# Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation

**Sandro Pezzelle, Ece Takmaz, Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
{s.pezzelle|e.takmaz|raquel.fernandez}@uva.nl

## Abstract

This study carries out a systematic *intrinsic* evaluation of the semantic representations learned by state-of-the-art pre-trained multimodal Transformers. These representations are claimed to be task-agnostic and shown to help on many downstream language-and-vision tasks. However, the extent to which they align with human semantic intuitions remains unclear. We experiment with various models and obtain *static* word representations from the *contextualized* ones they learn. We then evaluate them against the semantic judgments provided by human speakers. In line with previous evidence, we observe a generalized advantage of multimodal representations over language-only ones on concrete word pairs, but not on abstract ones. On the one hand, this confirms the effectiveness of these models to align language and vision, which results in better semantic representations for concepts that are *grounded* in images. On the other hand, models are shown to follow different representation learning patterns, which sheds some light on *how* and *when* they perform multimodal integration.

## 1 Introduction

Increasing evidence indicates that the meaning of words is multimodal: Human concepts are *grounded* in our senses (Barsalou, 2008; De Vega et al., 2012), and the sensory-motor experiences humans have with the world play an important role in determining word meaning (Meteyard et al., 2012). Since (at least) the first operationalizations of the distributional hypothesis, however, standard NLP approaches to derive meaning representations of words have solely relied on information extracted from large *text corpora*, based on the generalized assumption that the meaning of a word can be inferred from the effects it has on its linguistic context (Harris, 1954; Firth, 1957). Language-only semantic representations, from pioneering 'count' vectors (Landauer and Dumais, 1997; Turney and Pantel, 2010; Pennington et al., 2014) to either *static* (Mikolov et al., 2013) or *contextualized* (Peters et al., 2018; Devlin et al., 2019) neural network-based embeddings, have proven extremely effective in many linguistic tasks and applications, for which they constantly increased state-of-the-art performance. However, they naturally have no connection with the real-world referents they denote (Baroni, 2016). As such, they suffer from the symbol grounding problem (Harnad, 1990), which in turn limits their cognitive plausibility (Rotaru and Vigliocco, 2020).

To overcome this limitation, several methods have been proposed to equip language-only representations with information from concurrent modalities, particularly vision. Until not long ago, the standard approach aimed to leverage the complementary information conveyed by language and vision—for example, that bananas are *yellow* (vision) and *rich in potassium* (language)—by building richer multimodal representations (Beinborn et al., 2018). Overall, these representations have proved advantageous over purely textual ones in a wide range of tasks and evaluations, including the approximation of human semantic similarity/relatedness judgments provided by benchmarks like SimLex999 (Hill et al., 2015) or MEN (Bruni et al., 2014). This was taken as evidence that leveraging multimodal information leads to more human-like, full-fledged semantic representations of words (Baroni, 2016).

More recently, the advent of Transformer-based pre-trained models such as BERT (Devlin et al., 2019) has favored the development of a plethora of multimodal models (Li et al., 2019; Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2020) aimed to solve downstream language and vision tasks such as Visual Question



Transactions of the Association for Computational Linguistics, vol. 9, pp. 1563–1579, 2021. https://doi.org/10.1162/tacl_a_00443
Action Editor: Jing Jiang. Submission batch: 6/2021; Revision batch: 7/2021; Published 12/2021.

Answering (Antol et al., 2015) and Visual Dialogue (De Vries et al., 2017; Das et al., 2017). Similarly to the revolution brought about by Transformer-based language-only models to NLP (see Tenney et al., 2019), these systems have rewritten the recent history of research on language and vision by setting new state-of-the-art results on most of the tasks. Moreover, similarly to their language-only counterparts, these systems have been claimed to produce all-purpose, 'task-agnostic' representations ready-made for any task.

While there has been quite a lot of interest in understanding the inner mechanisms of BERT-like models (see the interpretability line of research referred to as *BERTology*; Rogers et al., 2020) and the nature of their representations (Mickus et al., 2020; Westera and Boleda, 2019), comparably less attention has been paid to analyzing the multimodal equivalents of these models. In particular, no work has explicitly investigated how the representations learned by these models compare to those by their language-only counterparts, which were recently shown to outperform standard *static* representations in approximating people's semantic intuitions (Bommasani et al., 2020).

In this work, we therefore focus on the representations learned by state-of-the-art multimodal pre-trained models, and explore whether, and to what extent, leveraging visual information makes them closer to human representations than those produced by BERT. Following the approach proposed by Bommasani et al. (2020), we derive *static* representations from the *contextualized* ones produced by these Transformer-based models. We then analyze the quality of such representations by means of the standard *intrinsic* evaluation based on correlation with human similarity judgments.

We evaluate LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020), ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), and Vokenization (Tan and Bansal, 2020) on five human judgment benchmarks[1] and show that: (1) in line with previous work, multimodal models outperform purely textual ones in the representation of concrete, but not abstract, words; (2) representations by Vokenization stand out as the overall best-performing multimodal ones; and (3) multimodal models differ with respect to *how* and *when* they integrate information from language and vision, as revealed by their learning patterns across layers.

[1]Data and code can be found at `https://github.com/sandropezzelle/multimodal-evaluation`.

## 2 Related Work

### 2.1 Evaluating Language Representations

Evaluating the *intrinsic* quality of learned semantic representations has been one of the main, long-standing goals of NLP (for a recent overview of the problem and the proposed approaches, see Navigli and Martelli, 2019; Taieb et al., 2020). In contrast to *extrinsic* evaluations that measure the effectiveness of task-specific representations in performing downstream NLU tasks (e.g., those contained in the GLUE benchmark; Wang et al., 2019), the former approach tests whether, and to what extent, task-agnostic semantic representations (i.e., not learned nor fine-tuned to be effective on some specific tasks) align with those by human speakers. This is typically done by measuring the correlation between the similarities computed on system representations and the semantic similarity judgments provided by humans, a natural testbed for distributional semantic models (Landauer and Dumais, 1997). Lastra-Díaz et al. (2019) provide a recent, comprehensive survey on methods, benchmarks, and results.

In the era of Transformers, recent work has explored the relationship between the *contextualized* representations learned by these models and the *static* ones learned by distributional semantic models (DSMs). On a formal level, some work has argued that this relation is not straightforward since only context-invariant—but not *contextualized*—representations may adequately account for expression meaning (Westera and Boleda, 2019). In parallel, Mickus et al. (2020) focused on BERT and explored to what extent the semantic space learned by this model is comparable to that by DSMs. Though an overall similarity was reported, BERT's next-sentence-prediction objective was shown to partly obfuscate this relation. A more direct exploration of the intrinsic semantic quality of BERT representations was carried out by Bommasani et al. (2020). In their work, BERT's *contextualized* representations were first turned into *static* ones by means of simple methods (see Section 4.2) and then evaluated against several similarity benchmarks. These representations were shown to outperform traditional ones, which revealed that pooling over many contexts improves embeddings' representational quality.

Recently, Ilharco et al. (2021) probed the representations learned by purely textual language models in their ability to perform *language grounding*. Though far from human performance, they were shown to learn nontrivial mappings to vision.

### 2.2 Evaluating Multimodal Representations

Since the early days of DSMs, many approaches have been proposed to enrich language-only representations with information from images. Bruni et al. (2012, 2014) equipped textual representations with low-level visual features, and reported an advantage over language-only representations in terms of correlation with human judgments. An analogous pattern of results was obtained by Kiela and Bottou (2014) and Kiela et al. (2016), who concatenated visual features obtained with convolutional neural networks (CNNs) with skip-gram linguistic representations. Lazaridou et al. (2015) further improved over these techniques by means of a model trained to optimize the similarity of words with their visual representations, an approach similar to that by Silberer and Lapata (2014). Extensions of these latter methods include the model by Zablocki et al. (2018), which leverages information about the visual context in which objects appear; and Wang et al. (2018), where three dynamic fusion methods were proposed to learn to assign importance weights to each modality. More recently, some work has explored the quality of representations learned from images only (Lüddecke et al., 2019) or by combining language, vision, and emojis (Rotaru and Vigliocco, 2020). In parallel, new evaluation methods based, for example, on decoding brain activity (Davis et al., 2019) or success on tasks such as image retrieval (Kottur et al., 2016) have been proposed. This mass of studies has overall demonstrated the effectiveness of multimodal representations in approximating human semantic intuitions better than purely textual ones. However, this advantage has been typically reported for concrete, but not abstract, concepts (Hill and Korhonen, 2014).

In recent years, the revolution brought about by Transformer-based multimodal models has fostered research that sheds light on their inner workings. One approach has been to use probing tasks: Cao et al. (2020) focused on LXMERT and UNITER and systematically compared the two models with respect to, for example, the degree of integration of the two modalities at each layer or the role of various attention heads (for a similar analysis on VisualBERT, see Li et al., 2020). Using two tasks (image-sentence verification and counting) as testbeds, Parcalabescu et al. (2021) highlighted capabilities and limitations of various pre-trained models to integrate modalities or handle dataset biases. Another line of work has explored the impact of various experimental choices, such as pre-training tasks and data, loss functions and hyperparameters, on the performance of pre-trained multimodal models (Singh et al., 2020; Hendricks et al., 2021). Since all of these aspects have proven to be crucial for these models, Bugliarello et al. (2021) proposed VOLTA, a unified framework to pre-train and evaluate Transformer-based models with the same data, tasks and visual features.

Despite the renewed interest in multimodal models, to the best of our knowledge no work has explored, to date, the *intrinsic* quality of the task-agnostic representations built by various pre-trained Transformer-based models. In this work, we tackle this problem for the first time.

## 3 Data

We aim to evaluate how the similarities between the representations learned by pre-trained multimodal Transformers align with the similarity judgments by human speakers, and how these representations compare to those by textual Transformers such as BERT. To do so, we need data that (1) is multimodal, that is, where some text (language) is paired with a corresponding image (vision), and (2) includes most of the words making up the word pairs for which human semantic judgments are available. In what follows, we describe the semantic benchmarks used for evaluation and the construction of our multimodal dataset.

### 3.1 Semantic Benchmarks

We experiment with five human judgment benchmarks used for *intrinsic* semantic evaluation in both language-only and multimodal work: RG65 (Rubenstein and Goodenough, 1965), WordSim-353 (Finkelstein et al., 2002), SimLex999 (Hill et al., 2015), MEN (Bruni et al., 2014), and SimVerb3500 (Gerz et al., 2016). These benchmarks have a comparable format, namely, they contain $N$ $\langle w_1, w_2, score \rangle$ samples, where $w_1$ and $w_2$ are two distinct words, and $score$ is a bounded value—that we normalize to range in

| | | | *original* | | | *found in VICO* | | |
|---|---|---|---|---|---|---|---|---|
| *benchmark* | *rel.* | *PoS* | *# pairs* | *# W* | *concr. (#)* | *# pairs (%)* | *# W (%)* | *concr. (#)* |
| RG65 | S | N | 65 | 48 | 4.37 (65) | 65 (100%) | 48 (100%) | 4.37 (65) |
| WordSim353 | R | N, V, Adj | 353 | 437 | 3.82 (331) | 306 (86.7%) | 384 (87.9%) | 3.91 (300) |
| SimLex999 | S | N, V, Adj | 999 | 1028 | 3.61 (999) | 957 (95.8%) | 994 (99.5%) | 3.65 (957) |
| MEN | R | N, V, Adj | 3000 | 752 | 4.41 (2954) | 2976 (99.2%) | 750 (99.7%) | 4.41 (2930) |
| SimVerb3500 | S | V | 3500 | 827 | 3.08 (3487) | 2890 (82.6%) | 729 (88.2%) | 3.14 (2890) |
| *total* | | | 7917 | 2453 | | 7194 (90.9%) | 2278 (92.9%) | |

Table 1: Statistics of the benchmarks before (*original*) and after (*found in VICO*) filtering them based on VICO: *rel.* refers to the type of semantic relation, i.e., (S)imilarity or (R)elatedness; *# W* to the number of unique words present; *concr.* to average concreteness of the pairs (in brackets, # found in Brysbaert et al., 2014). Within *found in VICO*, percentages in brackets refer to the coverage compared to *original*.

$[0, 1]$—which stands for the degree of semantic similarity or relatedness between $w_1$ and $w_2$: The higher the value, the more similar the pair. At the same time, these benchmarks differ in several respects, namely, (1) the type of semantic relation they capture (i.e., similarity or relatedness); (2) the parts-of-speech (PoS) they include; (3) the number of pairs they contain; (4) the size of their vocabulary (i.e., the number of unique words present); and (5) the words' degree of concreteness, which previous work found to be particularly relevant for evaluating the performance of multimodal representations (see Section 2.2). We report descriptive statistics of all these relevant features in Table 1 (*original* section). For concreteness, we report a single score for each benchmark: the higher, the more concrete. We obtained this score (1) by taking, for each word, the corresponding 5-point human rating collected by Brysbaert et al. (2014);[2] (2) by computing the average concreteness of each pair; and (3) by averaging over the entire benchmark.

### 3.2 Dataset

Previous work evaluating the *intrinsic* quality of multimodal representations has faced the issue of limited vocabulary coverage in the datasets used. As a consequence, only a subset of the tested benchmarks has often been evaluated (e.g., 29% of word pairs in SimLex999 and 42% in MEN, reported by Lazaridou et al., 2015). To overcome this issue, we jointly consider two large multimodal datasets: Common Objects in Contexts (COCO; Lin et al., 2014) and Visual Storytelling (VIST; Huang et al., 2016). The former contains samples where a natural image is paired with a free-form, crowdsourced description (or caption) of its visual content. The latter contains samples where a natural image is paired with both a description of its visual content (DII, Descriptions of Images in Isolation) and a fragment of a story invented based on a sequence of five images to which the target image belongs (SIS, Stories of Images in Sequences). Both DII and SIS contain crowdsourced, free-form text. In particular, we consider the entire COCO 2017 data (the concatenation of train and val splits), which consists of 616,767 $\langle image, description \rangle$ samples. As for VIST, we consider the train, val, and test splits of both DII and SIS, which sum up to 401,600 $\langle image, description/story \rangle$ samples.

By concatenating VIST and COCO, we obtain a dataset containing 1,018,367 $\langle image, sentence \rangle$ samples, that we henceforth refer to as VICO. Thanks to the variety of images and, in particular, the types of text it contains, the concatenated dataset proves to be very rich in terms of lexicon, an essential *desideratum* for having broad coverage of the word pairs in the semantic benchmarks. We investigate this by considering all the 7917 word pairs making up the benchmarks and checking, for each pair, whether both of its words are present at least once in VICO. We find 7194 pairs made up of 2278 unique words. As can be seen in Table 1 (*found in VICO* section), this is equivalent to around 91% of total pairs found (min. 83%, max. 100%), with an overall vocabulary coverage of around 93% (min. 88%, max. 100%). This is reflected in a pattern of average concreteness scores that is essentially equivalent to *original*. Figure 1 reports this pattern in a boxplot.

[2]Participants were instructed that *concrete* words refer to things/actions that can be experienced through our senses, while meanings of *abstract* words are defined by other words.

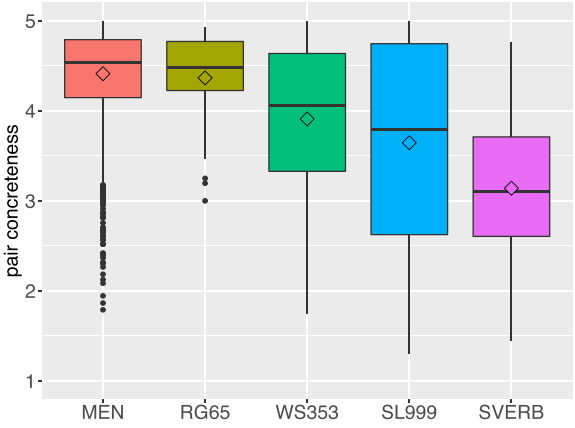



Figure 1: Concreteness of word pairs found in VICO. Concreteness ranges from 1 to 5. The horizontal line shows median; ◇, mean.

| | *# samples* | *# imgs* | *sent. L* | *# W* | *# WP* |
|---|---|---|---|---|---|
| COCO | 50452 | 39767 | 13.4 | 1988 | 6076 |
| VIST | 63256 | 40528 | 14.7 | 2250 | 7122 |
| *total* | 113708 | 80295 | 14.7 | 2278 | 7194 |

Table 2: Dataset statistics. *# imgs*: number of unique images; *sent. L*: average sentence length; *# W*, *# WP*: resp., number of words and word pairs.

Since experimenting with more than 1 million $\langle image, sentence \rangle$ samples turns out to be computationally highly demanding, for efficiency reasons we extract a subset of VICO such that: (1) all 2278 words found in VICO (hence, the vocabulary) and the corresponding 7194 pairs are present at least once among its sentences; (2) its size is around an order of magnitude smaller than VICO; (3) it preserves the word frequency distribution observed in VICO. We obtain a subcorpus including 113,708 unique $\langle image, sentence \rangle$ samples, that is, around 11% of the whole VICO. Since all the experiments reported in the paper are performed on this subset, from now on we will simply refer to it as our dataset. Some of its descriptive statistics are reported in Table 2.[3] Interestingly, VIST samples contain more vocabulary words compared to COCO (2250 vs. 1988 words), which is reflected in higher coverage of word pairs (7122 vs. 6076).

## 4 Experiments

In our experiments, we build representations for each of the words included in our semantic benchmarks by means of various language-only and multimodal models (Section 4.1). In all cases, representations are extracted from the samples included in our dataset. In the language-only models, representations are built based on only the sentence; in the multimodal models, based on the sentence and its corresponding image (or, for Vokenization, just the sentence but with visual supervision in pre-training, as explained later). Since representations by most of the tested models are *contextualized*, we make them *static* by means of an aggregation method (Section 4.2). In the evaluation (Section 4.3), we test the ability of these representations to approximate human semantic judgments.

### 4.1 Models

**Language-Only Models** We experiment with one distributional semantic model producing *static* representations, namely, GloVe (Pennington et al., 2014) and one producing *contextualized* representations, namely, the pre-trained Transformer-based BERT (Devlin et al., 2019). For GloVe, following Bommasani et al. (2020) we use its 300-d word representations pre-trained on 6B tokens from Wikipedia 2014 and Gigaword 5.[4] As for BERT, we experiment with its standard 12-layer version (BERT-base).[5] This is the model serving as the backbone of all the multimodal models we test, which allows for direct comparison.

**Multimodal Models** We experiment with five pre-trained Transformer-based multimodal models. Four of them are both pre-trained and evaluated using multimodal data, that is, they produce representations based on a sentence and an image (Language and Vision; $LV$) at both training and inference time: LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020), ViLBERT (Lu et al., 2019), and VisualBERT (Li et al., 2019). One of them, in contrast, is visually supervised during training, but only takes Language as input during inference ($L_V$): Vokenization (Tan and Bansal, 2020). All five models are similar in three main respects: (1) they have BERT as their backbone; (2) they produce *contextualized* representations; and (3) they have multiple layers from which such representations can be extracted.

[3] The average frequency of our vocabulary words is 171 (min. 1, max. 8440). 61 words (3%) have frequency 1.

[4] `http://nlp.stanford.edu/data/glove.6B.zip`.

[5] We adapt the code from: `https://github.com/rishibommasani/Contextual2Static`.

As for the $LV$ models, we use reimplementations by the VOLTA framework (Bugliarello et al., 2021).[6] This has several advantages since all the models: (1) are initialized with BERT weights;[7] (2) use the same visual features, namely, 36 regions of interest extracted by Faster R-CNN with a ResNet-101 backbone (Anderson et al., 2018);[8] and (3) are pre-trained in a controlled setting using the same exact data (Conceptual Captions; Sharma et al., 2018), tasks, and objectives, that is, Masked Language Model (MLM), masked object classification with KL-divergence, and image-text matching (ITM), a binary classification problem to predict whether an image and text pair match. This makes the four $LV$ models directly comparable to each other, with no confounds. Most importantly, each model is reimplemented as a particular instance of a unified mathematical framework based on the innovative gated bimodal Transformer layer. This general layer can be used to model both intra-modal and inter-modal interactions, which makes it suitable to reimplement both *single-stream* models (where language and vision are jointly processed by a single encoder; UNITER, VisualBERT) and *dual-stream* models (where the two modalities are first processed separately and then integrated; LXMERT, ViLBERT).

As for Vokenization, we use the original implementation by Tan and Bansal (2020). This model is essentially a visually supervised language model which, during training, extracts multimodal alignments to language-only data by contextually mapping words to images. Compared to $LV$ models where alignment between language and vision is performed at the $\langle sentence, image \rangle$ level, in Vokenization the mapping is done at the token level (the image is named *voken*). It is worth mentioning that Vokenization is pre-trained with less textual data compared to the standard BERT, the model used to initialize all $LV$ architectures. For comparison, in Table 3 we report the tasks and data used to pre-train each of the tested models. None of the tested $LV$ models were pre-trained with data present in our dataset. For Vokenization, we cannot exclude that some COCO samples of our dataset were also used in the TIM task.

[6] https://github.com/e-bug/volta.

[7] Including LXMERT, which was initialized from scratch in its original implementation.

[8] Our code to extract visual features for all our images is adapted from: https://github.com/airsplay/py-bottom-up-attention/blob/master/demo/demo_feature_extraction_attr.ipynb.

| | *pre-training task(s)* | *pre-training data* |
|---|---|---|
| GloVe | Unsupervised vector learning | Wikipedia 2014 + Gigaword 5 |
| BERT | Masked Language Model (MLM) + Next Sentence Prediction (NSP) | English Wikipedia + BooksCorpus |
| $LV$* | Masked Language Model (MLM) + Masked Object Classification KL + Image-Text Matching (ITM) | Conceptual Captions |
| Vok. | Token-Image Matching (TIM)* | COCO + Visual Genome |
| | Masked Language Model (MLM) | English Wikipedia + Wiki103 |

Table 3: Overview of tasks and data used to pre-train models. $LV$ refers to the 4 multimodal models; Vok. to Vokenization. *Initialized with BERT.

### 4.2 Aggregation Method

With the exception of GloVe, all our tested models build *contextualized* representations. We closely follow the method proposed by Bommasani et al. (2020) and, for each word in our benchmarks, we compute a single, *static* context-agnostic representation using the samples included in our dataset. This involves two steps: (1) subword pooling and (2) context combination. A schematic illustration of both these steps is shown in Figure 2, where we exemplify the general architecture of a $LV$ model. Subword pooling is the operation by which we construct a word representation from the tokens produced by the BERT tokenizer. Since, during tokenization, some words (e.g., 'donut') get decomposed into $N$ subwords (`don`, `##ut`), we apply a function to combine the representations $s_1, \ldots, s_N$ produced for each subword token $t_k, \ldots, t_{k+N-1}$ into a *contextualized* word-level representation $w_c$. We take the corresponding model hidden states as our representations, and use arithmetic mean as the combination function, following Bommasani et al. (2020):

$$w_c = mean(s_1, \ldots, s_N) \quad (1)$$

For each word, we then compute a *static* representation from its *contextualized* representations. This is done via context combination, where we aggregate the contextualized representations $w_{c_1}$, $\ldots$, $w_{c_M}$ of the same word found in $M$ sentences (contexts). We obtain a single *static* representation for the word, again using arithmetic mean:

$$w = mean(w_{c_1}, \ldots, w_{c_M}) \quad (2)$$

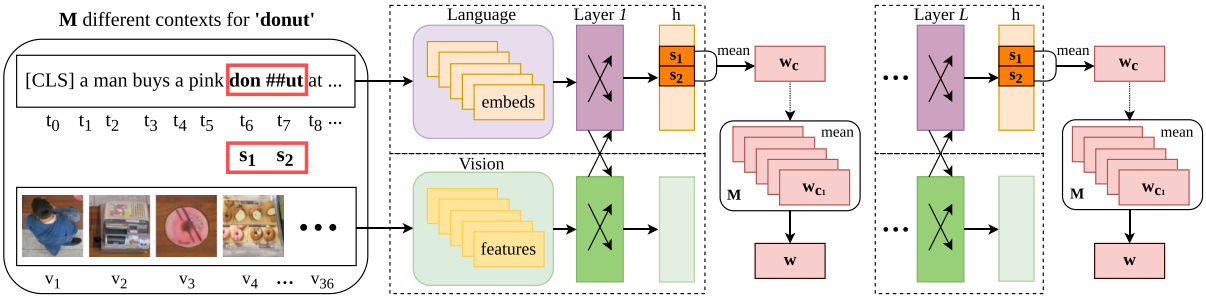



Figure 2: A schematic illustration of our method to obtain *static* representations of words, e.g., 'donut'.

As a result, for each of the 2278 words in our vocabulary we obtain a single static 768-d representation $w$. The operations of aggregating subwords (Eq. 1) and contextualized representations (Eq. 2) are carried out for each layer of each tested model.

For BERT, we consider 13 layers, from 0 (the input embedding layer) to 12. For Vokenization, we consider its 12 layers, from 1 to 12. For *LV* models, we consider the part of VOLTA's gated bimodal layer processing the language input, and extract activations from each of the feed-forward layers following a multi-head attention block. In LXMERT, there are 5 such layers: 21, 24, 27, 30, 33; in both UNITER and VisualBERT, 12 layers: 2, 4, . . . , 24; in ViLBERT, 12 layers: 14, 16, . . . , 36.[9] Representations are obtained by running the best snapshot of each pre-trained model[10] on our samples in evaluation mode, i.e., without fine-tuning nor updating the model's weights.

### 4.3 Evaluation

To evaluate the *intrinsic* quality of the obtained representations, we compare the semantic space of each model with that of human speakers. For each word pair in each semantic benchmark described in Section 3.1, we compute the cosine similarity between the representations of the words in the pair:

$$similarity = cosine(w_1, w_2) \quad (3)$$

For each benchmark, we then compute Spearman's rank $\rho$ correlation between the similarities obtained by the model and the normalized human judgments: The higher the correlation, the more aligned the two semantic spaces are.

[9]For reproducibility reasons, we report VOLTA's indexes.

[10]https://github.com/e-bug/volta/blob/main/MODELS.md.

## 5 Results

In Table 4, we report the best results obtained by all tested models on the five benchmarks. In brackets we report the number of the model layer.

**Language-Only Models** We notice that BERT evaluated on our dataset (hence, just BERT) systematically outperforms GloVe. This is in line with Bommasani et al. (2020), and replicates their findings that, as compared to standard *static* embeddings, averaging over *contextualized* representations by Transformer-based models is a valuable method for obtaining semantic representations that are more aligned to those of humans.

It is interesting to note, moreover, that the results we obtain with BERT actually outperform the best results reported by Bommasani et al. (2020) using the same model on 1M Wikipedia contexts (BERT-1M-Wiki). This is intriguing since it suggests that building representations using a dataset of visually grounded language, as we do, is not detrimental to the representational power of the resulting embeddings. Since this comparison is partially unfair due to the different methods employed in selecting language contexts, we also obtain results on a subset of Wikipedia that we extract using the method described for VICO (see Section 3.2),[11] and which is directly comparable to our dataset. As can be seen, representations built on this subset of Wikipedia (BERT-Wiki *ours*) turn out to perform better than those by BERT for WordSim353, SimLex999, and SimVerb3500 (the least concrete benchmarks—see Table 1); worse for RG65 and MEN (the most concrete ones). This pattern of results indicates that visually grounded language is different from encyclopedic one, which in turn has an impact on the resulting representations.

[11]This subset contains 127,246 unique sentences.

| *model* | *input* | *Spearman ρ correlation (layer)* | | | | |
|---|---|---|---|---|---|---|
| | | RG65 | WS353 | SL999 | MEN | SVERB |
| BERT-1M-Wiki* | $L$ | 0.7242 (1) | 0.7048 (1) | 0.5134 (3) | – | 0.3948 (4) |
| BERT-Wiki *ours* | $L$ | 0.8107 (1) | 0.7262 (1) | 0.5213 (0) | 0.7176 (2) | 0.4039 (4) |
| GloVe | $L$ | 0.7693 | 0.6097 | 0.3884 | 0.7296 | 0.2183 |
| BERT | $L$ | 0.8124 (2) | **0.7096 (1)** | **0.5191 (0)** | 0.7368 (2) | **0.4027 (3)** |
| LXMERT | $LV$ | 0.7821 (27) | 0.6000 (27) | 0.4438 (21) | 0.7417 (33) | 0.2443 (21) |
| UNITER | $LV$ | 0.7679 (18) | 0.6813 (2) | 0.4843 (2) | 0.7483 (20) | 0.3926 (10) |
| ViLBERT | $LV$ | 0.7927 (20) | 0.6204 (14) | 0.4729 (16) | 0.7714 (26) | 0.3875 (14) |
| VisualBERT | $LV$ | 0.7592 (2) | 0.6778 (2) | 0.4797 (4) | 0.7512 (20) | 0.3833 (10) |
| Vokenization | $L_V$ | **0.8456 (9)** | 0.6818 (3) | 0.4881 (9) | **0.8068 (10)** | 0.3439 (9) |

Table 4: Spearman's rank $\rho$ correlation between similarities computed with representations by all tested models and human similarity judgments in the five evaluation benchmarks: the higher the better. Results in **bold** are the highest in the column among models run on our dataset (that is, all but the top 2 models). Results underlined are the highest among $LV$ models. *Original results from Bommasani et al. (2020).

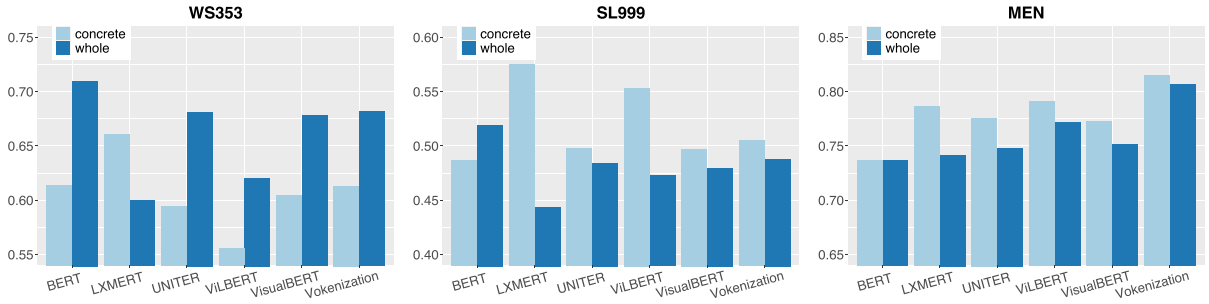



Figure 3: Highest $\rho$ by each model on WordSim353, SimLex999 and MEN. Each barplot reports results on both the *whole* benchmark (Table 4) and the most *concrete* subset of it (Table 5). Best viewed in color.

**Multimodal Models** Turning to multimodal models, we observe that they outperform BERT on two benchmarks, RG65 and MEN. Though Vokenization is found to be the best-performing architecture on both of them, all multimodal models surpass BERT on MEN (see rightmost panel of Figure 3; dark blue bars). In contrast, no multimodal model outperforms or is on par with BERT on the other three benchmarks (Figure 3 shows the results on WordSim353 and SimLex999). This indicates that multimodal models have an advantage on benchmarks containing more concrete word pairs (recall that MEN and RG65 are the overall most concrete benchmarks; see Table 1); in contrast, leveraging visual information appears to be detrimental for more abstract word pairs, a pattern that is very much in line with what was reported for previous multimodal models (Bruni et al., 2014; Hill and Korhonen, 2014). Among multimodal models, Vokenization stands out as the overall best-performing model. This indicates that *grounding* a masked language model is an effective way to obtain semantic representations that are intrinsically good, as well as being effective in downstream NLU tasks (Tan and Bansal, 2020). Among the models using an actual visual input ($LV$), ViLBERT turns out to be best-performing on high-concreteness benchmarks, while UNITER is the best model on more abstract benchmarks. This pattern could be due to the different embedding layers of these models, which are shown to play an important role (Bugliarello et al., 2021).

**Concreteness** Our results show a generalized advantage of multimodal models on more concrete benchmarks. This seems to indicate that visual information is beneficial for representing concrete words. However, it might still be that models are just better at representing the specific words contained in these benchmarks. To further investigate this point, for each benchmark we extract the subset of pairs where both words have concreteness

| *model* | *input* | *concr.* | *Spearman ρ correlation (layer)* | | | | |
|---|---|---|---|---|---|---|---|
| | | | RG65 | WS353 | SL999 | MEN | SVERB |
| BERT | $L$ | $\geq 4$ | 0.8321 (2) | 0.6138 (1) | 0.4864 (0) | 0.7368 (2) | 0.1354 (3) |
| LXMERT | $LV$ | $\geq 4$ | 0.8648 (27) | **0.6606 (27)** | **0.5749 (21)** | 0.7862 (33) | 0.1098 (21) |
| UNITER | $LV$ | $\geq 4$ | 0.8148 (18) | 0.5943 (2) | 0.4975 (2) | 0.7755 (20) | 0.1215 (10) |
| ViLBERT | $LV$ | $\geq 4$ | 0.8374 (20) | 0.5558 (14) | 0.5534 (16) | 0.7910 (26) | **0.1529 (14)** |
| VisualBERT | $LV$ | $\geq 4$ | 0.8269 (2) | 0.6043 (2) | 0.4971 (4) | 0.7727 (20) | 0.1310 (10) |
| Vokenization | $L_V$ | $\geq 4$ | **0.8708 (9)** | 0.6133 (3) | 0.5051 (9) | **0.8150 (10)** | 0.1390 (9) |
| *# pairs (%)* | | | 44 (68%) | 121 (40%) | 396 (41%) | 1917 (65%) | 210 (7%) |

Table 5: Correlation on the concrete subsets (concr. $\geq 4$) of the five evaluation benchmarks. Results in **bold** are the highest in the column. Results underlined are the highest among $LV$ multimodal models.

$\geq$4 out of 5 in Brysbaert et al. (2014). For each model, we consider the results by the layer which is best-performing on the whole benchmark. Table 5 reports the results of this analysis, along with the number (and %) of word pairs considered.

For all benchmarks, there is always at least one multimodal model that outperforms BERT. This pattern is crucially different from that observed in Table 4, and confirms that multimodal models are better than language-only ones at representing concrete words, regardless of their PoS. Zooming into the results, we note that Vokenization still outperforms other multimodal models on both RG65 and MEN (see rightmost panel of Figure 3; light blue bars), while LXMERT turns out to be the best-performing model on both WordSim-353 and SimLex999 (see left and middle panels of Figure 3; light blue bars). These results suggest that this model is particularly effective in representing highly concrete words, but fails with abstract ones, which could cause the overall low correlations in the full benchmarks (Table 4). ViLBERT obtains the best results on SimVerb-3500, thus confirming the good performance of this model in representing verbs/actions seen also in Table 4. However, the low correlation that all models achieve on this subset indicates that they all struggle to represent the meaning of verbs that are deemed very concrete. This finding appears to be in line with the generalized difficulty in representing verbs reported by Hendricks and Nematzadeh (2021). Further work is needed to explore this issue.

## 6 Analysis

We perform analyses aimed at shedding light on commonalities and differences between the various models. In particular, we explore how model performance evolves through layers (Section 6.1), and how various models compare to humans at the level of specific word pairs (Section 6.2).

### 6.1 Layers

Table 4 reports the results by the best-performing layer of each model. Figure 4 complements these numbers by showing, for each model, how performance changes across various layers. For BERT, Bommasani et al. (2020) found an advantage of earlier layers in approximating human semantic judgments. We observe the same exact pattern, with earlier layers (0–3) achieving the best correlation scores on all benchmarks and later layers experiencing a significant drop in performance. As for multimodal models, previous work (Cao et al., 2020) experimenting with UNITER and LXMERT revealed rather different patterns between the two architectures. For the former, a higher degree of integration between language and vision was reported in later layers; as for the latter, such integration appeared to be in place from the very first multimodal layer. Cao et al. (2020) hypothesized this pattern to be representative of the different behaviors exhibited by single-stream (UNITER, VisualBERT) vs. dual-stream (LXMERT, ViLBERT) models. If a higher degree of integration between modalities leads to better semantic representations, we should observe an advantage of later layers in UNITER and VisualBERT, but not in LXMERT and ViLBERT. In particular, we expect this to be the case for benchmarks where the visual modality plays a bigger role, i.e., the more concrete RG65 and MEN.

As can be seen in Figure 4, LXMERT exhibits a rather flat pattern of results, which overall

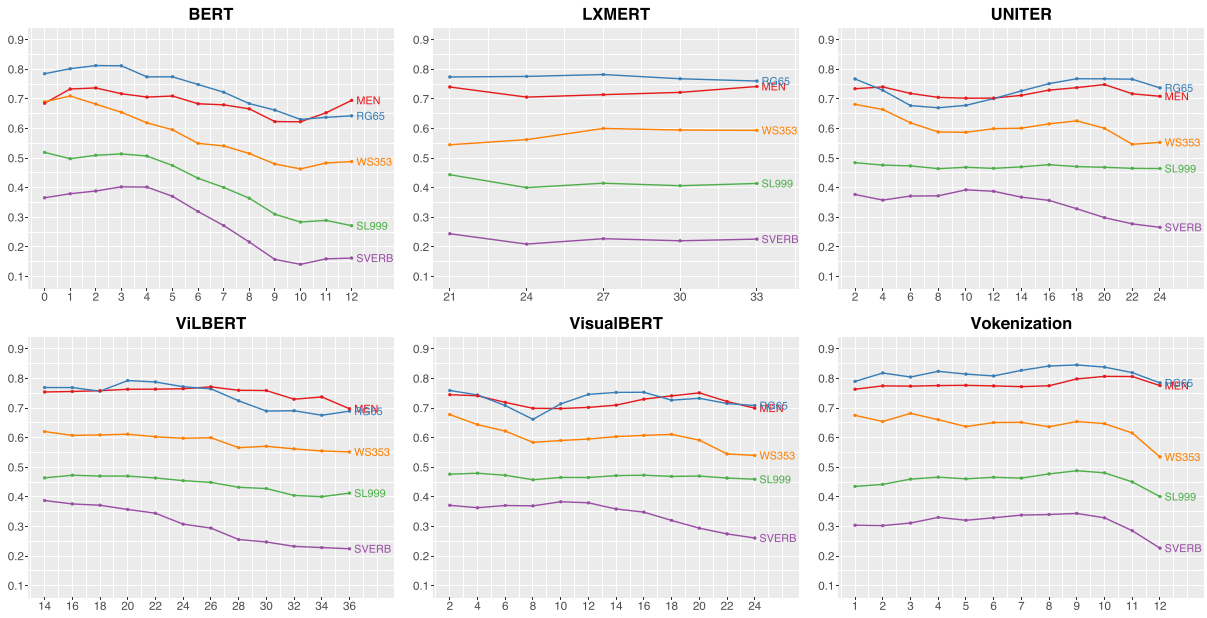



Figure 4: Correlation by all tested models on the five benchmarks across layers. Best viewed in color.

confirms the observation that, in this dual-stream model, integration of language and vision is in place from the very first multimodal layer. Conversely, we notice that single-stream UNITER achieves the best correlation on RG65 and MEN towards the end of its pipeline (at layers 18 and 20, respectively), which supports the hypothesis that later representations are more multimodal. The distinction between single- and dual-stream models appears less clear-cut in the other two architectures (not explored by Cao et al., 2020). Though ViLBERT (dual-stream) achieves generally good results in earlier layers, the best correlation on RG65 and MEN is reached in middle layers. As for VisualBERT (single-stream), consistently with the expected pattern the best correlation on MEN is achieved at one of the latest layers; however, the best correlation on RG65 is reached at the very first multimodal layer. Overall, our results mirror the observations by Cao et al. (2020) for LXMERT and UNITER. However, the somewhat mixed pattern observed for the other models suggests more complex interactions between the two modalities. As for Vokenization, there is a performance drop at the last two layers, but otherwise its performance constantly increases through the layers and reaches the highest peaks toward the end.

Taken together, the results of this analysis confirm that various models differ with respect to how they represent and process the inputs and to how and when they perform multimodal integration.

### 6.2 Pair-Level Analysis

Correlation results are not informative about (1) which word pairs are more or less challenging for the models, nor about (2) how various models compare to each other in dealing with specific word pairs. Intuitively, this could be tested by comparing the raw similarity values output by a given model to both human judgments and scores by other models. However, this turns out not to be sound in practice due to the different ranges of values produced. For example, some models output generally low cosine values, while others produce generally high scores,[12] which reflects differences in the density of the semantic spaces they learn. To compare similarities more fairly, for each model we consider the entire distribution of cosine values obtained in a given benchmark, rank it in descending order (from highest to lowest similarity values) and split it in five equally-sized bins, that we label *highest*, *high*, *medium*, *low*, *lowest*. We do the same for human similarity scores. Then, for each word pair, we check whether it is assigned the same similarity 'class' by both humans and the model. We focus on the three

[12] Differences also emerge between various model layers.

| | BERT | ViLBERT | Vokenization | BERT | ViLBERT | Vokenization |
|---|---|---|---|---|---|---|
| | *most similar pairs according to humans* | | | *least similar pairs according to humans* | | |
| **RG65** | gem, jewel | gem, jewel | gem, jewel | cord, smile | cord, smile | cord, smile |
| | midday, noon | midday, noon | midday, noon | noon, string | noon, string | noon, string |
| | automobile, car | automobile, car | automobile, car | rooster, voyage | rooster, voyage | rooster, voyage |
| | cemetery, graveyard | cemetery, graveyard | cemetery, graveyard | fruit, furnace | fruit, furnace | fruit, furnace |
| | cushion, pillow | cushion, pillow | cushion, pillow | autograph, shore | autograph, shore | autograph, shore |
| **WS353** | coast, shore | coast, shore | coast, shore | king, cabbage | king, cabbage | king, cabbage |
| | money, cash | money, cash | money, cash | chord, smile | chord, smile | chord, smile |
| | midday, noon | midday, noon | midday, noon | noon, string | noon, string | noon, string |
| | journey, voyage | journey, voyage | journey, voyage | professor, cucumber | professor, cucumber | professor, cucumber |
| | dollar, buck | dollar, buck | dollar, buck | rooster, voyage | rooster, voyage | rooster, voyage |
| **SL999** | stupid, dumb | stupid, dumb | stupid, dumb | cliff, tail | cliff, tail | cliff, tail |
| | creator, maker | creator, maker | creator, maker | container, mouse | container, mouse | container, mouse |
| | vanish, disappear | vanish, disappear | vanish, disappear | ankle, window | ankle, window | ankle, window |
| | quick, rapid | quick, rapid | quick, rapid | new, ancient | new, ancient | new, ancient |
| | insane, crazy | insane, crazy | insane, crazy | shrink, grow | shrink, grow | shrink, grow |
| **MEN** | sun, sunlight | sun, sunlight | sun, sunlight | feather, truck | feather, truck | feather, truck |
| | cat, kitten | cat, kitten | cat, kitten | angel, gasoline | angel, gasoline | angel, gasoline |
| | automobile, car | automobile, car | automobile, car | bakery, zebra | bakery, zebra | bakery, zebra |
| | river, water | river, water | river, water | bikini, pizza | bikini, pizza | bikini, pizza |
| | stair, staircase | stair, staircase | stair, staircase | muscle, tulip | muscle, tulip | muscle, tulip |
| **SVERB** | repair, fix | repair, fix | repair, fix | drive, breed | drive, breed | drive, breed |
| | triumph, win | triumph, win | triumph, win | die, grow | die, grow | die, grow |
| | build, construct | build, construct | build, construct | visit, giggle | visit, giggle | visit, giggle |
| | flee, escape | flee, escape | flee, escape | miss, catch | miss, catch | miss, catch |
| | rip, tear | rip, tear | rip, tear | shut, vomit | shut, vomit | shut, vomit |

Table 6: Similarity assigned by BERT, ViLBERT, and Vokenization to most (left) and least (right) similar pairs according to humans. Dark green indicates *highest* assigned similarity; light green, *high*; yellow, *medium*; orange, *low*; red, *lowest*. Best viewed in color.

overall best-performing models, namely, BERT ($L$), ViLBERT ($LV$), and Vokenization ($L_V$).

We perform a qualitative analysis by focusing on 5 pairs for each benchmark with the highest and lowest semantic similarity/relatedness according to humans. Table 6 reports the results of this analysis through colors. Dark green and red indicate alignment between humans and models on most similar and least similar pairs, respectively. At first glance, we notice a prevalence of dark green on the left section of the table, which lists 5 of the most similar pairs according to humans; a prevalence of red on the right section, which lists the least similar ones. This clearly indicates that the three models are overall effective in capturing similarities of words, mirroring the results reported in Table 4. Consistently, we notice that model representations are generally more aligned in some benchmarks compared to others: consider, for example, RG65 vs. SimLex999 or SimVerb3500. Moreover, some models appear to be more aligned than others in specific benchmarks: For example, in the highly concrete MEN, Vokenization is much more aligned than BERT on the least similar cases. In contrast, BERT is more aligned with humans than are multimodal models on the most similar pairs of SimLex999, to which ViLBERT (and, to a lesser extent, Vokenization), often assigns *low* and *medium* similarities. These qualitative observations are in line with the numbers reported in Table 7, which refer to the proportion of aligned cases between humans and the models within each benchmark. Interestingly, all models display

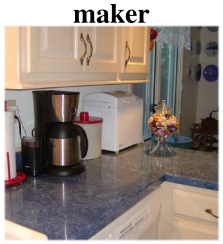
| | RG65 | WS353 | SL999 | MEN | SVERB |
|---|---|---|---|---|---|
| all | | | | | |
| BERT | 0.52 | 0.39 | 0.38 | 0.41 | 0.31 |
| ViLBERT | 0.49 | 0.37 | 0.35 | 0.43 | 0.30 |
| Vokenization | 0.60 | 0.39 | 0.35 | 0.45 | 0.29 |
| similar | | | | | |
| BERT | 0.62 | 0.45 | 0.41 | 0.44 | 0.33 |
| ViLBERT | 0.50 | 0.43 | 0.38 | 0.47 | 0.31 |
| Vokenization | 0.73 | 0.48 | 0.38 | 0.46 | 0.29 |
| dissimilar | | | | | |
| BERT | 0.46 | 0.43 | 0.39 | 0.42 | 0.32 |
| ViLBERT | 0.50 | 0.39 | 0.33 | 0.44 | 0.33 |
| Vokenization | 0.54 | 0.41 | 0.36 | 0.48 | 0.31 |

Table 7: Proportion of aligned cases between humans and the models when considering all pairs in the benchmarks (all), their *highest* + *high* partition (similar), and *lowest* + *low* partition (dissimilar).

**creator**


"In closing, our **creator** [male] speaks to the audience." (VIST)

**maker**



"A kitchen counter with coffee **maker**, bread **maker** and a candy dish visible." (COCO)

**bakery**

"A display in a **bakery** filled with lots of cakes and pies." (COCO)

**zebra**

"A couple of **zebra** walking across a grass covered field." (COCO)

Figure 5: Four $\langle image, caption \rangle$ samples from our dataset (in brackets, we indicate the source: VIST/ COCO). For the high-similarity SL999 pair $creator, maker$ (top), multimodal models perform worse than BERT. An opposite pattern is observed for the low-similarity MEN pair $bakery, zebra$ (bottom). The pair $bakery, zebra$ is highly concrete, while $creator, maker$ is not.

a comparable performance when dealing with semantically similar and dissimilar pairs; that is, none of the models is biased toward one or the other extreme of the similarity scale.

Some interesting observations can be made by zooming into some specific word pairs in Table 6: For example, $creator, maker$, one of the most similar pairs in SimLex999 (a pair with low concreteness), is assigned the *highest* class by BERT; *low* and *medium* by ViLBERT and Vokenization, respectively. This suggests that adding visual information has a negative impact on the representation of these words. As shown in Figure 5 (top), this could be due to the (visual) specialization of these two words in our dataset, where *creator* appears to be usually used to refer to a human agent, while *maker* typically refers to some machinery. This confirms that multimodal models effectively leverage visual information, which leads to rather dissimilar representations. Another interesting case is $bakery, zebra$, one of MEN's least similar pairs (and highly concrete), which is assigned to *low* and *lowest* by ViLBERT and Vokenization, respectively, to *medium* by BERT. In this case, adding visual information has a positive role toward moving one representation away from the other, which is in line with human intuitions. As for the relatively high similarity assigned by BERT to this pair, a manual inspection of the dataset reveals the presence of samples where the word *zebra* appears in bakery contexts; for example, ''There is a decorated cake with zebra and giraffe print'' or ''A zebra and giraffe themed cake sits on a silver plate''. We conjecture these co-occurrence patterns may play a role in the *non-grounded* representations of these words.

| | RG65 | WS353 | SL999 | MEN | SVERB |
|---|---|---|---|---|---|
| all | 0.31 | 0.20 | 0.18 | 0.19 | 0.13 |
| none | 0.22 | 0.43 | 0.44 | 0.32 | 0.51 |
| B only | 0.08 | 0.06 | 0.10 | 0.08 | 0.08 |
| MM only | 0.08 | 0.05 | 0.05 | 0.09 | 0.05 |

Table 8: Proportion of pairs where *all* / *none* of the 3 models or only either BERT (*B only*) or multimodal (*MM only*) models are aligned to humans.

To provide a more quantitative analysis of contrasts across models, we compute the proportion of word pairs in each benchmark for which *all* / *none* of the 3 models assign the target similarity class; BERT assigns the target class, but neither of the multimodal models do (*B only*); both multimodal models are correct but BERT is not (*MM only*). We report the numbers in Table 8. It can be noted that, in MEN, the proportion of *MM only* cases is higher compared to *B only*; that is, visual information helps more than harms in this benchmark. An opposite pattern is observed for, as an example, SimLex999.

## 7 Conclusion

Language is grounded in the world. Thus, *a priori*, representations extracted from multimodal data should better account for the meaning of words. We investigated the representations obtained by Transformer-based pre-trained multimodal models —which are claimed to be general-purpose semantic representations—and performed a systematic *intrinsic* evaluation of how the semantic spaces learned by these models correlate with human semantic intuitions. Though with some limitations (see Faruqui et al., 2016; Collell Talleda and

Moens, 2016), this evaluation is simple and interpretable, and provides a more direct way to assess the representational power of these models compared to evaluations based on task performance (Tan and Bansal, 2020; Ma et al., 2021). Moreover, it allows to probe these models on a purely *semantic* level, which can help answer important theoretical questions regarding how they build and represent word meanings, and how these mechanisms compare to previous methods (see Mickus et al., 2020, for a similar discussion).

We proposed an experimental setup that makes evaluation of various models comparable while maximizing coverage of human judgments data. All the multimodal models we tested—LXMERT, UNITER, ViLBERT, VisualBERT, and Vokenization—show higher correlations with human judgments than language-only BERT for more concrete words. These results confirm the effectiveness of Transformer-based models in aligning language and vision. Among these, Vokenization exhibits the most robust results overall. This suggests that the token-level approach to visual supervision used by this model in pre-training may lead to more fine-grained alignment between modalities. In contrast, the sentence-level regime of the other models may contribute to more uncertainty and less well defined multimodal word representations. Further work is needed to better understand the relation between these different methods.

## Acknowledgments

We kindly thank Emanuele Bugliarello for the advice and indications he gave us to use the VOLTA framework. We are grateful to the anonymous TACL reviewers and to the Action Editor Jing Jiang for the valuable comments and feedback. They helped us significantly to broaden the analysis and improve the clarity of the manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 819455).

## References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433. https://doi.org/10.1109/ICCV.2015.279

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13. https://doi.org/10.1111/lnc3.12170

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639, PubMed: 17705682

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339. Association for Computational Linguistics.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47. https://doi.org/10.1613/jair.4135

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1219–1228. https://doi.org/10.1145/2393347.2396422

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English

word lemmas. *Behavior Research Methods*, 46(3):904–911. https://doi.org/10.3758/s13428-013-0403-5, PubMed: 24142837

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00408

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer. https://doi.org/10.1007/978-3-030-58539-6_34

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer. https://doi.org/10.1007/978-3-030-58577-8_7

Guillem Collell Talleda 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 *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2807–2817. ACL.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Christopher Davis, Luana Bulat, Anita Lilla Verő, and Ekaterina Shutova. 2019. Deconstructing multimodality: Visual properties and visual context in human semantic processing. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 118–124.

Manuel De Vega, Arthur Glenberg, and Arthur Graesser. 2012. *Symbols and Embodiment: Debates on Meaning and Cognition*, Oxford University Press.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512. https://doi.org/10.1109/CVPR.2017.475

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35. https://doi.org/10.18653/v1/W16-2506

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131. https://doi.org/10.1145/503104.503110

John R. Firth. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182. https://doi.org/10.18653/v1/D16-1235

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162. https://doi.org/10.1080/00437956.1954.11659520

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal Transformers. *Transactions of the Association for*

*Computational Linguistics*. https://doi.org/10.1162/tacl_a_00385

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language Transformers for verb understanding. *arXiv preprint arXiv: 2106.09141*. https://doi.org/10.18653/v1/2021.findings-acl.318

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. https://doi.org/10.3115/v1/D14-1032

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. https://doi.org/10.1162/COLI_a_00237

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and et al.. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239. https://doi.org/10.18653/v1/N16-1147

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2021. Probing contextual language models for common ground with visual representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5367–5377, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.422

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. https://doi.org/10.3115/v1/D14-1005

Douwe Kiela, Anita Lilla Verő, and Stephen Clark. 2016. Comparing data sources and architectures for deep visual representation learning in semantics. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 447–456. https://doi.org/10.18653/v1/D16-1043

Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4994. https://doi.org/10.1109/CVPR.2016.539

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211. https://doi.org/10.1037/0033-295X.104.2.211

Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, 85:645–665. https://doi.org/10.1016/j.engappai.2019.07.010

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163. https://doi.org/10.3115/v1/N15-1016

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275. https://doi.org/10.18653/v1/2020.acl-main.469

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. https://doi.org/10.1007/978-3-319-10602-1_48

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Timo Lüddecke, Alejandro Agostini, Michael Fauth, Minija Tamosiunaite, and Florentin Wörgötter. 2019. Distributional semantics of objects in visual scenes in comparison to text. *Artificial Intelligence*, 274:44–65. https://doi.org/10.1016/j.artint.2018.12.009

Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, and Timothy Baldwin. 2021. On the (in)effectiveness of images for text classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 42–48.

Lotte Meteyard, Sara Rodriguez Cuadrado, Bahador Bahrami, and Gabriella Vigliocco. 2012. Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7):788–804. https://doi.org/10.1016/j.cortex.2010.11.002, PubMed: 21163473

Timothee Mickus, Mathieu Constant, Denis Paperno, and Kees van Deemter. 2020. What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. *Proceedings of the Society for Computation in Linguistics*, volume 3.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*.

Roberto Navigli and Federico Martelli. 2019. An overview of word and sense similarity. *Natural Language Engineering*, 25(6):693–714. https://doi.org/10.1017/S1351324919000305

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 'Beyond Language: Multimodal Semantic Representations' Workshop*.

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. https://doi.org/10.3115/v1/D14-1162

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202

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. https://doi.org/10.1162/tacl_a_00349

Armand S. Rotaru and Gabriella Vigliocco. 2020. Constructing semantic models from words, images, and emojis. *Cognitive Science*, 44(4):e12830. https://doi.org/10.1111/cogs.12830, PubMed: 32237093

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633. https://doi.org/10.1145/365628.365657

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565. https://doi.org/10.18653/v1/P18-1238

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. https://doi.org/10.3115/v1/P14-1068

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Mohamed Ali Hadj Taieb, Torsten Zesch, and Mohamed Ben Aouicha. 2020. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, 53(6):4407–4448. https://doi.org/10.1007/s10462-019-09796-3

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1514

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080. https://doi.org/10.18653/v1/2020.emnlp-main.162

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188. https://doi.org/10.1613/jair.2934

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. https://doi.org/10.18653/v1/W18-5446

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018. Learning multimodal word representation via dynamic fusion methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Matthijs Westera and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 120–133. https://doi.org/10.18653/v1/W19-0410

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.