

Visual Grounding of Inter-lingual Word-Embeddings

Wafaa Mohammed Hassan Shahmohammadi Hendrik P. A. Lensch R. Harald Baayen

University of Tübingen
wmohammed@aimsammi.org,
{hassan.shahmohammadi, hendrik.lensch, harald.baayen}@uni-tuebingen.de

Abstract

Visual grounding of Language aims at enriching textual representations of language with multiple sources of visual knowledge such as images and videos. Although visual grounding is an area of intense research, inter-lingual aspects of visual grounding have not received much attention. The present study investigates the inter-lingual visual grounding of word embeddings. We propose an implicit alignment technique between the two spaces of vision and language in which inter-lingual textual information interact in order to enrich pre-trained textual word embeddings. We focus on three languages in our experiments, namely, English, Arabic, and German. We obtained visually grounded vector representations for these languages and studied whether visual grounding on one or multiple languages improved the performance of embeddings on word similarity and categorization benchmarks. Our experiments suggest that inter-lingual knowledge improves the performance of grounded embeddings in similar languages such as German and English. However, inter-lingual grounding of German or English with Arabic led to a slight degradation in performance on word similarity benchmarks. On the other hand, we observed an opposite trend on categorization benchmarks where Arabic had the most improvement on English. In the discussion section, several reasons for those findings are laid out. We hope that our experiments provide a baseline for further research on inter-lingual visual grounding.

1 Introduction

Distributional Semantic Models (DSMs) have long been used to capture words' meaning. They estimate semantic representations from co-occurrences of words in text corpora. Even though embeddings are the dominant method for large scale data, from a psychological and cognitive point of view, distributional models suffer from the problem referred to as the *Symbol Grounding Problem* (Harnad, 1990):

the meaning of a symbol (word) is entirely accounted for in terms of other symbols without any links to the outside world. In the context of natural language processing (NLP), grounding is defined as “the process of linking the symbolic representation of language (e.g., words) into the rich perceptual knowledge of the outside world” (Shahmohammadi et al., 2021). Moreover, (Huang et al., 2021) have proved that multi-modal learning outperforms uni-modal learning as it has access to a better quality latent space representation.

Many studies have addressed grounding of language in vision, typically focusing on grounding for English (Bruni et al., 2014; Shahmohammadi et al., 2022). As a consequence, inter-lingual visual grounding is still poorly understood. This study investigates whether monolingual textual embeddings benefit from the knowledge of other languages in the process of visual grounding. We extend a state-of-the-art model for monolingual visual grounding (Shahmohammadi et al., 2022) by considering different combinations of three languages, namely, English, German, and Arabic. Using various word categorization benchmarks, our experiments show that the three languages profitably exchange inter-lingual knowledge across a simple linear vector space. To the best of our knowledge, we are the first to investigate the problem of visual grounding of inter-lingual word embeddings. Overall, our contributions are as follows:

a) We propose a simple extension of a state-of-the-art visual grounding model to integrate three different languages. b) We obtain zero-shot visually grounded embeddings in three languages. c) Using various benchmarks, we reveal how visual grounding changes textual vector space across languages and show that inter-lingual knowledge transfers to downstream tasks.

Our paper is structured as follows: Section 2 briefly highlights the related works. Section 3 introduces our problem of interest. In Section 4 our pro-

posed model is elaborated. Implementation details are covered in Section 5. The results are presented in Section 6, with further discussion in section 7. In Section 8, we conclude our research, and finally, we point out the limitations and future directions of our work.

2 Related Work

There have been many studies on language grounding in vision most of which focus on monolingual visual grounding. There have been also other works on cross-modal and cross-lingual representations tailored for specific downstream applications.

Monolingual grounding: The study of [Bruni et al. \(2014\)](#) was one of the first studies to obtain visually grounded embeddings by simple fusion such as applying SVD on the concatenation of word and image vectors. [Kiros et al., 2018](#) adopted a similar fusion approach using gating mechanisms. [Silberer and Lapata, 2014](#) and [Hasegawa et al., 2017](#) encoded the two modalities as vectors of attributes and combine them using autoencoders. [Kurach et al., 2017](#) and [Shahmohammadi et al., 2022](#) adopted a simple approach where textual embeddings are directly optimized to match image representations. They propose a grounding framework that depends on the alignment of textual and visual features.

Cross-modal cross-lingual representations: In the multilingual setting, the focus has largely been on cross-modal downstream tasks. [Burns et al., 2020](#) proposed a scalable multilingual aligned language representation using masked cross-language modelling objective. [Ni et al., 2021](#) proposed a multilingual multimodal model that combines different languages and different modalities into a shared space via multitask pre-training. Similarly, [Zhou et al., 2021](#) introduced a machine translation augmented model for cross-modal cross-lingual learning by introducing multi-modal losses. [Mohammadshahi et al., 2019](#) trained a multilingual multimodal model by optimizing the alignment between languages for image-description retrieval task.

The present study is inspired by both directions explored in the literature on visual grounding and multi-lingual representations. We propose a straight forward alignment technique informing textual representations about the visual space while also making use of inter-lingual features. We generate visually grounded inter-lingual word embed-

dings and evaluate their performance on similarity and categorization benchmarks.

A new direction of research that has been published in parallel with this paper is the work of [\(Chen et al., 2022\)](#). Their model, PaLI (Pathways Language and Image model), employs scaling of joint vision and language pre-training. They make use of the largest transformers to date to train the model. They were able to achieve state-of-the-art in multiple vision and language tasks such as captioning, visual question answering, and scene-text understanding.

3 Inter-lingual Visual Grounding

Multilingual-language models hold great promise for the development of embeddings for under-resourced languages ([Armengol-Estapé et al., 2021](#)). The central idea in this line of research is that different languages bring different perspectives (e.g., cultural information and grammar) which can inform each other, resulting in a richer model that has a better understanding of words’ meanings in any specific language. Moreover, since typical visual scenes are thought to produce similar information across different languages, integrating visual knowledge (e.g., images) into a multilingual model can contribute to obtaining a better quality grounded embedding space.

4 Model Architecture

Our model maps a textual description of an image into its corresponding image representation. It makes use of a linear alignment to preserve most of the textual knowledge in the word embeddings, allowing only subtle modifications by the error received from the image. It is trained using multilingual image captioning data. The model is given the task to match, for a given image, the multilingual captions to that image in such a way that language-specific features are preserved, and not overwhelmed by inter-lingual features, and image features.

Our model maps two (or three) languages to the grounded space using a shared linear alignment. For instance, [figure 1](#) introduces the model for the combination of English and Arabic languages. Let D be the dataset consisting of triple samples of $(I, S_{en}, S_{ar}) \in D$. Here I refers to an image, S_{en} and S_{ar} denote matching captions of I in English and Arabic respectively. As shown in [Figure 1](#), the two captions are passed through

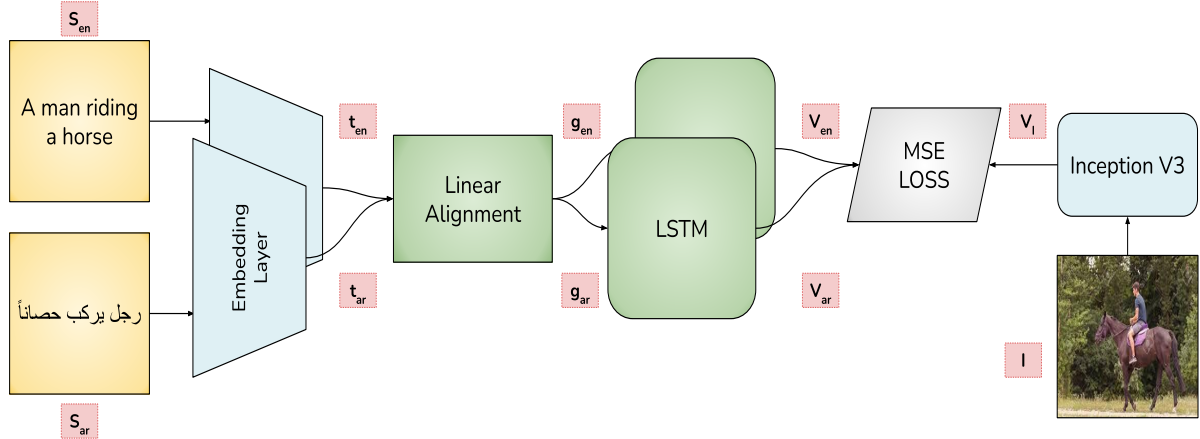


Figure 1: Model Architecture. sentences are first tokenized. Individual tokens are passed, one by one, to a pre-trained embedding layer, followed by a linear alignment that transfers the embeddings into the grounded space. Grounded vectors are encoded into a single vector by an LSTM encoder. The output of the LSTM is then optimized against the image vector generated via a pre-trained CNN model. Layers in blue are frozen during training.

a pre-trained embedding layer (GloVe) (Pennington et al., 2014) to obtain their textual representations t_{en}, t_{ar} which are then mapped to a visually grounded space through a linear transformation. We refer to this linear transformation as the alignment layer. The alignment layer is used to extract grounded embeddings after training. During training, grounded word vectors of each caption are encoded as a single vector using an LSTM layer as follows:

$$\begin{aligned} V_{en} &= LSTM_{en}(g_{en}, c_0, h_0 | \theta), \\ V_{ar} &= LSTM_{ar}(g_{ar}, c_0, h_0 | \theta) \end{aligned}$$

where, g_{en}, g_{ar} denote the grounded word vectors of the English and Arabic captions respectively. c_0, h_0 and θ represent the initial cell state, initial hidden state, and the trainable parameters of the LSTM. The parameters of the linear alignment and the LSTM layer are optimized to match the sentence representations in both languages to the same image vector V_I as follows:

$$\begin{aligned} \mathcal{L}_{en}(\theta_{en}) &= \frac{1}{|D|} \sum_{t=1}^n (V_I^t - V_{en}^t)^2, \\ \mathcal{L}_{ar}(\theta_{ar}) &= \frac{1}{|D|} \sum_{t=1}^n (V_I^t - V_{ar}^t)^2, \end{aligned}$$

where θ_{en} and θ_{ar} indicate the learning parameters for each language. The image vector V_I^* is generated using a pre-trained CNN model. The overall loss is simply the sum of the two losses: $\mathcal{L}_{all}(\Theta) = \mathcal{L}_{en}(\theta_{en}) + \mathcal{L}_{ar}(\theta_{ar})$

In this equation, Θ represents all the network’s learning parameters. After training, we generate grounded word embedding using the alignment layer. A given textual word embedding $w_t \in \mathbb{R}^d$ is passed through the trained alignment, after which its grounded version is extracted from the align-

ment layer: $g_t \in \mathbb{R}^c$ as $g_t = w_t.M$, where M denotes the trained alignment layer.

5 Implementation details

We used the Microsoft COCO 2017 dataset (Lin et al., 2014) for our experiments. This dataset consists of 123,287 images with 5 captions each. It is split into 118k training images and 5k validation images. We experimented with three languages for the captions: English, Arabic, and German. The original dataset provided by Microsoft contains the English captions. We obtained the German captions from (Biswas et al., 2021), who translated the English COCO captions using the Fairseq neural machine translator, and the Arabic captions from (Hashim, 2020), who generated the captions using Google’s advanced cloud translation API. For the Arabic version of COCO, we only had available to us translations of the captions for 82k samples, which we split into 77k samples for training and 5k samples for validation, and this is the set of images that we use for models that included Arabic. For fair comparisons, we also investigated model performance for English and German using the same 82k images. For all the experiments, we used TensorFlow as a development framework. The training environment is similar to the one used by Shahmohammadi et al. (2022). We used a batch size of 256 image-caption pairs. We trained for 20 epochs with 5 epochs as early stopping tolerance, using the NAdam optimizer (Dozat, 2016) with a learning rate of 0.001. The image vectors were obtained using pre-trained vectors from Inception-V3

(Szegedy et al., 2016), which are based on ImageNet (Deng et al., 2009). For pre-trained textual embeddings we used GloVe embeddings (Pennington et al., 2014). The vocabulary considered for training English comprised the 10k most frequent words. For German and Arabic, which have much richer inflectional systems compared to English, we took into account the 30k most frequent words. We set the dimension of grounded word embeddings to 1024 ($g_t \in \mathbb{R}^{1024}$), and matched the size of the LSTM’s output to that of the image vectors (both to 2048). Both the embedding layer and the pre-trained CNN were frozen during training.

6 Results

In this section, we explain our evaluation criteria and report the results of our experiments. We use various word similarity/relatedness and word categorization benchmarks and provide both quantitative and qualitative results.

6.1 Qualitative Evaluation

Figure 2 shows the difference between the nearest neighbours of words from the three languages in the textual and grounded spaces (using the grounding setup with separate grounding of each individual language). The representations in the grounded space are semantically much more precise, and are much less dependent on simple co-occurrence statistics. Our algorithm for visual grounding thus contributes to taking a step forward in solving the symbol grounding problem. For example, the word *car* in Arabic has its nearest neighbours as *airplane* and *explosion* in the textual space, while in the grounded space, the neighbours are different declensions of the word *car*.

6.2 Word Similarity/ Relatedness Evaluation

Following (Bruni et al., 2014; Shahmohammadi et al., 2022), we evaluated our visually grounded word embeddings using similarity/relatedness benchmarks. The task is to estimate the similarity/relatedness score of a pair of words using the Spearman correlation as evaluation metric. Relatedness is a measure of the extent to which two words are associated with each other, e.g. (pen, paper). Similarity quantifies how alike two concepts are based on their location within an is-a hierarchy (e.g., car, automobile). Some benchmarks differentiate between the two while others consider them similar when scoring pairs of words.

Tables 1, 2, 3 summarize the results of visually grounded embeddings on similarity/relatedness benchmarks for English, German, and Arabic. For English, we experimented with six similarity/relatedness benchmarks: WordSim353 (Finkelstein et al., 2001), MEN (Bruni et al., 2014), RW (Luong et al., 2013), MTurk (Radinsky et al., 2011), simVerb (Gerz et al., 2016), and SimLex999 (Hill et al., 2015). For German, evaluations are based on the Multilingual versions of WrdSim353 and simLex999 (Leviant and Reichart, 2015). For Arabic, similarity was evaluated using four benchmarks: Almarsoomi (Almarsoomi et al., 2013), MC30 (Hassan and Mihalcea, 2009), Saif40 (Saif et al., 2014), and WordSim (Hassan and Mihalcea, 2009).

Across the three languages, visual grounding yields embeddings that perform substantially better than embeddings that are based on text only. It is noteworthy that the grounded embeddings achieved superior results on all the similarity benchmarks, for all three languages.

For both English and German, adding German and English respectively as a second language to the model leads to a further improvement in performance on the benchmark tasks. Adding Arabic as a second language along with English or German, however, led to a reduction in accuracy. The experiments evaluating Arabic word embeddings revealed that fusing in English or German did not improve performance on the Arabic benchmarks. Furthermore, experiments implementing visual grounding for three languages jointly did not provide further accuracy.

The same findings can also be observed even when varying the size of the training and validation data. For example, for the same set of 82k images, adding German embeddings to English embeddings led to an improvement on benchmark tasks, whereas adding Arabic embeddings did not. In the discussion section, we provide a detailed discussion of why Arabic embeddings do not provide further precision for English or German grounded embeddings.

6.3 Word Categorization Evaluation

We also evaluated our embeddings on six categorization benchmarks: Battig (Battig and Montague, 1969), AP (Almuhareb and Poesio, 2005), BLESS (Baroni and Lenci, 2011), and three tasks published at (ESSLLI, 2009), (ESSLLI-a, 2009), which focuses on grouping concrete nouns into semantic

	WSim	MEN	RW	MTurk	SimVerb	SimLex	Mean
Textual	73.8	80.5	45.5	71.5	28.3	40.8	56.7
Grounded EN	77.7	84.8	51.9	73.3	38.02	52.2	62.9
Grounded EN (82k)	76.03	84.5	50.3	72.7	34.9	48.6	61.2
Grounded EN + DE	79.2	84.8	52.3	74.1	36.6	51.03	63
Grounded EN + DE (82k)	75.3	84.3	50.8	74.2	34.5	49.1	61.4
Grounded EN + AR	76.9	84.7	50.3	73.1	34.3	48.3	61.3
Grounded EN + DE + AR	76.7	84.3	51.1	73.9	33.3	48.04	61.2

Table 1: Performance of textual and grounded English embeddings on similarity/relatedness benchmarks. Results include different combinations of the three languages, English (EN), German (DE), and Arabic (AR). Inter-lingual grounding in English and German outperforms both the textual and monolingual grounded embeddings.

	WSim	SimLex	Mean
Textual	46.6	30.9	38.8
Grounded DE	56.2	36.9	46.6
Grounded DE (82k)	56.3	35.8	46.1
Grounded DE + EN	57.02	37.2	47.1
Grounded DE + AR	55.5	33.2	44.3
Grounded DE + EN (82k)	56.6	35.1	45.9
Grounded DE + EN + AR	54.1	33.2	43.7

Table 2: Performance of textual and grounded German embeddings on similarity/relatedness benchmarks. Results include different combinations of German embeddings with two other languages: English (EN), and Arabic (AR). Grounding in both German and English outperforms all other monolingual groundings.

	WSim	Almarsoomi	MC30	Saif40	Mean
Textual	30.7	65.9	49.9	71.8	54.6
Grounded AR	41.9	72.8	59.2	80.6	63.6
Grounded AR + EN	39.7	72.8	56.9	83.2	63.2
Grounded AR + DE	36.9	75.2	52.6	77.05	60.4
Grounded AR + EN + DE	39.6	73.9	56.2	75.5	61.3

Table 3: Performance of textual and grounded Arabic embeddings on similarity/relatedness benchmarks. Results include different combinations of Arabic embeddings with two other languages: English (EN), and German (DE).

	Battig	AP	BLESS	ESLLI-a	ESLLI-b	ESLLI-c	Mean
Textual	45.4	60.4	87.5	75.0	75.0	62.2	67.6
Grounded EN	47.03	60.7	80.5	75.0	75.0	64.4	67.1
Grounded EN + DE	48.6	62.4	87	84.1	77.5	60.0	69.9
Grounded EN + FA	47.1	64.4	85.5	81.8	80.0	64.4	70.5
Grounded EN + AR	49.8	64.9	79.5	84.1	75.0	64.4	69.6
Grounded EN + DE + AR	47.5	64.7	85.5	75.0	75.0	62.2	68.3
Grounded EN + DE (82k)	47.1	65.9	81.5	84.1	77.5	55.6	68.6

Table 4: Performance of textual and grounded English embeddings on Categorization benchmarks. Results include different combinations of the three languages, English (EN), German (DE), Arabic (AR), and Persian (FA).

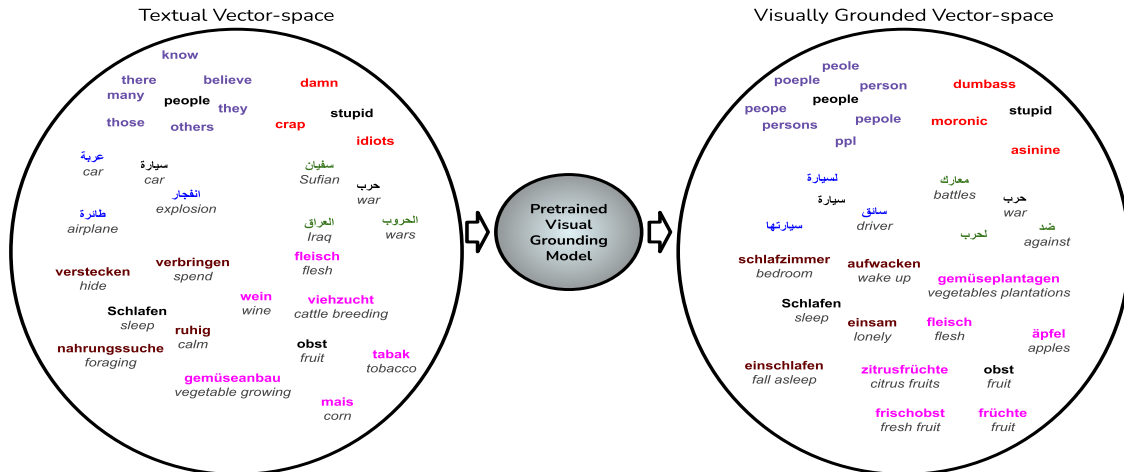


Figure 2: Comparisons of the textual and grounded vector spaces for English, German, and Arabic. For each query word (in black), out of the 10 nearest neighbours, the neighbours unique to each space are displayed. Visual grounding better captures a word’s meaning and reduces the dependency on just co-occurrence statistics.

categories; (ESLLI-b, 2009), which tests computational models for their ability to discriminate between abstract and concrete nouns; and (ESLLI-c, 2009), which groups verbs into semantic categories.

The concept-categorization task requires clustering a set of nouns expressing basic-level concepts into gold standard categories. To evaluate on this task, clustering is performed using a k-means clustering algorithm (Likas et al., 2003). Performance is evaluated using a purity score between the truth and predicted cluster labels. Results are presented in Table 4. Monolingual grounding did not result in improvements on this benchmark; grounded English embeddings revealed worse performance on BLESS compared to the textual embeddings. However, adding a second language solved this problem. Incorporation of both German and Arabic embeddings resulted in improved performance of the English embeddings on all benchmarks. However, combining the three languages did not give rise to further improvements. Interestingly, for the smaller dataset size (82k images), Arabic had a better performance than German, a result that contrasts with those obtained for the similarity benchmarks.

More Languages: We further extended our experiments by using the Persian language. For this aim, we translated the COCO captions using google translate API¹ and made use of a pre-trained GloVe word embeddings model² train on OSCAR (Abadji

et al., 2022). Similar to other languages grounding textual Persian embeddings significantly boosted the result (Spearman’s correlation) by more than 10% (from 36.7 to 47) on the SemEval2017 benchmark (Camacho-Collados et al., 2017). Due to time constraints, we only trained the grounded embeddings from English + Persian and evaluated them on the word categorization benchmarks. As shown in Table 4, Adding Persian (denoted as FA) results in the best mean performance.

To further analyze the interaction of visual grounding with multiple languages, we made use of the BLESS (Baroni and Lenci, 2011) dataset. This dataset consists of tuples of the format (*concept-relation-relatum*). For example, *lizard-attri-striped*: the concept *lizard* is linked to the relatum *striped* via the *attribute* relation. BLESS focuses on a set of basic concrete nouns and explicit semantic relations. Additionally, it contains a number of random relatum words that are not semantically related to any of the concepts. The tasks that come with this dataset it to detect which words are related to a given concept, as well as determining the type of relation involved. The dataset comprises 200 concepts grouped into 17 classes.

BLESS includes 5 types of relations, in-addition to the random relations: **COORD**: the relatum is a noun that is a co-hyponym (coordinate) of the concept: *dishwasher-coord-oven*. **HYPER**: the relatum is a noun that is a hypernym of the concept: *dishwasher-hyper-appliance*. **MERO** : the relatum is a noun referring to a

¹<https://libraries.io/pypi/googletrans>

²<https://github.com/taesiri/PersianWordVectors>

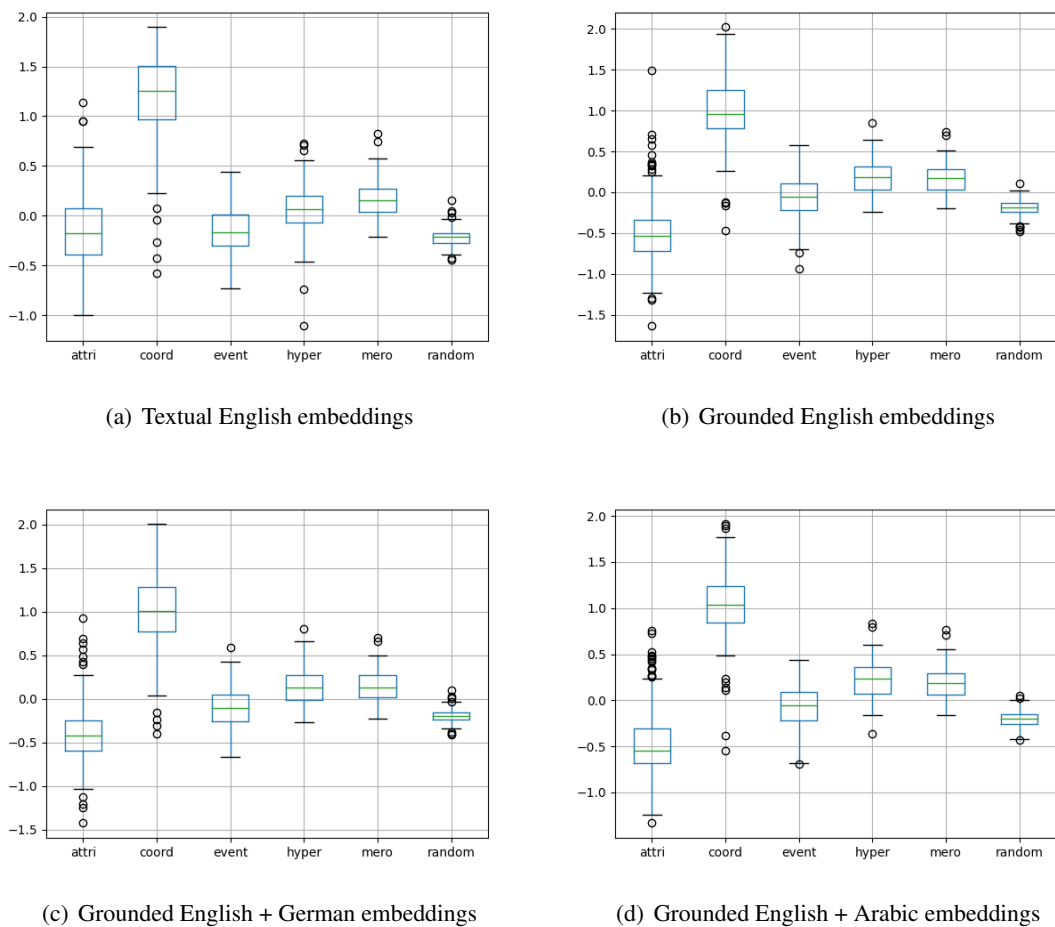


Figure 3: BLESS (Baroni and Lenci, 2011) Analyses of textual and grounded English embeddings with the combination of other languages. Visual grounding clearly reduces the variance on *attri* and *coord* categories resulting in more refined clusters and higher word categorization scores.

part/component/organ/member of the concept, or something that the concept contains or is made of: *dishwasher-mero-button*. **ATTRI** : the relatum is an adjective expressing an attribute of the concept: *dishwasher-attri-full*. **EVENT** : the relatum is a verb referring to an action/activity/happening/event the concept is involved in or is performed by/with the concept: *dishwasher-event-use*.

Using our embeddings, we calculated the mean cosine similarity score of each concept to all its relata across all relations. For each of the 200 BLESS concepts, we obtain six cosine similarity scores, one per relation: $C_{ir} = \frac{1}{n} \sum_{j=1}^n \cos(C_i, Rel_{rj})$ where C_{ir} denotes the mean cosine score of concept i for relation r and n indicates the number of words per relation. The scores are then normalized across each concept as: $C_{ir} = \frac{C_{ir} - \mu_i}{\sigma_i}$, where μ_i and σ_i denote the mean and the standard deviation of the scores of C_i across all relations.

Figure 3 presents the distribution of scores per relation across the 200 concepts. While the coarse structures of all the embeddings are relatively similar with respect to the scores (cosine similarity) across relations, the figures reveal interesting properties. For instance, the distributions in both *attri* and *coord* are more compact when visual grounding is applied. That is, the model is more certain about the similarity between the words and hence creates a more refined cluster of words. Another interesting point is the increased mean in the *hyper* category, especially for Arabic, in line with the results reported in Table 4.

Moreover, visual grounding lowers the mean score on *coord* category across all languages; this is probably because of the visually different word pairs in *coord* category. For example, (*turtle, alligator*) and (*toaster, stove*) are not visually similar. Therefore, their word vectors diverge as the

result of grounding. These findings are in line with previous findings that visual grounding prioritizes similarity over relatedness (Shahmohammadi et al., 2021). Surprising at first sight is that the mean score of *attri* category is lower in all grounding setups. This, however, may be due to the rather different sets of attributes in BLESS and in our image captions. Many of the attributes used in BLESS rarely occur in image captions, examples are *antarctic*, *amphibious*, *aquatic*, and *noisy*.

In order to statistically validate these findings, we applied a Gaussian Location-Scale Generalized Additive Mixed Model (GAMM) (Wood, 2017), with word as random-effect factor, and main effects for *embedding type* and *relation* for both mean and variance. This analysis revealed that the grounded English embeddings (monolingual grounding) had the highest mean score, followed by the grounded English embeddings generated by integrating English and German, followed closely by the English + Arabic embeddings. Interestingly, compared to the textual embeddings, the variance for grounded embeddings is reduced, and even more reduced for inter-lingual grounded embeddings with Arabic and German. Thus, there seems to be a trade-off between mean and variance. While monolingual grounding had the highest mean score, inter-lingual grounding helped more in reducing the variance, resulting in more refined clusters of semantically related words.

Comparing the mean of scores with respect to the different relations, with the *random* relation as the baseline, we noticed that the mean decreases for *attri*, but increases for all other relations, and noticeably so for the *hyper* and *mero* relations. The variance, on the other hand, increases for all relations and to the greatest extent for *attr* and *coord*. These statistical results dovetail well with our previously mentioned conclusions about visually different word pairs in *coord* category and the difference in *attributes* between the BLESS data and our image captions. Overall, the boxplots indicate that inter-lingual visual grounding creates more refined clusters of word vectors in the vector space based on visual clues in the training sets.

7 Discussion

We proposed an inter-lingual visual grounding model on textual word embeddings. Our model thus far supports the benefit of visual grounding and inter-lingual visual grounding on various word

similarity and word categorization benchmarks. Some of the results in Section 6 however are hard to interpret. In this section, we will discuss possible explanations for the model’s behavior on different tasks across different languages.

On the word similarity benchmarks (Tables 1, 2, and 3) we observe that German and English seem to interact more efficiently than Arabic with either. We believe the slight degradation in performance when adding Arabic might be due to the fact that the Arabic language structure is quite different: much more information is packed into its verbs, and pronouns are used differently and more sparingly. Moreover, its orthography leaves out a lot of phonological information (hardly any vowels), so word embeddings are much more ambiguous relative to English or German. Therefore, the semantic spaces that are constructed are much less similar to that in the two other languages. Apart from the evident differences between Arabic and the other two languages, it is worth mentioning that adding Arabic is far from detrimental. That is, the resulting embeddings (Arabic added) still outperform the textual embeddings significantly. This implies that there exists a linearly aligned common core between the three languages (vector spaces) which as observed in section 6.3, yielded the lowest variance and more pure vector space. Table 4 further supports these findings. Interestingly, the monolingual grounding of English does not seem to improve the categorization performance, inter-lingual knowledge, on the other hand, results in obvious improvements with respect to the mean score. The opposing impact of adding Arabic on the similarity/relatedness results in contrast to the categorization results indicates the need for further investigation on the evaluation criteria of inter-lingual embeddings.

Furthermore, it is not clear why monolingual visual grounding is more beneficial for word similarity compared to word categorization. We think cultural biases might play a role. For example, our training set (the COCO image dataset) is likely culture-specific, with a strong bias toward the US culture, and our benchmarks are compiled with various purposes across different languages. We, therefore, believe that current evaluation benchmarks only shine light on some facets of the complex interplay of different languages in visual grounding, and further investigation is required for more coherent interpretations.

8 Conclusions

The main purpose of this study is to shed light on the problem of inter-lingual visual grounding. We stated the importance of grounding in language understanding and the cognitive plausibility of text representations. We also suggested a baseline architecture for inter-lingual visual grounding and analyzed the performance of the resulting embeddings on word similarity and categorization benchmarks.

Our findings indicate that inter-lingual features lead to improvements on both similarity and categorization benchmarks with a more significant effect on categorization. Our results on the similarity benchmarks indicate that inter-lingual visual grounding is more beneficial for related languages such as English and German, but can lead to reduced performance when unrelated languages, such as English and Arabic, or German and Arabic, are considered jointly. On the other hand, Arabic provided the most improvement on categorization benchmarks for grounded English embeddings.

We hope that these initial steps towards inter-lingual visual grounding inspire further research. Low-resourced languages might benefit from joint processing with high-resourced languages in multilingual models but one has to make sure that their unique characteristics are not overwhelmed and masked by datasets acquired in different cultural settings.

Limitations

The architecture that we made use of for exploring multi-lingual visual grounding has the limitation that embeddings from different languages, which define high-dimensional spaces that are in all likelihood not congruent, constitute the input for visual grounding. One direction for future research is to first align the embeddings of different languages. A large multilingual language model such as XLM (Lample and Conneau, 2019) may help to better capture shared inter-lingual features, while at the same time retaining the linear alignment that restricts the extent to which vision can affect text-based semantics. Another possibility is to use an unsupervised technique (Conneau et al., 2017) to generate cross-lingual embeddings, which can then be used as initializers for our grounding architecture.

Acknowledgements

This work has been supported by EXC number 2064/1 – Project number 390727645, the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and the European research council, project WIDE-742545. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hassan Shahmohammadi. We appreciate the contribution of (Biswas et al., 2021) from the German Research Center of Artificial Intelligence (DFKI) who have provided us with the German COCO captions upon request.

References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. *Towards a Cleaner Document-Oriented Multilingual Crawled Corpus*. *arXiv e-prints*, page arXiv:2201.06642.
- Faaza A Almarsoomi, James D OShea, Zuhair Bandar, and Keeley Crockett. 2013. Awss: An algorithm for measuring arabic word semantic similarity. In *2013 IEEE international conference on systems, man, and cybernetics*, pages 504–509. IEEE.
- Abdulrahman Almuahreb and Massimo Poesio. 2005. Concept learning and categorization from the web. In *proceedings of the annual meeting of the Cognitive Science society*, volume 27.
- Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? a comprehensive assessment for catalan. *arXiv preprint arXiv:2107.07903*.
- Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10.
- William F Battig and William E Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*, 80(3p2):1.
- Rajarshi Biswas, Michael Barz, Mareike Hartmann, and Daniel Sonntag. 2021. Improving german image captions using machine translation and transfer learning. In *International Conference on Statistical Language and Speech Processing*, pages 3–14. Springer.
- Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

- Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *European Conference on Computer Vision*, pages 197–213. Springer.
- Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. Association for Computational Linguistics.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Timothy Dozat. 2016. Incorporating nesterov momentum into adam.
- ESLLI. 2009. <https://esslli2009.labri.fr/>.
- ESLLI-a. 2009. http://wordspace.collocations.de/doku.php/data:esslli2008:concrete_nouns_categorization.
- ESLLI-b. 2009. http://wordspace.collocations.de/doku.php/data:esslli2008:abstract_concrete_nouns_discrimination.
- ESLLI-c. 2009. http://wordspace.collocations.de/doku.php/data:esslli2008:verb_categorization.
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.
- 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*.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- Mika Hasegawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2017. Incorporating visual features into word embeddings: A bimodal autoencoder-based approach. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.
- Muhammad Hashim. 2020. Arabic coco. <https://github.com/canese-project/Arabic-COCO>.
- Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1192–1201.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.
- Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956.
- Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933.
- Karol Kurach, Sylvain Gelly, Michal Jastrzebski, Philip Haeusser, Olivier Teytaud, Damien Vincent, and Olivier Bousquet. 2017. Better text understanding through image-to-text transfer. *arXiv preprint arXiv:1705.08386*.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.
- Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, Sofia, Bulgaria.
- Alireza Mohammadshahi, Rémi Lebret, and Karl Aberer. 2019. Aligning multilingual word embeddings for cross-modal retrieval task. *arXiv preprint arXiv:1910.03291*.
- Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 3977–3986.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346.

A Saif, MJ Ab Aziz, and N Omar. 2014. Evaluating knowledge-based semantic measures on arabic. *International Journal on Communications Antenna and Propagation*, 4(5):180–194.

Hassan Shahmohammadi, Maria Heitmeier, Elnaz Shafaei-Bajestan, Hendrik Lensch, and Harald Baayen. 2022. Language with vision: a study on grounded word and sentence embeddings. *arXiv preprint arXiv:2206.08823*.

Hassan Shahmohammadi, Hendrik Lensch, and R Harald Baayen. 2021. Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. *arXiv preprint arXiv:2104.07500*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

S. N. Wood. 2017. *Generalized Additive Models*. Chapman & Hall/CRC, New York.

Mingyang Zhou, Luwei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.