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.

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
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 contextu-

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.

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 know-ledge 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 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, 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. 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. 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. With a gate and memory mechanism, they reason globally on the relationships-enhanced features of the GCN.

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 contextual 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 through 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

3. Contextualised 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 contextual 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 through 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
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 import 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 $\bar{o}_i$ towards an enriched, contextualised meaning $\bar{o}_i^h$ by computing the mean over the embeddings of a node $\bar{o}_i$ and its direct neighbour concepts $\{k_i\}$. 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, hyponyms and hypernyms. 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 (hyponymy, hypernymy, 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. 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 \( \gamma_i \) between a phrase \( p \) and an image region \( o_i \) by combining (i) cosine similarity between word embeddings of labels of the visual representation \( o_i \) and the embedding of the query phrase \( p \) multiplied by (ii) the maximum WordNet path similarity score \( \text{path}_\text{sim}(o_i, p_{\text{syns}}) \) between the query sentence and 1000 candidate images (following the standard protocol for Flickr30k).

\[
\gamma_i = \cos(o_i^h, p^h) \cdot \max_{\text{syns}}(\text{path}_\text{sim}(o_i, p_{\text{syns}}))
\]

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).

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 \( \Gamma_I \) between the query sentence and 1000 candidate images (following the standard protocol for Flickr30k). \( \Gamma_I \) is defined as the sum of the grounding scores between the \( N \) phrases in the sentence \( \Gamma_I = \sum_{n=1}^{N} \gamma_n \).

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 \( \text{IoU} \geq 0.5 \).

Footnotes:
1. Each node \( o_i \) in the scene graph is annotated with its WordNet synset using the Visual Genome [21] sense annotations, hence the scene graph nodes are disambiguated and we use a single synset per node label.
2. In this setting, normalisation is not necessary because each sentence has the same number of phrases.
Table 1. Results on Flickr30k Entities. UB: upper bound. Var: variance over categories (cf. Tbl. 2). sg: scene graph. \emph{wn\textsubscript{path,sim}}: WordNet path similarity, \emph{tfoid}: Faster-RCNN trained on Open Images, \emph{hier\textsubscript{kn}}: Open Images hierarchical knowledge, \emph{visgen}: Faster-RCNN trained on Visual Genome.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (Var) %</th>
<th>UB %</th>
</tr>
</thead>
<tbody>
<tr>
<td>No training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tfoid+CC+PL [38]</td>
<td>50.49 (5.37)</td>
<td>57.81</td>
</tr>
<tr>
<td>tfoid [38]</td>
<td>44.69</td>
<td>50.04</td>
</tr>
</tbody>
</table>

| No training – ours                         |             |      |
| tfoid (reimpl. of [38])                    | 46.08 (7.02)| 61.17|
| tfoid + sg                                 | 46.62 (6.50)| 61.17|
| tfoid + hier\textsubscript{kn}            | 46.82 (6.73)| 61.17|
| tfoid + hier\textsubscript{kn} + \emph{wn\textsubscript{path,sim}} | 47.74 (6.78)| 61.17|
| tfoid + sg + \emph{wn\textsubscript{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 + \emph{wn\textsubscript{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       | ≥ 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|

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 \emph{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.$^5$

\emph{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.

\emph{tfoid+sg} includes labels (but no bounding boxes beyond the \emph{tfoid} detections) and relationships corresponding to the neighbours in the scene graph and \emph{tfoid+sg+\emph{wn\textsubscript{path,sim}}} extends \emph{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.

\emph{tfoid + hier\textsubscript{kn}} performs grounding with the visual representation without scene graph information, but enriched with the Open Images label hierarchy (in Fig. 4) and \emph{tfoid + hier\textsubscript{kn}+\emph{wn\textsubscript{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:

\emph{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 \emph{visgen+sg+\emph{wn\textsubscript{path,sim}}}.

Our best performing model in the \emph{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 \emph{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

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

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

\footnotetext[7]{With our own reimplementation. Compared to [38]’s published results using \emph{tfoid} as single detector we observe an increase of 3.5 pp. accuracy.}
Table 2. Accuracy per category on Flickr30k Entities. \textit{var}: variance over categories, \textit{sg}: scene graph, \textit{wn}_\text{path} \textit{sim}: WordNet path similarity.

<table>
<thead>
<tr>
<th>Method</th>
<th>people</th>
<th>clothing</th>
<th>bodyparts</th>
<th>animals</th>
<th>vehicles</th>
<th>instruments</th>
<th>scene</th>
<th>other</th>
<th>overall</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>tfoid</td>
<td>66.94</td>
<td>36.95</td>
<td>21.03</td>
<td>81.27</td>
<td>82.02</td>
<td>57.41</td>
<td>18.73</td>
<td>25.62</td>
<td>46.09</td>
<td>7.02</td>
</tr>
<tr>
<td>tfoid+sg+wn_path_sim</td>
<td>67.35</td>
<td>43.37</td>
<td>21.80</td>
<td>81.47</td>
<td>82.27</td>
<td>57.41</td>
<td>21.45</td>
<td>26.95</td>
<td>47.92</td>
<td>6.53</td>
</tr>
<tr>
<td>visgen</td>
<td>67.68</td>
<td>56.98</td>
<td>34.76</td>
<td>73.75</td>
<td>73.65</td>
<td>6.17</td>
<td>60.23</td>
<td>38.29</td>
<td>56.30</td>
<td>5.51</td>
</tr>
<tr>
<td>visgen+sg+wn_path_sim</td>
<td>67.92</td>
<td>57.11</td>
<td>34.61</td>
<td>73.75</td>
<td>73.65</td>
<td>8.02</td>
<td>60.69</td>
<td>38.52</td>
<td>57.08</td>
<td>5.31</td>
</tr>
</tbody>
</table>

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

Impact of the contextualised visual representation. In Table 1 we show with \textit{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 \textit{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 \textit{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 \textit{men}, generating the right proposal for the query \textit{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 \textit{tfoid+\_hier}_{kn}, we show that already a small amount of linguistic knowledge can boost the performance (see Fig. 5 for examples). Our methods \textit{tfoid+\_hier}_{kn}+\textit{wn\_path\_sim}, \textit{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.

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.

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).

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>mdR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DVSA CVPR15 [17]</td>
<td>15.2</td>
<td>57.7</td>
<td>50.5</td>
<td>9.2</td>
</tr>
<tr>
<td>SCO CVPR15 [14]</td>
<td>56.7</td>
<td>87.5</td>
<td>94.8</td>
<td>-</td>
</tr>
<tr>
<td>ViLBERT NeurIPS19 [25]</td>
<td>58.2</td>
<td>84.9</td>
<td>91.5</td>
<td>-</td>
</tr>
<tr>
<td>No training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tfoid</td>
<td>8.7</td>
<td>19.8</td>
<td>27.5</td>
<td>52</td>
</tr>
<tr>
<td>visgen</td>
<td>14.6</td>
<td>34.0</td>
<td>43.7</td>
<td>17</td>
</tr>
<tr>
<td>visgen + wn_path_sim</td>
<td>15.3</td>
<td>33.9</td>
<td>43.7</td>
<td>16</td>
</tr>
</tbody>
</table>

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 un-

References

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


