# Visual Grounding Strategies for Text-Only Natural Language Processing

Damien Sileo

Department of Computer Science (CS)
KU Leuven
damien.sileo@kuleuven.be

## Abstract

Visual grounding is a promising path toward more robust and accurate Natural Language Processing (NLP) models. Many multimodal extensions of BERT (e.g., VideoBERT, LXMERT, VL-BERT) allow a joint modeling of texts and images that lead to state-of-the-art results on multimodal tasks such as Visual Question Answering. Here, we leverage multimodal modeling for purely textual tasks (language modeling and classification) with the expectation that the multimodal pretraining provides a grounding that can improve text processing accuracy. We propose possible strategies in this respect. A first type of strategy, referred to as *transferred grounding* consists in applying multimodal models to text-only tasks using a placeholder to replace image input. The second one, which we call *associative grounding*, harnesses image retrieval to match texts with related images during both pretraining and text-only downstream tasks. We draw further distinctions into both strategies and then compare them according to their impact on language modeling and commonsense-related downstream tasks, showing improvement over text-only baselines.

## 1 Introduction

Representation of text with transferable encoding is a central task of artificial intelligence. Transfer from larger and larger transformer-based text encoders trained on masked language modeling has become a standard way to achieve state-of-the-art results. Progress is shown in tasks such as natural language inference and semantic similarity estimation when evaluated on natural language understanding benchmark datasets (Kaplan et al., 2020) such as GLUE and SuperGLUE (Wang et al., 2019). However, these scores do not tell the whole story. Firstly, marginally better benchmark scores can come at the price of impractical GPU requirements. Secondly, super-human scores can be obtained by exploiting spurious dataset-specific correlations instead of more generalizable reasoning (Niven and Kao, 2019). Mastering commonsense reasoning is regarded as a requirement for "true" language understanding (Bisk et al., 2020), and grounding representations of natural language on other modalities such a visual perception is a privileged strategy in that endeavor. Since the meaning of language stems from the physical world, visual grounding[1] is a valuable way to guide the training of NLP models. Thus, we hypothesize that text encoders, even already pretrained on a massive amount of text-only data, can be improved by a further multimodal pretraining stage.

Many visuolinguistic transformer architectures such as LXMERT (Tan and Bansal, 2019) VilBERT (Lu et al., 2019) or VL-BERT (Su et al., 2019) have been proposed to augment BERT (Devlin et al., 2019) with joint text-image understanding in order to tackle multimodal tasks including visual question answering and image retrieval, each leading to substantial performance improvement over the previous state of the art. To that end, these models generalize BERT's Masked Language Modeling (MLM) objective to a multimodal setting. More specifically, they perform the MLM task on image captions while allowing the model to use features of the paired image[2], and also perform image modeling while letting the model attend to textual features. Nevertheless, it is not clear whether joint pretraining is better than text-only pretraining when transferring to text-only tasks (e.g., language modeling, text classification, similarity estimation) and how it should be done. In this work, we address this question by breaking down the strategies toward

[1] Here, we define visual grounding as learning language representations from explicit visual associations.

[2] An image is represented as a sequence of regions features extracted with a Convolutional Neural Network.



*Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 19–29
Kyiv, Ukraine (Online), April 20, 2021

that endeavor.

A first proposed way to leverage visuolinguistic models for textual tasks is to perform the aforementioned joint modeling pretraining and then finetuning on text-only tasks while using a placeholder as image input. However, this only trains the text encoder on the caption domain (which is different from the wider domain usually used to train text encoders) and the model is not exposed to any training example without images. To alleviate this problem, it is possible to combine joint masked modeling on captioned images with unimodal MLM on a broader domain text-only corpus during pretraining by associating an image placeholder with the text-only examples. We call this strategy *transferred grounding*.

We also propose another technique named *associative* grounding where for each textual input, an association module retrieves the most relevant images from a large external images collection. The text encoder parameters can then harness these externally provided images for better text understanding instead of having to internally model them.

Since we are bringing an additional computational cost by using images and not only pure text, we will strive to reduce the incurred additional memory usage in our models.

The goal of this paper is to identify and compare techniques that bridge visuolinguistic multimodal modeling with text-only tasks. To that aim, our contributions are the following: (i) 4 visuolinguistic grounding strategies for text-only tasks transfer; (ii) An evaluation of the size/accuracy trade-off of image representation to reduce the memory usage of multimodal transformers; and (iii) A systematic comparison of the proposed strategies for transformers on English text-only tasks.

## 2 Grounding strategies

In this section we present the two main grounding strategies, that is *transferred grounding* and *associative grounding*, of which Figure 1 shows an overview. We first describe the multimodal model architecture used in both grounding strategies.

### 2.1 Multimodal model structure

We rely on a setup shared by several multimodal extensions of BERT (Tan and Bansal, 2019; Su et al., 2019; Lu et al., 2019). The text encoding takes place in two stages. The output of the text unimodal encoder (e.g., BERT) provides an intermediate representation of the text input. The visual input has a vector representation (typically CNN-extracted embedding possibly contextualized with a transformer model). Then, both are fed to a cross-modal encoder that learns to compose the modalities in order to perform visuolinguistic modeling tasks. This multimodal model structure is illustrated in both Figures 2 and 3.

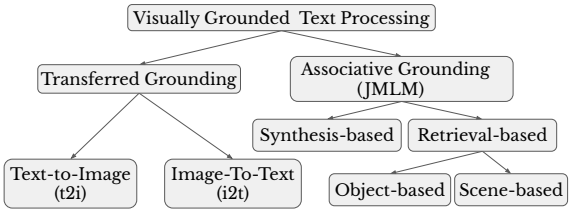



Figure 1: A categorization of the proposed grounding strategies.

This architecture could directly be used in order to perform downstream tasks. However, a multimodal pretraining, for instance through masked modeling, is key to achieve high performance (Tan and Bansal, 2019).

### 2.2 Masked modeling of multimodal features

In our grounding strategies we rely on masking. In Joint Masked Language Modeling (JMLM), random tokens of the input text $t$ are replaced by a `[masked]` token. A softmax JMLM head on top of the cross-modal encoder has to predict the original value of the masked tokens through cross-entropy (H) loss minimization. The cross modal encoder is thus incentivized to use the visual input regions $r$ when they are relevant for the masked tokens prediction. Joint Masked Region Modeling (JMRM) is the visual counterpart of JMLM. Here, image region features are masked and have to be predicted, for example, through a $\mathrm{L}_p$ loss minimization, based on both the text and non-masked visual cues. The corresponding losses are the following:

$$\mathcal{L}_{\text{JMLM}} = \sum_{k=1}^{|t|} m_k \mathrm{H}(t_k, \hat{t}_k(\bar{t}, r))$$

$$\mathcal{L}_{\text{JMRM}} = \sum_{k=1}^{|r|} m_k \mathrm{L}_p(r_k, \hat{r_k}(t, \bar{r}))$$

where $m_k = 1$ when a token/region is masked, $\bar{t}/\bar{r}$ represents the masked tokens/regions sequence, $\hat{t}$ the predicted probability distributions, and $\hat{r}$ the predicted region features.

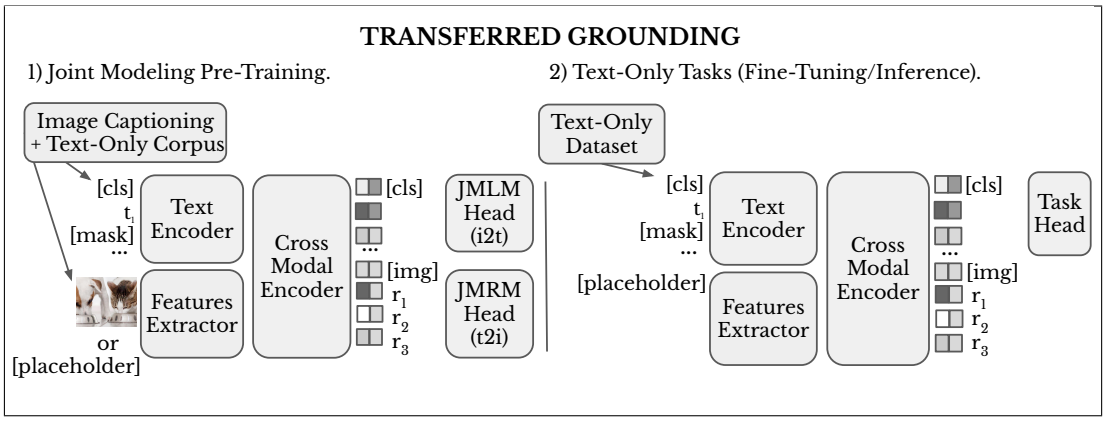



Figure 2: Overview of the Transferred grounding strategy.

### 2.3 Transferred grounding

In *transferred grounding* the masked modeling of multimodal features is used in pretraining as shown in Figure 2.

The corpus used for pretraining contains images paired with captions, which can also be extended with another text-only corpus. Adding examples from a text-only corpus allows the encoder to be exposed to a wider domain of text during pretraining. Since no image is available for these examples, we replace the image features by a single trainable embedding $h_{\text{[placeholder]}} \in \mathbb{R}^d$ where $d$ is the dimension of the cross-modal encoder inputs.

Cross-modal prediction can occur in the two following directions, each providing a different way to ground language understanding.

#### 2.3.1 Transfer from Text-To-Image prediction (t2i)

Pretraining a model to perform JMRM incentivizes the text representations to abstract the visual knowledge involved in visual region modeling. This provides a form of grounding that could be useful for textual downstream tasks. Here the model learns to predict masked image aspects from text, which might help language understanding by visually imagining the language content.

#### 2.3.2 Transfer from Image-To-Text prediction (i2t)

A different way to perform pretraining is to perform JMLM in the presence of visual input. The model can learn to use visual information for textual modeling, thus developing useful abstractions. We hypothesize that these abstractions can help text-only tasks when the visual input they relied on is missing.

#### 2.3.3 Text-only tasks

Once pretrained, the multimodal architecture can be applied to text-only tasks. To do so, the placeholder features $h_{\text{[placeholder]}}$ are used to replace the missing visual input.

### 2.4 Associative Grounding

In an alternative grounding strategy, called *associative grounding*, we do not rely on a text corpus that is a priori paired with images, but the pairing is part of the model. As seen in Figure 3, for each text input[3] $t$ (e.g., sentence or paragraph), a visual association module retrieves the most relevant images from an archive of images.

Here, we instantiate the visual association module in the following way. A query encoder provides representation $q_t$ for the masked input text $t$ using the cross-modal encoder described above, and a relevance metric R($q_t$,$k_i$) such as the cosine similarity identifies the $K$ most relevant images in a collection $M$ of images $i$, which are encoded into key vector representations $\{k_i, i \in M\}$.

Associated images can be visual scenes that match the input text as a whole. However, the space of situations matching the content of the input text is immense, and in many cases, the most relevant images in $M$ will only be loosely associated to the textual input. Decomposing the situation evoked in the text into objects that play an important role can help narrowing down the space of possible relevant images, thus leading to closer associations, albeit more partial. We call the first substrategy *scene-based* and the second one *object-based*.

[3] During JMLM pretraining, the masked text is used to find associations. This to prevent trivial prediction of the masked text tokens from their presence in the images.

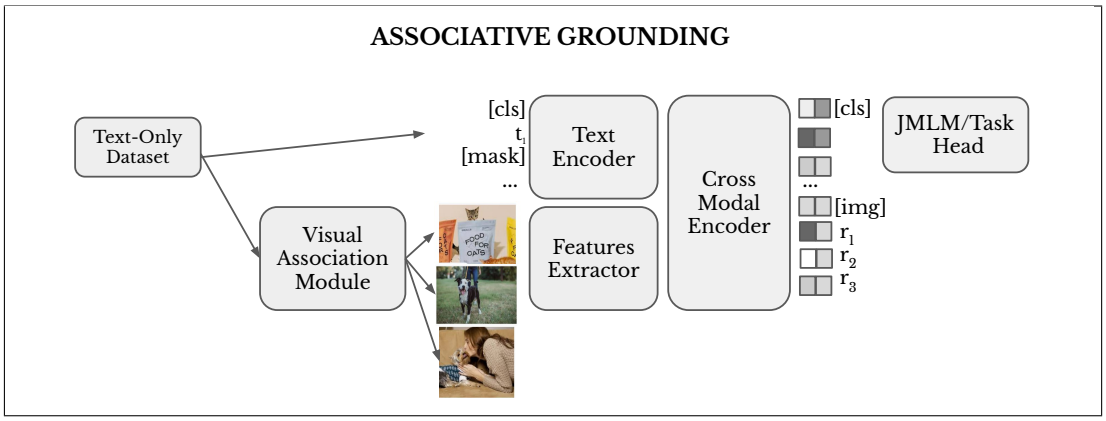



Figure 3: Overview of the Associative grounding strategy.

#### 2.4.1 Scene-based association

Here we propose to use textual captions to index images.Each input text $t$ is then associated with the $K$ most similar images, according to the relevance computation of a caption given the input text. Of course, retrieved images may not be sufficiently relevant, and it is not clear whether some sentences can be illustrated by existing images. But we hypothesize that the cross-modal encoder of the captions is able to ignore non-relevant information. Kiela et al. (2014) have shown that image retrieval systems map abstract words to high dispersion results, that is, retrieved images are different to each other. This dispersion could be informative for the cross-modal encoder. Besides, concrete images could also ground more abstract concepts (Lakoff and Johnson, 1980); a very simple illustration of this is that images captioned with references to *sadness* tend to be darker than those containing the word *happiness*.

#### 2.4.2 Object-based associations

Another possibility is to rely on objects mentioned in the text. Here, we extract all common nouns in the input text and perform a Gaussian Mixture Clustering to extract $\kappa \leq K$ clusters among the noun representations to find the key concepts. For each centroid, the closest noun is chosen as a representative. A noun-indexed database, here ImageNet22K (Deng et al., 2009), can be used to map these nouns to images.

The two association systems used in the experiments are illustrated in Figure 4 in the appendix.

#### 2.4.3 Image synthesis

It would also be possible to synthesize images that represent the input text with a dedicated model. We experimented with the *DeepAI* system[4] and as shown in Figure 3, the results suggests that more work is needed to get conclusive results. Consequently, we do not further consider image synthesis here.

## 3 Experiments

We evaluate the proposed grounding strategies in language modeling and downstream tasks.

### 3.1 Text-only corpus

Following (Devlin et al., 2019), we use a combination of Wikipedia English pages and the BookCorpus (Zhu et al., 2015) as language modeling corpus. We extract the body of pages from the Wikipedia pages with at least 20 views in the 2013 dump[5], totaling 463k documents, and we sample the same number of passages from the BookCorpus. We call this combination Wiki-BC. We select $90\%$ of the pages/passages as training data, $5\%$ for validation and $5\%$ for test.

### 3.2 Visuolinguistic corpus

For the **scene-based associations**, we populate the scenes bank $M$, with the combination of two image captioning datasets. The Stony Brook University corpus (SBU) (Ordonez et al., 2011) is composed of 860k images from the Flickr website, filtered to ensure that the caption literally describes the image. ConceptualCaptions (CC) (Sharma et al., 2018) gathers 3M images that come from web pages that also were heuristically filtered. We use a $90/5/5\%$

[4] https://deepai.org/machine-learning-model/text2img

[5] Available at https://storage.googleapis.com/lateral-datadumps/wikipedia_utf8_filtered_20pageviews.csv

train/validation/test split for SBU and the standard split for CC.

When performing **object-based associations**, $M$ is populated with ImageNet (Deng et al., 2009). We keep the $15k$ synsets that are associated to at least 10 images, and we randomly sample the images when more are available. Each image is indexed with the average of the lemmas embeddings of its synset, averaged with the embedding of the synset definition.

When performing **transferred grounding**, we concatenate the captioned images (CC-SBU) with our text-only corpus. We sample CC-SBU so that its size matches the text only corpus size (840k passages) to balance the train set. Since our end-goal is text-only processing, we do not use the captioned images in the validation and test set but only in the train set.

In order to test image representations and language modeling on literal text, we also run experiments on the COCO (Lin et al., 2014) dataset made of 400k image/captions pairs. When we do so, we discard the original images when using the associative setup, but we always keep the images in the transferred setup.

### 3.3 Setup and hyperparameters

We build upon the LXMERT (Tan and Bansal, 2019) code and experimental setup, and rely on its hyperparameters values because of the state-of-the-art results and code availability of this model. Our text encoder is an Albert model pretrained from *albert-base-v2* checkpoint (Lan et al., 2019). Its weights are kept fixed during pretraining and fine-tuned during downstream task training. We use a different copy of this Albert model as the cross-modal encoder initialization and always keep its weights trainable.

In this work, our key/query encoder is a continuous bag of words.We use fastText (Mikolov et al., 2018) embeddings[6] which are competitive with state-of-the-art models on semantic similarity tasks (Reimers and Gurevych, 2019). Queries and keys are matched according to cosine distance with the Faiss (Johnson et al., 2017) library.

Following LXMERT, each image is represented with a sequence $N$ region features. A linear projection maps the region features to the input space of the cross-modal encoder. We also provide $K$ rank embeddings which are added to the cross-modal encoder input image representation. Input text is lower-cased and token sequence length is clipped to 64. We use the Adam (Kingma and Ba, 2014) optimizer with a batch size of 32 and a learning rate of $10^{-4}$ for downstream tasks until convergence on the validation accuracy, with a maximum of 4 epochs. When we retrieve $K = 16$ images [7] with object-based associations, we use $\kappa = 8$ clusters[8]. Following (Tan and Bansal, 2019), we mask 15% of text tokens during the JMLM pretraining; and we use a linear regression head with $L_1$ regularization for JMRM.

We first compare the expressivity of the image representations in function of their size, then delve into the comparison of grounding strategies, according to JMLM pretraining and downstream tasks scores.

### 3.4 Efficient image representation

We want to be able to provide $K > 1$ images to the cross-modal encoder. However, the cross-modal encoder memory usage scales quadratically with the number of regions $K.N$, so we will evaluate to what extent $N \gg 1$ is necessary. Previous work use object detection on a single image to find 36 Regions of Interest (RoI) and represents them by ResNet-50 (He et al., 2016) features with 2048 dimensions. We run the JMLM pretraining setup with several image representations. We evaluate two representations: (1) the previously used ResNet region representations provided by (Tan and Bansal, 2019) with various values of $N$; (2) Single representation of the whole image, using EfficientNet-B1 Noisy Student (Xie et al., 2020) pre-logit max-pooled features. We finetune neither of them.

We use the COCO dataset for this evaluation because its annotations are manual and not heuristically filtered, and because the image features used by (Tan and Bansal, 2019) are publicly available. We evaluate the image representations with JMLM pretraining perplexity. Table 1 shows the results of this experiment. Overall, the presence of image features ($N > 0$) significantly improves the JMLM performance. While the standard FastR-CNN(Girshick, 2015)+ResNet with 16 objects representation achieves the best perplexity (even bet-

[6] We use the CommonCrawl version https://fasttext.cc/docs/en/english-vectors.html. and discard stop-words with NLTK.

[7] We have tested values $\{1, 4, 8, 16, 24\}$ and the lowest Wiki-BC validation perplexity for both object and scene based setups is 16.

[8] We capped this value to 8 since some short texts do not refer to many objects.

| N | Features | COCO perplexity |
|---|---|---|
| 0 | - | 8.4 |
| 1 | EfficientNet-1280 | 4.6 |
| 8 | ResNet-2048 | 4.6 |
| 16 | ResNet-2048 | **4.0** |
| 36 | ResNet-2048 | 4.2 |

Table 1: The effect of image representations on cross-modal JMLM perplexity on the COCO dataset with $K = 1$ image per example; $N$ is the number of visual regions per image.

ter than with 36 features, which can be attributed to a reduced over-fitting), while this performance breaks down with few recognized objects. The performance of EfficientNet representations using the whole image as a single region is competitive while being much lighter. Consequently, in the experiments described in the following section we represent images with EfficientNet-B1 features and $N = 1$ region per image.

## 4 Empirical comparison of grounding strategies

### 4.1 JMLM training

We run the pretraining for each strategy on the Wiki-BC corpus, and also run a separate pretraining on the COCO corpus in order to assess language modeling performance on a literal, descriptive domain. Table 2 reports the influence of the grounding strategies on test perplexities. (Zhang et al., 2020) is an image retrieval system ranking the images according to the number of text keywords they contain. It was originally used to improve Neural Machine translation (see Related Work). On the COCO dataset, using the original image paired with the text captions leads to the lowest perplexity, but surprisingly, using $K = 16$ associated object or scenes leads to a competitive perplexity. This suggests than the many distantly related images can be a good substitution to a canonically associated image, even though associating $K = 1$ object seems to be a too noisy signal.

The results on the Wiki-BC dataset reveal a similar pattern. Here, the transferred strategies cannot use images at test time and suffer from a domain shift which leads to inferior results. However, the associative strategies and $K = 16$ lead to the lowest perplexity even though the gap between ungrounded and grounded methods is small than with the COCO dataset due to abstractness of the open-domain texts. We thereby show that

| Strategy | K | Wiki-BC ppl. | COCO ppl. |
|---|---|---|---|
| No Grounding | 0 | 8.7 | 7.6 |
| Transferred i2t | 0/1 | 30.0 | **4.6** |
| Transferred t2i | 0/1 | 41.6 | - |
| Transferred t2i+i2t | 0/1 | 25.3 | - |
| Associative (Zhang et al., 2020) | 16 | 8.9 | 8.6 |
| Associative (scene-based) | 16 | **8.5** | **4.6** |
| Associative (scene-based) | 1 | 9.1 | 7.3 |
| Associative (object-based) | 1 | 9.0 | 35.6 |
| Associative (object-based) | 16 | 8.6 | 4.7 |

Table 2: JMLM test perplexity on the Wiki-BC+CC-SBU dataset and on the COCO captions. On the COCO dataset, COCO images are used at test time for the transferred i2t strategy. On the Wiki-BC dataset, the image placeholder is used at test time for the transferred strategies while associative strategies always use associated images.

even language modeling can benefit from grounding, and the masked language modeling perplexity might translate to yet further applications in text generation (Wang and Cho, 2019). We will now investigate how grounding translates to downstream classification tasks accuracy.

### 4.2 Downstream evaluation

We hypothesize that grounded models should be better equipped to perform common sense related downstream tasks. A first component of our evaluation on tasks RTE, PDTB and JOCI will test that claim. But we also expect the concreteness to have an influence on the behavior of grounded models. Thus, we also perform a more targeted evaluation on concrete-only examples (SICK) and metaphoricity classification (VUA tasks) to better interpret our results.

**Recognizing Textual Entailment (RTE)** (Dagan et al., 2006) is a Natural Language Inference (NLI) task. Its dataset gathers sentence pairs with a premise and a hypothesis. The labels describe the logical relationship between the two, that is, entailment and non-entailment.

**JHU Ordinal Common-sense Inference (JOCI)** (Zhang et al., 2017) The dataset for this task also consists of premise/hypothesis sentence pairs, but the labels are numerical scores from 1 to 5 that reflect the plausibility of the hypothesis given the premise according to human annotators, relying on their own common sense.

**Sentences Involving Compositional Knowledge (SICK)** (Kiela et al., 2018) is also a NLI dataset with entailment, neutral and contradiction classes

| Strategy | Pretraining Text | K | JOCI | PDTB | RTE | SICK | VUA | VUAN | AVG |
|---|---|---|---|---|---|---|---|---|---|
| Albert-base | - | - | 6.7 | 49.6 | **68.0** | **88.2** | 82.7 | 85.0 | 63.4 |
| No-Grounding | Wiki-BK | 0 | 8.7 | 46.0 | 56.8 | 87.4 | 82.2 | 83.8 | 60.8 |
| Transferred i2t | CC-SBU+Wiki-BK | 0/1 | -0.4 | 31.1 | 60.2 | 85.7 | 82.1 | 84.5 | 57.2 |
| Transferred t2i | CC-SBU+Wiki-BK | 0/1 | 8.3 | 19.3 | 65.6 | 87.4 | 82.9 | 84.6 | 58.0 |
| Transferred t2i+i2t | CC-SBU+Wiki-BK | 0/1 | **11.1** | **52.0** | 61.1 | 88.0 | 82.7 | **85.2** | 63.4 |
| Associative (Zhang et al., 2020) | Wiki-BK | 16 | 2.4 | 48.4 | 54.3 | 87.6 | 79.6 | 81.9 | 59.0 |
| Associative (scene-based) | Wiki-BK | 16 | 1.2 | 49.3 | 54.5 | 86.1 | 83.0 | 84.0 | 59.7 |
| Associative (object-based) | Wiki-BK | 16 | 8.6 | 50.8 | 67.6 | 87.1 | **83.1** | 84.0 | **63.5** |

Table 3: Downstream tasks transfer results. Reported score is Spearman's correlation percentage for the JOCI task and accuracy percentage, otherwise. Albert-base is the pretrained model from (Lan et al., 2019) without architectural change or further pretraining before the downstream fine-tuning. Our models are based on Albert-base but have an additional cross-modal transformer that underwent an additional pretraining stage. $K$ is the number of images per example. The No-Grounding setup is equivalent to the transferred setup on Wiki-BK, or the associative setup with $K = 0$.

but it differs from the previous two in that its premises are only composed of image captions and video descriptions. This allows a more specific evaluation of concrete language, as opposed to abstract domain language in the RTE tasks.

**Penn Discourse TreeBank (PDTB)** (Prasad et al., 2008) contains a collection of fine-grained implicit (i.e., not signaled by a discourse marker) relations between sentences from the news domain in PDTB2.0, which signal the purpose of an utterance given a context utterance. We select level 2 implicit relations as categories. The task involves presupposition recognition and the ability to deal with non-literal meaning.

**VU Amsterdam Metaphor Corpus (VUA)** (Krennmayr and Steen, 2017) annotates the uses of verbs in sentences of the British National Corpus according to their level of metaphoricity. For instance, in the sentence *The alligator's teeth are like white daggers*, the use of the word *daggers* is metaphorical while *teeth* is not. We use the dataset of a shared task on verb metaphoricity detection (Klebanov et al., 2020) as well as another version where we kept only the nouns which we call VUAN.

Table 4 in appendix A.2 shows the dataset sizes, and Table 3 reports the results of the Wiki-BC trained models on the above tasks. We perform 8 runs for each task and report the median score.

As baselines we propose the Albert-base model alongside an ablated model without visual grounding and a state-of-the-art association model which uses visual input as described by Zhang et al. (2020). The Albert-base model (Lan et al., 2019) is a pretrained transformer that takes the text tokens as inputs and provides a representation of the text examples at the output of the $[CLS]$ token which is used to perform the logistic regression. In the No-Grounding model, the cross-modal encoder is trained on text-only corpus always using the placeholder as visual input. (Zhang et al., 2020) is an image retrieval system that ranks images according to the number of text keywords their captions contain. In this work, the retrieved image features are combined with the source text features to perform neural machine translation. Here, we use this retrieval model with our associative strategy as a baseline.

The No-Grounding model performance does not match the Albert-base model, which indicates that our linguistic pretraining is not as well tuned. However, grounding still yields improvements on the JOCI, PDTB and VUA tasks. The image-to-text (i2t) pretraining does not transfer well in the absence of images. But combining both text-to-image and image-to-text training leads to higher results. This suggests that the resulting model is able to use the obtained multimodal features to better perform the downstream tasks. The performance of the associative strategies depends on the image retrieval system but the comparison suggests that object-based and scene-based retrieval perform well enough to yield meaningful results.

Overall, the object-based associative strategy has the best performance and also performs well on language modeling, especially when generating descriptions. But the transferred strategy seems to be a better choice for the downstream NLP tasks, since it does not require images during fine-tuning and at test time.

## 5 Related work

Visual grounding has been repeatedly applied in NLP. The specificity of our work lies in the systematic categorization and comparison of grounding strategies, alongside the proposition of image-to-text and associative grounding.

It was shown to improve machine text translation (Specia et al., 2016; Elliott et al., 2017). However, these improvements only affected multimodal translation tasks. Zhang et al. (2020) showed that text-only translation tasks could benefit from external visual knowledge through image search, which we also demonstrate in our associative grounding strategy when applied to different NLP tasks.

Visual grounding has also been applied to transferable word representation learning. Bruni et al. (2014) generalize the distributional hypothesis by extracting discrete visual words from images and using them as context of text while relying on dimensionality reduction. This latter idea is reused with an autoencoder (Silberer and Lapata, 2014) and a SkipGram model (Lazaridou et al., 2015). Kiela and Bottou (2014) combine word embeddings with ImageNet visual features associated to words. Conversely, Collell et al. (2017) learn to predict ImageNet visual features from words and use the predicted *imagined* visual representation as auxiliary features. However, the above techniques are not applicable beyond the word-level. Image captioning datasets do provide sentences associated with relevant images, but the multimodal models trained on these are not commonly used in downstream NLP tasks. Kiela et al. (2018) train sentence embeddings for text processing tasks by learning to predict images from captions and report marginal improvements on SentEval (Conneau and Kiela, 2018) downstream tasks. This approach is similar to our text-to-image transferred grounding. Concurrent work incorporate grounding into Transformers-based pretrained language models (Tan and Bansal, 2020). They match each word in a text-only corpora with an image, and perform a masked image modeling with associated tokens. In our framework, this could be Transferred Associative object-based Grounding.

Other work targets grounded language understanding, but mostly in the context of robotics (Matuszek, 2018) where the text mostly regards task-oriented interactions in a closed-world.

Numerous text-level encoders were recently proposed to leverage images but were only applied to tasks involving both text and images. They all take BERT (Devlin et al., 2019) as a starting point and represent images with region features. They adapt masked language modeling to images with a form of masked image modeling, and each have more specific contributions. For instance, LXMERT (Tan and Bansal, 2019) demonstrates the value of visual question answering as a transferable pretraining task. UNITER (Chen et al., 2020) shows that masked image modeling and masked language modeling are best done separately. VL-BERT (Su et al., 2019) uses an additional text-only training. VilBERT (Lu et al., 2019) proposes a KL-divergence loss for masked image modeling.

Ideas comparable to associative visual grounding have also previously been used with external graph and text data. Liu et al. (2020) leverage triplets found in knowledge bases to better find entities in text. Guu et al. (2020) use language modeling pretraining in order to learn to retrieve and use relevant passages in question answering tasks.

## 6 Conclusion

In this paper we have proposed visual grounding strategies to make joint text-image models applicable to text-only processing. We have demonstrated that the associative strategy leads to consistent improvements when performing NLP tasks such as masked language modelling, plausibility estimation, metaphoricity detection and discourse relation prediction. Results could be further improved by refining the image representation and retrieval model. Since relying on image captions limited the number of images we could use, it would be interesting to investigate the use of larger-scale image datasets. Further work is needed to refine the effect of multimodality on NLP tasks. It is also interesting to study how NLP performance scales with the sizes of datasets used in pretraining the multimodal representations as is also suggested in (Kaplan et al., 2020).

## 7 Acknowledgements

This work is part of the CALCULUS5 project, which is funded by the ERC Advanced Grant H2020-ERC-2017 ADG 788506[9].

[9] https://calculus-project.eu/

## References


Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

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 *Proceedings of ECCV*.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Proceedings of AAAI*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of ICCV*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45. ACL.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *Proceedings of NAACL HLT, Volume 1 (Long Papers)*, pages 408–418. ACL.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 835–841. ACL.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors. 2020. *Proceedings of the Second Workshop on Figurative Language Processing*. ACL.

Tina Krennmayr and Gerard Steen. 2017. Vu amsterdam metaphor corpus. In *Handbook of linguistic annotation*, pages 1053–1071. Springer.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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 NAACL-HLT*, pages 153–163. ACL.

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 *Proceedings of ECCV*, pages 740–755. Springer.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of AAAI 2020*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of NIPS*, pages 13–23.

Cynthia Matuszek. 2018. Grounded language learning: Where robotics and nlp meet. In *Proceedings of IJCAI*, pages 5687–5691.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of ACL*, pages 4658–4664. ACL.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of NIPS*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*. ACL.

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 ACL (Volume 1: Long Papers)*, pages 2556–2565. ACL.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 721–732. ACL.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

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, Hong Kong, China. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36. ACL.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

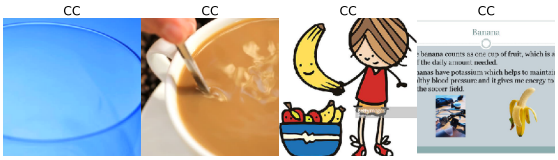
Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *Proceedings of ICLR*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*.

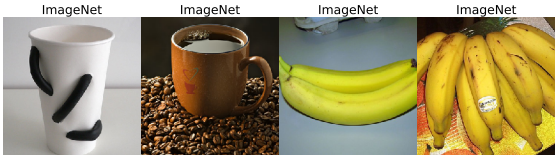
# A Appendix

## A.1 Image association systems

*a) A close up of a banana next to a cup with liquid.* (Scene-based)




*b) A close up of a banana next to a cup with liquid.* (Object-based)




*c) A close up of a banana next to a cup with liquid.* (Synthesis)

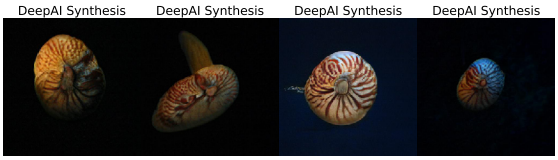



Figure 4: Association of images with a random text sample from the COCO dataset, with scene-based and object-based matching with $K = 4$ and $\kappa = 2$. Scene-based matching $(a)$ indexes images by their caption embedding to match it with the input text. CC refers to the Conceptual Captions dataset. Object-based matching $(b)$ identifies key concepts and matches them with ImageNet images. The Synthesis $(c)$ example illustrates the failure of the DeepAI image synthesis system when the input text is out of domain.

## A.2 Downstream tasks sizes

| Task | #Examples (Train/Val/Test) | #Labels |
|---|---|---|
| JOCI | 2.4k/299/298 | - |
| RTE | 2.2k/249/277 | 2 |
| SICK | 4.5k/500/4.9k | 3 |
| PDTB | 12.9k/1.2k/1.1k | 16 |
| VUAN | 5.0k/1.3k/2.2k | 2 |
| VUA | 16k/1.7k/5.9k | 2 |

Table 4: Dataset sizes for each downstream task