Images Supporting Language Models: Selected Models

The following table contains a summary of the selected language models augmented with visual components.
For each model, the following information are reported:
- Pure language model and pretraining data
- Visual features and pretraining data
- Fusion strategy of the two modalities
- Benchmarks/baselines for evaluation:
▲Better performance over baseline(s)
● Mixed performance results over baseline(s)
▼Worse performance over baseline(s)

Year	Paper	Language model (LM)	LM-Pre-training sources	Visual elements (IMG)	IMG-Pre-training sources	Multimodal representation and model description	Testset/Fine-tuning	Baseline(s)/model settings/comparison to other models	Results
2014	Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. "Multimodal distributional semantics." Journal of artificial intelligence research 49 (2014): 1-47.	Distributional model expressed as a matrix with rows as "semantic vectors" representing the meaning of a set of target words. The model is based on co-occurrence counts of words (as a result, the matrix is a squared one).	- ukWaC, 1.9B tokens - Wackypedia, 820M tokens.	(i) Local descriptors to extract low-level visual features (ii) Assign local descriptors to cluster of visual words (bag of words) to build the vector representation of an image (iii) Sum up visual words co-occurrence to across all images/instances to get co-occurrence counts related to a target word (the resulting matrix is a squared one).	ESP-Game dataset, 100K images.	Only words for which there is a related image are considered. Two steps to build multimodal representations: (i) Textual and visual matrices are concatenated and projected into a common latent multimodal space with a singular value decomposition. From this matrix, the "textual **mixed** matrix" and the "visual **mixed** matrix" are extracted (ii) Association between words is assessed with cosine similarity Two fusion methods to estimate similarity of pairs: - Feature level fusion: linear combination of textual and visual mixed matrix and then similarity estimation - Scoring level fusion: word similarity computed on both textual and visual mixed matrices separately and then the final score is a linear combination of the two In both methods the weights in the linear combinations are hyperparameter.	- WordSim353 - MEN.	- Text mixed embeddings only - Visual mixed embeddings only - Equally weighted versions of feature and scoring level fusion model settings - Several "fine tuned" versions of fusion and scoring level fusion model settings.	▲Multimodal word representations enhance performance of purely textual or visual embeddings ● No alternative model used as a means of comparison.
2014	Hill, Felix, and Anna Korhonen. "Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.	Skipgram.	400m word Text8 Corpus.	Mapping of words w to a bag of perceptual features b(w), extracted from external sources and encoded in an associative array P. Generation of pseudo sentences based on these perceptual features to be fed into the language model.	- ESP-Game (100K images) - CSLB Property Norms.	Extension of the Skipgram injecting perceptual information by generating pseudo-sentences based on a bag-of-visual-words. A hyperparameter α controls the level of perceptual information relative to linguistic input.	USF Dataset.	- Concatenation of linguistic and perceptual features - Canonical Correlation Analysis applied on vectors of both modalities - SVD of matrix of concatenated multimodal representations.	▲Concepts, which can **directly** be represented in the perceptual modality (e.g. concrete verbs and nouns) ▲Propagation of perceptual input from concrete concepts (nouns and verbs) to enhance the representation of abstract verbs, those for which no direct representation in the visual space is available ▼Abstract nouns (for which is more difficult to find a concrete visual representation) are still more efficiently learned from language-only models.
2014	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, Doha, Qatar. Association for Computational Linguistics.	Skipgram.	- Text8 Corpus (400M words) - British National Corpus (100M words).	Seventh layer of a CNN to extract 6144-d features vectors for images, obtained in two ways: - CNN-Mean (average of all features vectors representing images) - CNN-Max (component-wise maximum of all features vectors)	- ImageNet (12.5M images) - Esp-Game (100K images).	Concatenation of visual and textual embeddings.	- MEN - WordSim353 (it captures not only "relatedness" but also "similarity".	- Skipgram (text-only baseline) - Embeddings - visual only.	▲CNN-Mean better on MEN: averaging might capture relatedness better. CNN-Max better on WordSim353.
2014	Silberer, Carina, and Mirella Lapata. "Learning grounded meaning representations with autoencoders." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.	Vectors of textual attributes are extracted.	- McRae et al.'s (2005).	Vectors of visual attributes are extracted.	Same dataset as in Silberer et al. (2013): taxonomy of 636 visual attributes (e.g., has wings, made of wood) and nearly 700K images from ImageNet (Deng et al., 2009) describing more than 500 of McRae et al.'s (2005) nouns.	Stacked (denoising) autoencoders for each single modality and the outputs are concatenated and fed to a stacked bimodal autoencoder which map the inputs to a joint hidden layer.	With McRae et al.'s (2005), two tasks: - word similarity - word categorization.	- Unimodal autoencoders only - Kernelized Canonical Correlation, Hardoon et al. (2004) - Bruni et al. (2014).	▲Bimodal models outperform unimodal ones ▼Training is on attribute-based inputs. Not widely used in the field.
2015	Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. "Combining language and vision with a multimodal skip-gram model." arXiv preprint arXiv:1501.02598 (2015).	Skipgram.	Wikipedia 2009, 800M Tokens.	Visual information for 5100 words with an entry in ImageNet, occur >500 times in the text corpus and have a concreteness score ≥0.5; sample 100 images for each word and extract a 4096-d array with a CNN; average the vectors of 100 pictures associated to each word to get visual representation.	ImageNet.	The objective function is a linear composition of the language objective L-ling from the Skipgram and a visual objective L-vision. For the L-vision objective two variants are proposed: - MM Skipgram A (MMSA): aligning vectors of visual and linguistic representations (1:1 correspondence assumed) - MM Skipgram B (MMSB): estimate a cross-modal mapping matrix from linguistic onto visual representations.	- MEN - SemSim - VisSim.	- Kiela and Bottou (2014) - Bruni et al. (2014) - Silberer & Lapata (2014) - Skipgram (text-only baseline) - Embeddings - visual only - Concatenation - SVD.	▲Both MMSA and MMSB better than simpler models (linguistic/vision only, concatenation SVD) ● MMSA and MMSB competitive in relatedness and visual similarity, despite having often less training data than other models ● Visual grounding less effective with abstract words
2017	Collell, Guillem, Ted Zhang, and Marie-Francine Moens. "Imagined visual representations as multimodal embeddings." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1. 2017.	300-d GloVe.	Common Crawl corpus, 840B tokens, 2.2M words.	To extract visual features, the last hidden layer of a CNN is taken. For each concept, two different ways to combine the extracted visual features: - Averaging (averaging of al features vectors) - Maxpooling (component-wise maximum).	Imagenet.	Mapping from language to vision. No need of 1:1 correspondence between linguistic and visual inputs. Two different mappings are considered: - Linear (MAP-Clin) - Neural Network (MAP-Cnn).	- MEN - WordSim353 - SemSim - Simlex999 -SimVerb3500 - VisSim.	- Kiela and Bottou (2014) - Lazaridou et al. (2015) - Silberer & Lapata (2014) - GloVe (text-only baseline) - Concatenation.	▲Outperformance in all instances where words have associated images in the training set ▼Performance on the zero-shot learning still inferior in many instances to the textual baselines.

2018	Kiela, Douwe, et al. "Learning visually grounded sentence representations." arXiv preprint arXiv:1707.06320 (2017).	- GloVe for word embeddings - Bidirectional LSTM for sentence representation.	WebCrawl	Image features obtained from the final layer of a ResNet-101.	MS COCO.	Word embeddings are projected to a ground space with a linear mapping. Linear mapping and Bi-LSTM are trained jointly. Three methods to ground sentences in images, captions or both: - Cap2Img: predict latent features of an image from its caption by mapping the (final) hidden state h(T) of the Bi-LSTM to the latent representation of the image. A ranking loss is to be minimized - Cap2Cap: given the caption pair (x,y) describing the same image, the goal is to maximize the probability of y given x. Negative log-likelihood as loss. - Cap2Both: Goal is to minimize the two loss functions above. In another setting, grounded and sentence-only (Skipthought) representations are concatenated with layer normalization to get the final sentence representations. Goal is to include information on less concrete concepts which are not likely to be represented in image-captioning databases but are present in language corpora.	Intrinsic evaluation of word embeddings: - MEN - SimLex 999 - Rare Words - WordSim-353 Extrinsic evaluations: - Movie review Sentiment (MR) - Product reviews (CR) - Subjectivity classification (SUBJ) - Opinion polarity (MPQA) - Paraphrase identification (MSRP) - Sentiment classification (SST) - SNLI (Entailment) - SICK (Entailment)	- Skipthought (text-only baseline).	▲Word embeddings are of higher quality than those obtained with GloVe, measured on the following similarity benchmarks: MEN, SimLex999, Rare Words and WordSim-353 ▲In extrinsic evaluations, grounding increases performance but it is not clear which one of the three grounding strategies considered is dominant ● Performance seems to be driven in a smaller amount of instances by a larger number of parameters rather than effectiveness of grounding ● Performance is better when dataset have a higher level of concreteness.
2020	Bordes, Patrick, et al. "Incorporating visual semantics into sentence representations within a grounded space." arXiv preprint arXiv:2002.02734 (2020).	- Skipthought.	Toronto Book Corpus: 11M books, 74M ordered sentences, 13 words per sentence on average.	Processing of visual elements with a pre-trained Inception v3 network (Szegedy et al., 2016).	MS COCO: 118K/5K/41K (train/val/test) images.	The objective function is composed of: - a textual objective Lt - a grounding objective Lg, which among its parameters has also those of the textual objective, which in turn profit from both objective functions. Lg is not applied directly on the sentence embeddings; it is trained on an intermediate space called the "grounded space". The sentence embeddings are projected to the grounded space with the projection function being a multi-layer perceptron. The goal is to move away from the 1:1 correspondence between textual and visual space. Lg the can be decomposed in two components, whose individual contribution is controlled by two hyperparameters: - Cluster Information (Cg): sentences associated with the same image(s) should be similar. The visual space is thus used to asses sentence similarity. The Max-margin ranking loss is used - Perceptual information (Pg): similarity between sentences in the grounded space should be correlated with similarity between corresponding images in the visual space. The loss is based on the negative Pearson correlation. Model scenarios include many compositions of the above mentioned elements.	Intrinsic evaluation of word embeddings: - STS - SICK Extrinsic evaluations: - Movie review Sentiment (MR) - Product reviews (CR) - Subjectivity classification (SUBJ) - Opinion polarity (MPQA) - Paraphrase identification (MSRP) - Sentiment classification (SST) - SNLI (Entailment) - SICK (Entailment).	- Skipthought (text-only baseline) For extrinsic evaluations: - Kiros et al. (2014) - Kiela et al. (2018) - Lazaridou et al. (2015) - cross-modal - Collell et al. (2017) - sequential/concatenation.	▲Word embeddings are better than the textual benchmark for data with a high level of concreteness and are similar in performance with respect to more abstract concepts ▲Projections on the grounded space are more effective than cross-modal projection and concatenation ● Not always best performance on entailment tasks (benchmarks SNLI, SICK).
2020	Tan, Hao, and Mohit Bansal. "Vokenization: Improving language understanding with contextualized, visual-grounded supervision." arXiv preprint arXiv:2010.06775 (2020).	BERT, but it can be adapted to any language model (through Revokenization).	English Wikipedia.	ResNeXt.	MS COCO.	Language model with visual supervision. Each token in a sentence obtains a corresponding image (voken) assigned from a finite set of images. The voken is the image which maximize a Relevance Score Function between a token and all images in the aforementioned finite set of images. With this token-voken pairs a voken classification pre-training task is performed that can be built in pure language models alongside other pre-training tasks such MLM or Next-Sentence Prediction.	- GLUE (only SST-2, QNLI, QQP, MNLI) - SQuAD - SWAG.	- BERT (various versions) - VilBert - VL-BERT - VisualBERT - Oscar - LXMERT.	▲ Improvement over the purely self-supervised language model on multiple language tasks.
2021	Hu, Ronghang, and Amanpreet Singh. "Unit: Multimodal multitask learning with a unified transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.	BERT-base with a learned task-specific vector (to capture task-specific information) as additional input, which is positioned at the beginning of the embedded token sequence.	Pretrained version of the embeddings	CNN (ResNet-50) to extract visual features map + transformer encoder to encode the features map to a set of hidden states. A learned task-task specific vector (to capture task-specific information) is concatenated to the beginning of the visual feature list before entering the encoder (architecture inspired by DETR).	- MS COCO - Visual Genome.	To both modalities is then applied a domain agnostic transformer architecture. As input the transformer takes the hidden states of either language or visual encoders or concatenation of both together with a task specific query embedding sequence. Self attention is applied in each layer among decoder hidden states and cross attention is applied to the encoded input modalities. The output is a set of decoded hidden states to which a task-specific head is applied (two-layer MPLP with GeLU activation and cross entropy loss). Training is done jointly on multiple tasks. At each training iteration, a task is randomly selected. Three settings (third one is the model described above): 1.) Single-task training: each model is trained separately on each task 2.) Multi-task training with separate decoders: a specific decoder for each task and jointly trained on all tasks 3.) Multi-task training with shared decoder. In this setting, there are still task-specific heads for each task.	Extrinsic evaluation.: GLUE: - QNLI - QQP - MNLI - SST2.	BERT (text-only baseline).	▲Model setting (1), single task training, outperforms all other settings and is comparable to the text-only baseline ▼Model setting (3), domain-agnostic, multi-task training with shared decoder across modalities exhibits a lower performance compared to domain-specific transformer models like BERT, the text-only baseline.
2021	Shahmohammadi, Hassan, Hendrik Lensch, and R. Harald Baayen. "Learning zero-shot multifaceted visually grounded word embeddings via multi-task training." arXiv preprint arXiv:2104.07500 (2021).	- GloVe, 300d, 2.2M words - fastText, 300d, 2M.	Pretrained version of the embeddings	Image vectors obtained by transferring the penultimate layer of pretrained Inception-V3 trained on ImageNet. A neural network with one hidden layer and tanh activation is used to project the image vectors into the initial hidden state of the GRUs employed in the model.	MS COCO.	Given embeddings originating from a pretrained text-only model, the goal is to generate a mapping matrix M to ground word embeddings visually (the mapping matrix is used in both directions, to map text to grounded space and to map grounded embeddings back to the textual space) This is obtained by performing three different tasks: (i) Next word prediction with a GRU, given previous words in the sentence provided as image caption, together with the related image embedding vector (ii) Same as (i) but the sentence is provided backwards to another GRU (iii) Binary classification task if the representation of a given sentence in the grounded space obtained from (i) and (ii) matched the associated image.	Limited to intrinsic evaluation: - MEN - SimLex999 - Rare Words - MTurk771 - WordSim353 - SimVerb3500.	- GloVe (text-only baseline) - fastText (text-only baseline) - Collell et al. (2017) - Park & Myaeng (2017) - Kiros et al. (2018) - Kiela et al. (2018).	▲Textual baselines and related models are outperformed and the model seems to improve the textual vector space by aligning it with real-world relations from the images (similarity appears to be favoured by the model over relatedness) ▲Embeddings related to less concrete words exhibit good quality compared to baselines.
2022	Hsu, Chan-Jan, Hung-yi Lee, and Yu Tsao. "XDBERT: Distilling Visual Information to BERT from Cross-Modal Systems to Improve Language Understanding." arXiv preprint arXiv:2204.07316 (2022).	- BERT - ELECTRA.	Wikipedia.	CLIP as image-text matching system: two components, a text encoder (CLIP-T) and an image encoder (CLIP-ViT).	not specified.	3 adaptive tasks: - Joint Masked Language Modelling (MLM) - Same Sentence Prediction (MATCH) - CLIP Token Classification After that, concatenation with cross-modal encoder is performed.	- GLUE - SWAG - READ.	- BERT - ELECTRA.	▲Better performance than pure language models, in particular in smaller datasets, which suggests that visual inputs improve generalization when the amount of training data is limited.

2022	Lu, Yujie, et al. "Imagination-Augmented Natural Language Understanding." arXiv preprint arXiv:2204.08535 (2022).	- BERT-base - RoBERTa.	Wikipedia.	Same as in VOKENIZATION paper.	MS COCO.	The framework iACE is composed of two modules: 1.) Imagination generator G: for each text input, VQGAN generates an "imagined visual" and CLIP is used to test how the generated image corresponds to the text and encodes the text and the image in a cross-modal embedding space and the objective function Lgan is to minimize the distance between these two embeddings 2.) Imagination augmented cross-modal encoder. Specifically, CLIP is used, with the embeddings from the textual and visual encoder (fed with the visualized imaginations) within CLIP are then "late fused". The output is a set of "imagination-augmented" language representations. Learning procedure: 1.) pre-training of a visually-supervised transformer following the Vokenization method 2.) Imagination-augmented fine tuning, composed of two losses to be minimized: (i) L-imagine where the iACE framework tries to minimize the cross-entropy loss based on text embeddings and visually imagination embeddings, given a testsets and a number of classes to predict (ii)L-lang, where the visually supervised transformer only relies on textual inputs The "imagination-augmented" is a composition of (i) and (ii) and the relative contribution of each loss is controlled with an hyperparameter.	From GLUE and SWAG: - SST-2 - QNLI - QQP - MultiNLI - MRPC - STS-B Focus is on few-shots learning (considering from 0.1% to 5% of the training dataset).	- BERT (text-only baseline) - RoBERTa (text-only baseline) With and w/o Vokenization.	▲Better performance of iACE over visually supervised transformers (VOKEN) in **all** instances of few-shots learning. Imagination can help existing language models to perform better in a setting with small training set (which means "less human annotated data").