Exploring Phrase Grounding without Training: Contextualisation and Extension to Text-Based Image Retrieval

Letitia Parcalabescu and Anette Frank
Computational Linguistics Department, Heidelberg University
{parcalabescu,frank}@cl.uni-heidelberg.de

Abstract

Grounding phrases in images links the visual and the textual modalities and is useful for many image understanding and multimodal tasks. All known models heavily rely on annotated data and complex trainable systems to perform phrase grounding – except for a recent work [38] that proposes a system requiring no training nor aligned data, yet is able to compete with (weakly) supervised systems on popular phrase grounding datasets. We explore and expand the upper bound of such a system, by contextualising both the image and language representation with structured representations. We show that our extensions benefit the model and establish a harder, but fairer baseline for (weakly) supervised models. We also perform a stress test to assess the further applicability of such a system for creating a sentence retrieval system requiring no training nor annotated data. We show that such models have a difficult start and a long way to go and that more research is needed.

1. Introduction

When integrating vision and language in a multimodal task (such as Visual QA, Dialogue or Commonsense Reasoning), it is essential to align textual phrases with the image regions they refer to. This is called *phrase grounding* or *phrase localisation*. Phrase grounding is important because by grounding the textual modality in the image we link knowledge and context from both modalities and can expect improved model performance in joint vision and language tasks. Evaluating a system's ability of phrase grounding also opens the door to interpretability: We can infer what regions or phrases mattered for system predictions, inspect whether the system made a decision informed on both vision & language simultaneously and test whether the model aligned the two modalities without confusion.

Phrase grounding is one step towards solving general vision & language tasks [18, 21, 30]. While CV object detectors are trained to recognise (a fixed class of) objects from a closed vocabulary (*e.g. cat*), phrase grounding is expected to localise objects in an image referred to with free-form phrases (*e.g. a newborn Siamese kitten*). Phrase grounding can thus be considered a generalised object recognition task that considerably extends the visual-linguistic knowledge captured in pre-trained object detectors, and that requires systems to have or exploit additional linguistic knowledge.

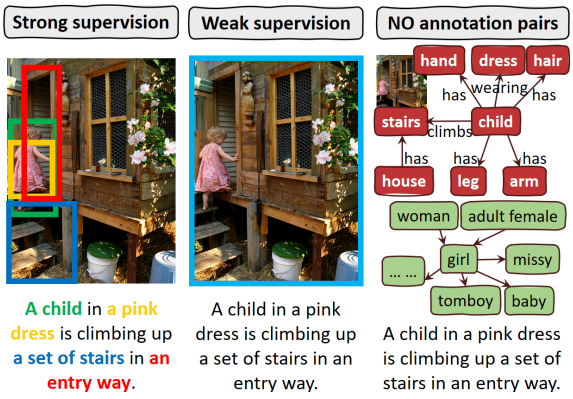

Figure 1. Strong, weak and no supervision. Our approach belongs to the latter. Scene graph in red, external knowledge in green.

A related application of phrase grounding is sentence-to-image alignment. Here the main interest is to detect images that correspond to a linguistic description, and phrase grounding scores can be re-purposed for aligning images and textual descriptions [16, 30].

The majority of phrase grounding methods have been strongly supervised with annotated phrase-region pairs [2, 3, 11, 12, 13, 23, 28, 29, 34, 35, 40, 39, 46] or weakly supervised using sentence-image pairs [1, 4, 41, 43, 47, 6] (cf. Fig. 1 left and middle). In recent work, Wang&Specia [38] propose to perform phrase grounding without any annotation pairs, by aligning caption phrases with object labels proposed by multiple object detectors and selecting a 'best' alignment by pairing the linguistically most similar phrase–label pairs. Their approach outperforms weakly supervised settings and defines a strong baseline for fully supervised

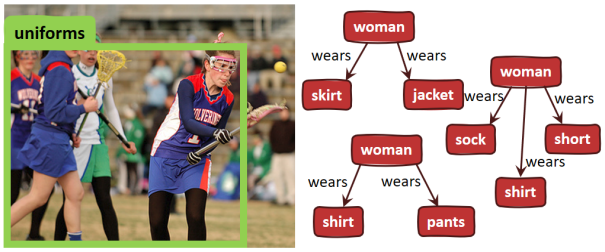

Figure 2. Simplified scene graph (SG) for the image on the left. Nodes in SG represent objects with associated bounding box information (labels and coordinates).

methods. Wang&Specia's method can be characterised as a 'bag of words (BoW)' approach that exploits the linguistic similarity of paired caption phrases and object labels assigned to region proposals – without considering structural properties of text or image. Moreover, their method combines outputs of different detectors, and can be regarded as being unfair against (weakly) supervised models, since the latter rely on only one region proposal backbone with a bounded coverage. Another weakness of their model is that it does not consider context. Their distributed word embeddings are not contextualised and can suffer from undesired associations, e.g. mapping synonyms and antonyms to similar space regions – a displeasing property when searching for *most similar* object labels and phrases.

Still, Wang&Specia's baseline requiring no training or supervision is an interesting approach and worth being further explored in comparison to systems that require supervision from large training sets. Thus, our work has two objectives: we aim (i) to explore and expand the upper bound of an unsupervised approach by extending it with contextualisation in the image and language representation, and (ii) to assess its performance on the related task of sentence-to-image retrieval. Our contributions relate to four aspects:

i) *Label set size and upper bound of the underlying object detector:* Wang&Specia employ a collection of object detectors that jointly improve system results by considerable degrees. In order to ensure fair comparison to competing (weakly) supervised systems, we aim to employ a single object detector with a finer-grained label set. Also, while Wang&Specia's object detectors offer a proposal upper bound of ca. 50%, we establish a more realistic improved baseline using a single object detector with a proposal upper bound close to 90%.

ii) *Structuring the visual representation:* While Wang&Specia compare unstructured label sets (BoW), we contextualise the visual modality by exploiting structured information about the image provided by a scene graph. This allows us to address imperfections of object detectors and small class label sets.

iii) *Knowledge injection:* We contextualise the semantic labels of image region proposals in the linguistic modality by grounding them in structured background knowledge sources and thereby extend the model's linguistic capacities. We (a) compensate the deficiencies of object detectors by enriching the vocabulary coverage of the upstream object detection system (using the Open Images v5 [20] label hierarchy) and (b) compute word similarity on the WordNet [9] graph structure, as an alternative to only measuring cosine similarity over distributional word embeddings as in Wang&Specia.

iv) *Unsupervised phrase grounding for image retrieval:* We explore the limitations of unsupervised phrase grounding by applying our method to the sentence-image retrieval task. We do this by re-purposing phrase grounding scores as image ranking scores.

Experiments on the Flickr30kEntities dataset [30] show that (a) our method – **using no training nor supervision, but scene graphs and external knowledge** – outperforms state-of-the-art weakly supervised models by a large margin and surpasses the majority of supervised models, establishing a strong baseline for strongly supervised models. Our model that uses only one object detector can be a good alternative to engineering a suitable ensemble of object detectors: it outperforms Wang&Specia's system on Flickr30k-Entities, and we show that it benefits from leveraging knowledge and context in both modalities. Yet, while our method is competitive when tasked to ground entities in the same image, (b) our experiments on text-based image retrieval on Flickr30k [44] show that it its results can not compete with recent supervised state-of-the-art systems. Nonetheless, our method sets a noteworthy baseline as the (to the best of our knowledge) only sentence-image retrieval method requiring no supervision and no training data.

2. Related work

Increasingly accurate object recognition systems are being developed that extract bounding boxes of detected objects and label these with classes from a fixed class label set (vocabulary) [8, 10, 24, 31, 32, 33, 48]. The task of phrase grounding clearly profits from enhanced object recognition systems, yet it needs to solve an extended task: grounding phrases from an open vocabulary to objects in an image, by aligning phrases with proposed bounding boxes.

Current research on phrase grounding can be divided into *strongly supervised* and *weakly supervised* approaches.

Strongly supervised approaches [2, 3, 7, 11, 12, 13, 23, 28, 29, 34, 35, 40, 39, 46] make use of different techniques: some project the visual and textual modalities onto the same space [29, 27, 40, 39], others attend to the correct image region and reconstruct the corresponding phrase [34]. The

architectures are becoming ever more intricate: Hinami & Satoh [12] train object detectors with an open vocabulary, Plummer *et al*. [28] condition the textual representation on the phrase category. We also find approaches that treat phrases not individually, but as a sequential and contextual process [7]. Recent approaches integrate visual features from object detectors into Transformer models [23, 25].

Weakly supervised systems [1, 4, 41, 43, 47, 6] do not have access to paired bounding boxes and phrases in training. They learn to ground phrases implicitly by solving downstream tasks such as caption-to-image retrieval [6], use external region proposals and knowledge [1, 47], attention maps [41] or co-occurrence statistics [43].

The regular techniques for text-based image retrieval methods (TBIR) are mapping text and image features to a learnt joint space in order to compute distances between vectors from the two different modalities [14, 17, 22, 30]. Early approaches [17] adopt a CNN-based region proposal network to encode the image at the level of objects and a bidirectional RNN [36] for text processing. Latest approaches have a finer degree of refinement of visual and textual features. *E.g*. [22] use a Graph Convolutional Neural Network (GCN) [19] to reason about the relationships of the proposed image regions. With a gate and memory mechanism, they reason globally on the relationships-enhanced features of the GCN. The latest innovation to this line of work is the Visual Transformer model [5, 25] with the self-attention mechanism. It is pre-trained in a multi-task fashion (masked multi-modal modelling and multimodal alignment prediction) on even larger training data and is applicable to many downstream tasks beyond image retrieval. The most similar approach to ours classifies image regions into objects, actions and properties and uses these labels and visual features to learn their semantic ordering [14]. The ordering is supervised by sentence generation through an LSTM and sentence similarity scoring. In our work, we aim to test the performance of an unsupervised system that has access only to the classified object labels and noun phrases in the text – while the aforementioned systems are guided by a loss function, have access to visual features and process these and textual embeddings with high sophistication.

Recently, Wang&Specia [38] developed the (to the best of our knowledge) only other approach on phrase grounding that does not need paired training examples. They combine the object detections of four different systems and one colour detector into labelled bounding box candidates. They embed all labels and the phrase with word vectors and rank the labels by cosine similarity. As grounding proposals they choose the bounding box of the highest-ranked label. Their method can be considered a 'bag of objects (BoO)' approach. We, by contrast, will make use of a single object detector and exploit structured context in the linguistic and the visual modality.

3. Contextualising phrase localisation with knowledge and scene graphs

The phrase grounding task measures a system's capacity of identifying an area in a given image I that a phrase p is referring to. The system is tasked to deliver the bounding box b in image I that circumscribes the location of that area.

Grounding without training Our approach to the task does not rely on annotated pairs $\{p, b\}$ (p a phrase, b a bounding box in I), as in strongly supervised settings, nor on pairs $\{p, I\}$ as in weakly supervised settings. Our model uses information from (i) out-of-the-box object detectors, (ii) out-of-the-box scene graph generators, (iii) external linguistic knowledge bases and (iv) word embeddings obtained from language embedding models trained on generic text, not specifically on phrases or vocabulary of a phrase grounding dataset. Even though models used for generating (i-iv) may be trained with supervision, our phrase grounding system is completely unsupervised and all information it uses is extracted from readily available models or knowledge bases.

Contextualisation When creating a visual representation for phrase grounding, our aim is to exhaustively capture the visual content. Wang&Specia [38] employ multiple object detectors to extract a 'bag of objects (BoO)': a set of detected objects $\{b; l\}$ consisting of their coordinates b and predicted labels l. But a bag of objects does not model any dependencies or neighbourhood between objects in the image, and risks to ignore important contextual information. We propose to use a structured representation of the image content in form of a scene graph that models relationships between objects. This design puts objects into context and thus, the visual context can offer distinguishing context information that may, *e.g*., disambiguate cases where the same coarse-grained category label is assigned to different entities, while the visual neighbours may offer additional hints for correctly grounding the phrase. An example is seen in the top left image of Fig. 5.

Another advantage of a structured visual and linguistic representation is that it can bridge missing explicitness in the language. This is seen in Fig. 2, where we ground the phrase *uniforms*. A BoO approach will lack the information that a set of *shirts, jackets, pants* belonging to *women* jointly constitute what is called *uniforms*. By contrast, an LKB such as WordNet captures the relatedness of these concepts trough a relation path *e.g. uniform – clothing – skirt*.

Model overview We build a **visual representation** of the image I consisting of a set of object proposals $\Omega = \{o_i\}$ in the form of bounding boxes and a linguistic label $o_i = \langle b_i, l_i \rangle$. We extend prior work by (i) including relational information r_{ij} between object detections that form a *scene graph* $SG = \{(o_i, r_{ij}, o_j)\}$ (red nodes in Fig. 3). We also (ii) *enrich the linguistic components of the object*

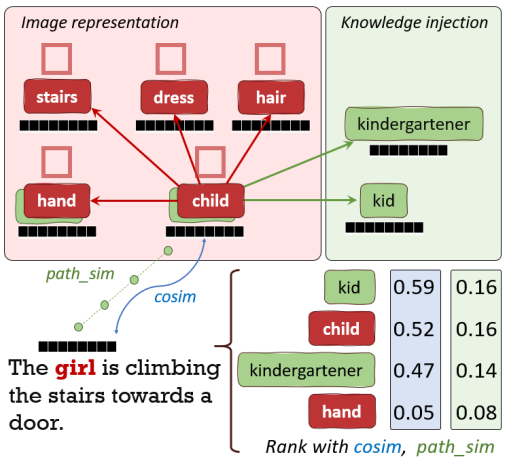

Figure 3. Sketch of our approach. The scene graph (red) and knowledge nodes (green) – retrieved through *is-a* relations – contain information about bounding boxes and word embeddings. The embeddings are compared with cosine similarity to the word embedding of the phrase and with WordNet path similarity between label and phrase. After ranking, the bounding box related to the maximum score is the grounding result.

representations l_i by linking them to *structured semantic knowledge* from a **linguistic knowledge base** (green nodes in Fig. 3). Specifically, we (a) map labels l_i of the scene graph nodes to small sub-graphs $\{k_i\}$ that connect l_i with its direct neighbours in LKB, using selected relations, such as *hyponymy* in WordNet, or else (b) use the full LKB to enhance the search for the best-fitting phrase-object pair, using a shortest path method to compute path similarity.

To obtain **linguistic representations for the phrases** p_i in the text, we use a language embedding model to encode them into vectors p_i^h. We use same embedding model to encode the labels r_{ij} and lemmas l_i in the **enriched visual representation** $SG \cup \{k_i\}$ (*i.e.*, the labelled object nodes and edges in the scene graph and the linked subgraphs $\{k_i\}$ from the LKB), converting them to vectors in SG = $\{(\langle b_i, l_i^h\rangle, r_{ij}^h, \langle b_j, l_j^h\rangle)\}$ (and similarly for the nodes in subgraphs $\{k_i\}$) using the word embedding model. Finally, we map the phrases p_i to all concepts s_i in the LKB (using their lemma) and select the one with shortest connecting path.

With these extensions, we can measure the similarity between phrase representations p^h and any linguistic component l^h, r^h or k^h in the (enriched) visual representation $SG \cup \{k_i\}$ using (i) the cosine similarity metric over vectors as well as (ii) a score based on the shortest path connecting the concept representation of the phrase and the object label in the extended linguistic representation. By ranking the distances of all pairings of phrase p_i and visual object o_j representations, we select the highest-ranking visual proposal o_i for the query phrase p_i and generate the grounding result.

3.1. Structured visual-linguistic representations

Scene graph generation We first extract object detections with bounding boxes $\Omega = \{o_i\}$. We then extract a scene graph from the image to build a graph of structured proposals. The graph $SG = \Omega \cup R = \{(o_i, r_{ij}, o_j)\}$ (represented in red in all figures) is described by the set of object nodes $\Omega = \{o_i\}$ containing bounding box information and labels and the set of labelled edges that model visual relationships $R = \{r_{ij}\}$ between detected objects.[1]

We extract scene graphs from images using the generator of Zellers *et al*. [45]. We choose this model because it performs almost state-of-the-art and includes ready-to-run code. The scene graph generation model [45] was trained on Visual Genome [21] with 150 object labels and 50 relationships. For generating the scene graph, we use the 50 most confident relationships from the generator's output.

Enhancing visual with structured linguistic representations When grounding an open-vocabulary phrase to an image object labelled with a coarse category, we need to inform the nodes o_i of the scene graph with semantic knowledge in order to make correct predictions. When using distributional word embeddings to encode object labels in a vector space, one has to reckon with unintuitive side-effects, *e.g.* when the vector representation of *groom* is closer to the one for *woman* than the one for *man*. To counter such effects, we create an enhanced representation of *man* by aggregating it with the meaning of neighbouring concepts $\{k_i\}$ in the linguistic ontology that further characterise the entity. For this, we map each object label l_i to a concept $k_i \in LKB$ and retrieve the direct neighbours $\{k_i\}$. We then shift the vector o_i^h towards an enriched, contextualised meaning $\tilde{o}_i^h$ by computing the mean over the embeddings of a node o_i^h and its direct neighbour concepts $\{k_i^h\}$. *E.g.*, the neighbours *sir, guy, adult male* guide the system towards choosing the correct answer *man*, instead of *woman*. We apply similar aggregations to SG nodes and relations, to obtain contextualised visual object representations $\bar{o}_i^h$.

In this work, we experiment with two LKB's: WordNet [9] and the Open Images (OI) v5 [20] class label hierarchy. When using WordNet as a LKB, the mapping of l_i to $\{k_i\}$ is facilitated, since the object labels are annotated with WordNet senses. The direct neighbours in $\{k_i\}$ consist of synonyms, hypernyms and hyponyms. For the OI label hierarchy, the neighbourhood graph $\{k_i\}$ consists only of the direct hypernyms and hyponyms, as illustrated in Fig. 4. The mapping between l_i and k_i is also unambiguous, because we use Faster-RCNN [33] trained on Open Images v4 [20] to predict objects as grounding proposals, thus the l_i labels are all linked to the hierarchy.[2]

[1] For comparability with Wang&Specia we remove any detections (and relationships between them) from the scene graph that are not covered by the object detector.

[2] We only report the results with the OI hierarchy. While we also ex-

3.2. Text representation

For phrase grounding we map query phrases to a vector representation. For this we perform part-of-speech tagging with the Stanford Tagger [37] using the NLTK package. We extract and lower-case adjectives and nouns, perform spell checking and embed them using word embeddings. For multi-word phrases we compute the mean over all token embeddings in the phrase to obtain the final phrase vector. We use 300-dimensional word2vec [26] embeddings.

As an alternative to word embeddings, we utilise WordNet to compare labels and phrases. Lexical meaning in WordNet is represented in terms of *synsets*, *i.e.* sets of synonyms for a given word sense. Meaning relations (hypernymy, hyponymy, *etc.*) are defined between synsets. We link phrases to labelled nodes in the visual representation by mapping all words in a phrase to all their possible WordNet senses and search for the shortest paths that connect any synset of the phrase p to any candidate labelled object o_i.[3]

3.3. Grounding by ranking proposals

In the final phrase grounding step we rank the proposals from the visual representation and select the highest-ranking candidate according to their semantic similarity. For ranking (Fig. 3), we compute the **grounding score** γ_i between a phrase p and an image region o_i by combining (i) *cosine similarity* between word embeddings of labels of the visual representation o_i^h and the embedding of the query phrase p_i^h multiplied by (ii) the *maximum WordNet path similarity score* $\text{path\_sim} = \frac{1}{\text{d}+1} \in [0, 1]$ based on the shortest path distances d connecting the label synset o_i and any synset p_{syns} of the phrase in the WordNet hypernym/hyponym taxonomy. The node or relation with maximum grounding score γ_i is chosen and its bounding box is predicted as the phrase localisation result.

$$\gamma_i = \text{cosim}(o_i^h, p^h) \cdot \max_{syns}(\text{path\_sim}(o_i, p_{syns})) \qquad (1)$$

We adopt several policies: If a knowledge node is chosen by the process, the predicted image coordinates are defined by the bounding box of the scene graph node to which the knowledge is attached. If two nodes obtain equal scores, we predict the union of their bounding boxes. But if a scene graph node and a knowledge node score the same, we consider only the scene graph node, to minimise the amount of equally scoring nodes (*e.g.*, when the object detector predicts *person, woman* and knowledge appends again *person*).

perimented with WordNet [9] manifesting similar results, we decided to encode WordNet knowledge only in the language representation (Section 3.2) because the (i) bigger size of the WordNet ontology allows for computing similarities between phrases and nodes, while the OI label hierarchy is too small for this purpose, delivering many out-of-vocabulary errors. (ii) In the visual representation, by contrast, WordNet expands the visual graph with many superfluous fine-grained labels.

[3]Each node o_i in the scene graph is annotated with its WordNet [9] synset using the Visual Genome [21] sense annotations, hence the scene graph nodes are disambiguated and we use a single synset per node label.

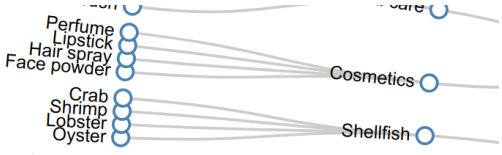

Figure 4. Excerpt from the Open Images v5 hierarchy [15].

3.4. Image retrieval by ranking proposal scores

When considering phrase grounding alignments between all phrases in a sentence and an image, we can use the sum of their grounding scores for sentence-to-image retrieval. By doing so, we can assess the practicality of phrase grounding methods requiring no training nor paired annotated data on the well-established benchmark of sentence-image retrieval. We do not expect these methods to perform very well, but neither did we expect them to perform competitively on the phrase grounding task.

We create an unsupervised image-from-text search engine without involving any training data, by ranking the grounding score per image Γ_I between the query sentence and 1000 candidate images (following the standard protocol for Flickr30k). Γ_I is defined as the sum[4] of the grounding scores between the N phrases in the sentence $\Gamma_I = \sum_n^N \gamma_n$.

4. Experiments

We test our method on the established **Flickr30k Entities** [30] dataset for **phrase grounding**. It is based on Flickr30k [44] and offers annotated noun phrases in image captions that are aligned with bounding box coordinates in the image. The vocabulary of the phrases is relatively diverse, comprising over 5,000 words on the whole dataset and 2806 on the test split. The test set amounts to 1,000 images, 16,576 phrases, of which 7,180 are unique.

In addition we perform an experiment on **sentence-image retrieval** on the Flickr30k[44] test set, which comprises 1000 images with five captions each. For each caption the system is tasked to choose the image with which it was paired from the full set of 1000 images. The data contains around 5 phrases per image in average.

Phrase grounding metric: In order to be comparable to previous work, we choose the accuracy metric for evaluation. A bounding box is considered to be correctly detected if the intersection over union (IoU) with the ground truth is greater than or equal to 0.5. As an upper bound **UB** we report the percentage of the cases in which the ground truth can be be found in the proposal set with $IoU \geq 0.5$.

[4]In this setting, normalisation is not necessary because each sentence has the same number of phrases.

Method	Acc (Var) %	UB %
No training		
tfoid+CC+PL [38]	50.49 (5.37)	57.81
tfoid [38]	44.69	50.04
No training – ours		
tfoid *(reimpl. of [38])*	46.08 (7.02)	61.17
tfoid + sg	46.62 (6.50)	61.17
tfoid + hier$_{kn}$	46.82 (6.73)	61.17
tfoid + hier$_{kn}$ + wn$_{path_sim}$	47.74 (6.78)	61.17
tfoid + sg + wn$_{path_sim}$	**47.92** (6.53)	61.17
visgen	56.30 (5.51)	87.88
visgen + sg	56.40 (5.42)	87.88
visgen + sg + wn$_{path_sim}$	**57.08** (5.30)	87.88
Weakly supervised		
GroundeR [34]	28.94	77.90
KAC Net + Soft KBP [1]	38.71 (8.41)	
Strongly supervised		
GroundeR [34]	47.81	77.90
CCA [30]	50.89	$\geq$ 75.73
Wang *et al.* [39]	51.05	84.58
QRC-Net [3]	65.14 (3.77)	89.61
VisualBERT [23]	71.33	87.45

Table 1. Results on Flickr30k Entities. **UB**: upper bound, **Var**: variance over categories (cf. Tbl. 2), **sg**: scene graph, **wn**$_{path\_sim}$:WordNet path similarity, **tfoid**: Faster-RCNN trained on Open Images, **hier**$_{kn}$: Open Images hierarchical knowledge, **visgen**: Faster-RCNN trained on Visual Genome.

Image retrieval metric: For sentence-image retrieval we report performance as median rank (mdR) and Recall@k, *i.e*. the percentage of instances for which the ground truth was ranked among the top k proposals.

4.1. Phrase grounding on Flickr30kEntities

We compare our method to Wang&Specia [38], the only other approach requiring no paired training data. We also compare to weakly supervised [1, 34] and fully supervised settings [3, 23, 34, 39]. Hereby we put our method into perspective: while we do not rely on any annotated pairs, weakly and strongly supervised settings use annotated data. Furthermore, they are unfairly overpowered: besides bounding box proposal generators they compute visual features on the image and refine word vectors during training.

Table 1 and 2 report our results for phrase grounding on Flickr30kEntities. We show results for different procedures of computing the visual representation, which fall into two main categories:

(a) *Coarse-grained labels*: Here we rely on an object detector with a relatively low number (545) of labels compared to the phrase vocabulary. This **tfoid** (OI) detector is also used in Wang&Specia. We re-implement their approach for two reasons: Firstly, we want to create a fair unsupervised baseline for phrase grounding that uses a single detector, the same amount as the (weakly) supervised systems we compare to. Secondly, we consider a recalculation of their results necessary, since we extract 16,576 phrases from Flickr30kEntities, while Wang&Specia's results are based on only 14,481 phrases.[5]

tfoid is our re-implementation of Wang&Specia using only object detection proposals with confidence higher than 0.1 from a Faster-RCNN trained on Open Images on 545 object categories.

tfoid+sg includes labels (but no bounding boxes beyond the tfoid detections) and relationships corresponding to the neighbours in the scene graph and **tfoid+sg+wn**$_{path\_sim}$ extends **tfoid+sg** in that it computes the grounding score using the combination of cosine similarity with WordNet path similarity between the scene graph representation and the phrases.

tfoid + hier$_{kn}$ performs grounding with the visual representation without scene graph information, but enriched with the Open Images label hierarchy (in Fig. 4) and **tfoid + hier**$_{kn}$**+wn**$_{path\_sim}$ combines cosine similarity with the WordNet path similarity.

(b) *Fine-grained labels*: Here the object detector is trained with a relatively big vocabulary for object detectors:

visgen represents a Faster-RCNN model trained on Visual Genome (mAP = 4.4) with a label set of 1600 objects available at [42]. All detections have a confidence higher than 0.1. We combine it with the scene graph and WordNet path similarity in **visgen+sg+wn**$_{path\_sim}$.[6]

Our best performing model in the *detector with low vocabulary* setting (a) integrates scene graphs, the WordNet structure and slightly outperforms our direct competitor in the single-detector setting [38].[7] We observe small and comparable increases when using the contextualised representations based on the scene graph and the label hierarchy as external knowledge, which go along with a reduction of variance over phrase categories (see Table 2). The extension of similarity-based ranking using WordNet path similarity shows consistent improvements over the contextualised representations and over the BoO baseline.

When using a detector with *large vocabulary*, we set a new baseline for models without training nor annotated data. We observe a higher BoO baseline, which is also reflected in the raised upper bound (ca. 88%), exploring the limits of unsupervised phrase grounding systems.

Comparison to systems using weakly supervised methods shows that both our and Wang&Specia's method that

[5]For extracting phrases, we use functionality available on the GitHub page of the Flickr30K Entities dataset [27].

[6]We do not report experiments of **visgen** enriched with the label hierarchy because its vocabulary of 1600 classes is much richer than the hierarchy with around 600 labels.

[7]With our own reimplementation. Compared to [38]'s published results using tfoid as single detector we observe an increase of 3.5 pp. accuracy.

Method	people	clothing	bodyparts	animals	vehicles	instruments	scene	other	overall	var
tfoid	66.94	36.95	21.03	81.27	82.02	57.41	18.73	25.62	46.09	7.02
tfoid+sg+wn$_{path_sim}$	67.35	43.37	21.80	81.47	82.27	57.41	21.45	26.95	47.92	6.53
visgen	67.68	56.98	34.76	73.75	73.65	6.17	60.23	38.29	56.30	5.51
visgen+sg+wn$_{path_sim}$	67.92	57.11	34.61	73.75	73.65	8.02	60.69	38.52	**57.08**	5.31

Table 2. Accuracy per category on Flickr30k Entities. **var**: variance over categories, **sg**: scene graph, **wn**$_{path\_sim}$: WordNet path similarity.

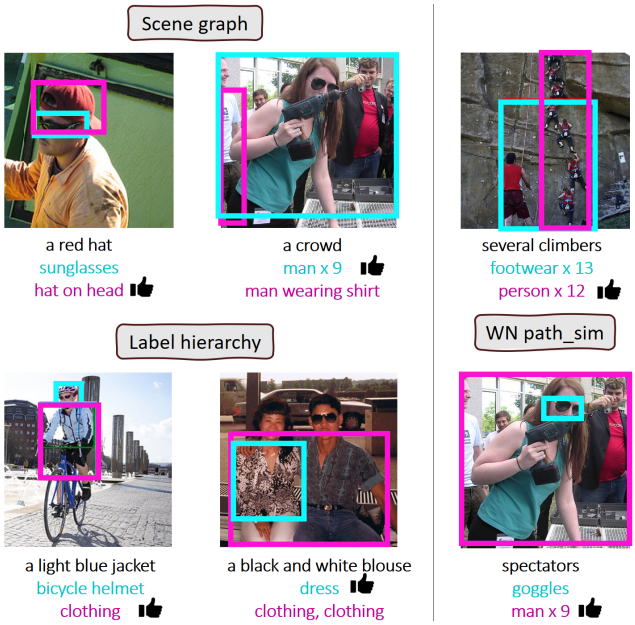

Figure 5. Grounding output example on Flickr30kEntities. The query phrases are represented in black and are localised by **tfoid** in cyan. Magenta stands for the response of our system enabled with the contextualisation method indicated in the grey box.

do not require training nor paired annotated data, outperform these systems by a large margin. We also stress that our results outperform many supervised systems, thus setting a strong baseline for state-of-the-art supervised models.

We also compute the variance of the accuracy over the eight phrase categories in the dataset (see Table 2). This shows how the performance throughout all classes varies when employing a different object detector, but also the consistent gains in accuracy we obtain by enriching the representation with additional knowledge – along with reduction of the variance of results for each category.

Impact of the size of the label set. In the ***coarse-grained*** label setting (a) we compare against Wang&Specia and the re-implementation **tfoid** of their best performing single object detector method and show that our method **visgen+sg+wn**$_{path\_sim}$ outperforms their ensemble model that relies on three different object detectors. The significantly higher bound of our object detector (88% compared to 58% for Wang&Specia), can not be the only explanation for the observed performance improvement: Table 1 shows that accuracy does not directly correlate with the upper bound. While a higher bound increases the possibility of finding the searched object, more coverage of the image generates more proposals, thus more confounders for the ranking step. We thus suspect that the reason for the accuracy increase lies in the more fine-grained label set rather than lower coverage of **tfoid**, and in the added scene graph structure together with the ranking based on the WordNet path similarity score.

Impact of the contextualised visual representation. In Table 1 we show with **tfoid+sg** that a *contextual representation* of the image (compared to BoO) brings only minimal benefit (little above half of percentage point) for grounding. This improvement might seem small at first, but there are important gains hidden by the accuracy statistics: In Fig. 5 on the top left we see one positive and one negative example of the scene graph impact. The system's pink response in the top left corner is informed by the scene graph context of *hat*, shifting the vector representation towards the correct meaning. In the second example, we see the effect of the evaluation procedure where a union of the bounding boxes representing one phrase is created: The ground truth boundary comprises all men in the picture. As an effect of the scene graph contextualisation, the nine *man* detections in the picture are unique due to their different neighbourhood. The result in cyan of the system not informed by the scene graph structure, does not differentiate between the *men*, generating the right proposal for the query *crowd*. The system having access to the scene graph (in pink), ranks unique candidates and detects only one part of the crowd. An additional reason for a lower than expected impact of the scene graph lies in errors of the generated scene graph. Another reason is the loss of some generated scene graph parts when mapping it to the existing object detections, in order to be comparable to previous work by not including additional bounding boxes.

Impact of knowledge injection and WordNet path similarity computation. By injecting the already existing knowledge of the label hierarchy of **tfoid+hier**$_{kn}$, we show that already a small amount of linguistic knowledge can boost the performance (see Fig. 5 for examples). Our methods **tfoid+hier**$_{kn}$**+wn**$_{path\_sim}$, **tfoid+sg+wn**$_{path\_sim}$ comprising the WordNet path similarity score together with the cosine similarity show the greatest jump in performance (around 1 pp. and 1.5 pp. respectively) and represent our second way to introduce the structure of a linguistic know-

Figure 6. Sentence-image retrieval examples on Flickr30k. Two positive (green bounding box) with the system's second guess (no bounding box) over the same coloured background. One negative example (red) where the second guess is the ground truth example.

ledge base. By these means, we show that a complementary way of computing word similarities besides just using word embeddings benefits our purposes. An analysis of the number of hops in WordNet between phrase and proposed grounding shows that around 80% of the correct proposals are made by number of hops no greater than 3 and that they lead to better predictions than larger numbers of hops.

4.2. Text-based image retrieval by proposal ranking

We performed experiments with both the **tfoid** detector having a relatively low upper bound (around 60%), as well as with **visgen** with a higher upper bound (roughly 88%). We select representative results of a selection of sentence-retrieval systems and report our experiments in Table 3 with a BoO approach with **visgen** and with integrating WordNet path similarity **visgen+wn**$_{path\_sim}$, hereby improving in R@1. We observe that while our method requiring no training is no match to the state-of-the-art systems trained with supervision, it is nonetheless decent for the only training-less sentence-image retrieval method known to us.

Discussion While our phrase grounding method is keeping pace with (weakly) supervised methods, we are stress-testing our phrase grounding system by extending it to sentence-image retrieval. Simultaneously, our method is an experiment for what a sentence-image retrieval method without any training nor supervision potentially delivers. The system is (a) requiring no training – but off-the-shelf object detectors; (b) uses no visual features – only the language representation of phrases and object detection labels; (c) no language context – but the noun phrases in the sentence, and yet **visgen** can select in 14% of the cases the right picture (among 999) in the first trial. While this unsupervised method represents strong competition to the supervised neural computer vision systems of 2015, it is no challenge to recent supervised models. This demonstrates that while object detectors can be used to sufficiently discriminate between different image regions when performing phrase grounding without regarding the image pixels, visual features are necessary when being challenged with the much harder task to rank 1000 pictures. In this task, several pictures can be characterised by the same BoO (cf. the negative examples in Fig. 6, where both pictures contain a *man, woman, cars* and distinguishing features like ethnicity and the glass are not captured by the object detector). Furthermore, full-fledged processing of the linguistic modality incorporated in the sentence is crucial: only looking at the noun phrases misses discriminating clues in the language.

Method	R@1	R@5	R@10	mdR
DVSA$_{\text{CVPR15}}$[17]	15.2	37.7	50.5	9.2
SCO$_{\text{CVPR18}}$ [14]	56.7	87.5	94.8	-
ViLBERT$_{\text{NeurIPS19}}$ [25]	58.2	84.9	91.52	-
No training				
tfoid	8.7	19.8	27.5	52
visgen	14.6	34.0	43.7	17
visgen + wn$_{path_sim}$	15.3	33.9	43.7	16

Table 3. Results on sentence-based image retrieval on Flickr30k in terms of Recall@k/**R@k** (high is good) and median rank **mdR** (low is good).

5. Conclusion

We propose a method that tackles the phrase grounding task without using annotated image-language pairs. We show that a structured representation of images and injection of linguistic knowledge are beneficial in a system that requires no training nor loss function to guide the attention to relevant input regions. Our model surpasses the performance of all weakly supervised and many supervised models on Flickr30kEntities and establishes a more serious baseline for (weakly) supervised models than prior work. A crucial factor are extensions for visual and linguistic contextualisation, which may be further enhanced in future work. We also stress-test our alignment system on the challenging Flickr30k sentence-image retrieval task and achieve first noteworthy results for a system without a training phase.

The strong performance of phrase grounding methods requiring no training casts doubt on the adequacy of supervised architectures trained on annotated phrase-region pairs, since these highly complex and over-parameterised trainable systems do not improve much over our approach, which does not require a training stage.

References

[1] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4042–4050, 2018. 1, 3, 6

[2] Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. Msrc: Multimodal spatial regression with semantic context for phrase grounding. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 23–31. ACM, 2017. 1, 2

[3] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832, 2017. 1, 2, 6

[4] Lei Chen, Mengyao Zhai, Jiawei He, and Greg Mori. Object grounding via iterative context reasoning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 3

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 3

[6] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 3

[7] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019. 2, 3

[8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. *CoRR*, abs/1904.08189, 2019. 2

[9] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 2, 4, 5

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[11] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. Open-vocabulary object retrieval. *Robotics: science and systems*, 2(5):6, 2014. 1, 2

[12] Ryota Hinami and Shin'ichi Satoh. Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2615, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 1, 2, 3

[13] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 1, 2

[14] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018. 3, 8

[15] Neil Alldrin Vittorio Ferrari SamiAbu-El-Haija Alina Kuznetsova Hassan Rom Jasper Ui-jlings Stefan Popov Andreas Veit et al. Ivan Krasin, Tom Duerig. Open Images v5 Flare Dendogram. https://storage.googleapis.com/openimages/2017_07/bbox_labels_vis/bbox_labels_vis.html. [Online; accessed 10-November-2019]. 5

[16] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 3, 8

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[20] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2:3, 2017. 2, 4

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. 1, 4, 5

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019. 3

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 2, 3, 6

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 3, 8

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc. 5

[27] Bryan A. Plummer. Flickr30k GitHub. https://github.com/BryanPlummer/flickr30k_entities. [Online; accessed 10-November-2019]. 2, 6

[28] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–264, 2018. 1, 2, 3

[29] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937, 2017. 1, 2

[30] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017. 1, 2, 3, 5, 6

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[32] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 4

[34] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 1, 2, 6

[35] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4694–4703, 2019. 1, 2

[36] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 3

[37] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for computational Linguistics, 2003. 5

[38] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, Oct. 2019. IEEE. 1, 3, 6

[39] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 1, 2, 6

[40] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *European Conference on Computer Vision*, pages 696–711. Springer, 2016. 1, 2

[41] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 1, 3

[42] Jianwei Yang. Faster-RCNN trained on Visual Genome. https://github.com/jwyang/faster-rcnn.pytorch. [Online; accessed 14-November-2019]. 6

[43] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6125–6134, 2018. 1, 3

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 2, 5

[45] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018. 4

[46] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 557–566, 2017. 1, 2

[47] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018. 1, 3

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2