantic vector space models. Except for SVD, all investigated methods showed significant improvements in single - and multi-modal setups on the task of predicting similarity and relatedness.

Acknowledgments. The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper, Kim-Anh Nguyen), the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde) and the Ministry of Education and Training of the Socialist Republic of Vietnam (Scholarship 977/QD-BGDDT; Kim-Anh Nguyen). We would like to thank the four anonymous reviewers for their comments and suggestions.

References

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

Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behav. Res. Methods **44**, 890–907 (2012)

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**, 391–407 (1990)

Feng, Y., Lapata, M.: Visual information in semantic representation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 91–99 (2010)

Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore (1996)

Harnad, S.: The symbol grounding problem. Physica D **42**, 335–346 (1990)

Hill, F., Reichart, R., Korhonen, A.: Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Comput. Linguist. **41**(4), 665–695 (2015)

Kiela, D., Bottou, L.: 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, Doha, Qatar, pp. 36–45 (2014)

Kiela, D., Clark, S.: Multi - and cross-modal semantics beyond vision: grounding in auditory perception. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 2461–2470. Association for Computational Linguistics (2015)

Kiela, D., Hill, F., Korhonen, A., Clark, S.: Improving multi-modal representations using image dispersion: why less is sometimes more. In: Proceedings of the 52nd Annual Meeting of the Association for Computational, Baltimore, USA, pp. 835–841 (2014)

Kiela, D., Verő, A.L., Clark, S.: Comparing data sources and architectures for deep visual representation learning in semantics. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 447–456 (2016)

Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates Inc., (2012)

Lazaridou, A., Pham, N.T., Baroni, M.: Combining language and vision with a multimodal skip-gram model. In: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, USA, pp. 153–163 (2015)

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, pp. 2278–2324 (1998)

Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature **401**, 788–791 (1999)

Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Proceedings of the 27th International Conference on Advances in Neural Information Processing Systems, Montréal, Canada, pp. 2177–2185 (2014)

Lin, C.-J.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**, 2756–2779 (2007)

Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1150–1157, Washington, DC, USA (1999)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **26**, 3111–3119 (2013)

Nguyen, K.A., im Walde, S.S., Vu, N.T.: Neural-based noise filtering from word embeddings. In: Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, pp. 2699–2707 (2016)

Roller, S., im Walde, S.S.: A multimodal LDA model integrating textual, cognitive and visual modalities. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, pp. 1146–1157 (2013)

Silberer, C., Lapata, M.: Grounded models of semantic representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Jeju Island, Korea, pp. 1423–1433 (2012)

Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv.org, abs/1409.1556 (2014)

Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision, Nice, France, pp. 1470–1477 (2003)

Steiger, J.H.: Tests for comparing elements of a correlation matrix. Psychol. Bull. **87**, 245–251 (1980)
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (2015)