Grounding of Textual Phrases in Images by Reconstruction

Anna Rohrbach[1(✉)], Marcus Rohrbach[2,3], Ronghang Hu[2], Trevor Darrell[2], and Bernt Schiele[1]

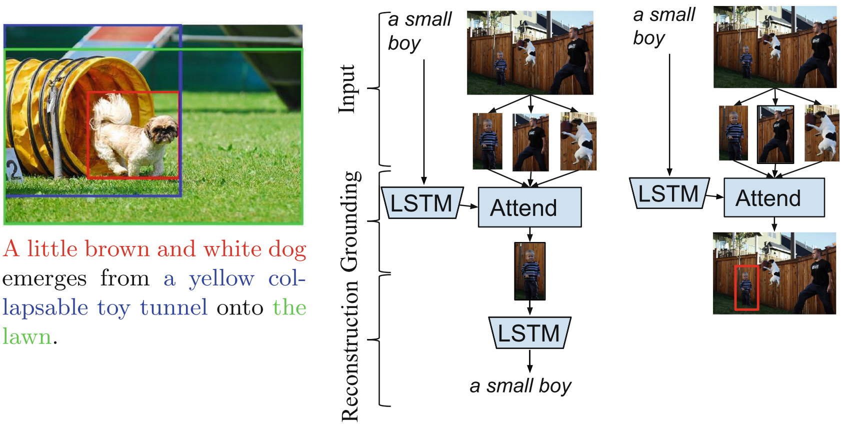
[1] Max Planck Institute for Informatics, Saarbrücken, Germany
{arohrbach,schiele}@mpi-inf.mpg.des
[2] UC Berkeley EECS, Berkeley, CA, USA
{rohrbach,ronghang,trevor}@eecs.berkeley.edu
[3] ICSI, Berkeley, CA, USA

Abstract. Grounding (i.e. localizing) arbitrary, free-form textual phrases in visual content is a challenging problem with many applications for human-computer interaction and image-text reference resolution. Few datasets provide the ground truth spatial localization of phrases, thus it is desirable to learn from data with no or little grounding supervision. We propose a novel approach which learns grounding by reconstructing a given phrase using an attention mechanism, which can be either latent or optimized directly. During training our approach encodes the phrase using a recurrent network language model and then learns to attend to the relevant image region in order to reconstruct the input phrase. At test time, the correct attention, i.e., the grounding, is evaluated. If grounding supervision is available it can be directly applied via a loss over the attention mechanism. We demonstrate the effectiveness of our approach on the Flickr30k Entities and ReferItGame datasets with different levels of supervision, ranging from no supervision over partial supervision to full supervision. Our supervised variant improves by a large margin over the state-of-the-art on both datasets.

1 Introduction

Language grounding in visual data is an interesting problem studied both in computer vision [18,24,25,28,35] and natural language processing [29,34] communities. Such grounding can be done on different levels of granularity: from coarse, e.g. associating a paragraph of text to a scene in a movie [41,52], to fine, e.g. localizing a word or phrase in a given image [18,35]. In this work we focus on the latter scenario. Many prior efforts in this area have focused on rather constrained settings with a small number of nouns to ground [28,31]. On the contrary, we want to tackle the problem of grounding arbitrary natural language phrases in images. Most parallel corpora of sentence/visual data do not provide localization annotations (e.g. bounding boxes) and the annotation process is costly. We propose an approach which can learn to localize phrases relying only on phrases associated with images without bounding box annotations but

B. Leibe et al. (Eds.): ECCV 2016, Part I, LNCS 9905, pp. 817–834, 2016.
DOI: 10.1007/978-3-319-46448-0_49


A little brown and white dog emerges from a yellow collapsable toy tunnel onto the lawn.
Input
Grounding
Reconstruction
a small boy
LSTM
Attend
LSTM
a small boy
a small boy
LSTM
Attend

(a) Predicted grounding. (b) Training time. (c) Test time.

Fig. 1. (a) Without bounding box annotations at training time our approach GroundeR can ground free-form natural language phrases in images. (b) During training our latent attention approach reconstructs phrases by learning to attend to the correct box. (c) At test time, the attention model infers the grounding for each phrase. For semi-supervised and fully supervised variants see Fig. 2.

which is also able to incorporate phrases with bounding box supervision when available (see Fig. 1).

The main idea of our approach is shown in Fig. 1(b, c). Let us first consider the scenario where no localization supervision is available. Given images paired with natural language phrases we want to localize these phrases with a bounding box in the image (Fig. 1c). To do this we propose a model (Fig. 1b) which learns to attend to a bounding box proposal and, based on the selected bounding box, reconstructs the phrase. As the second part of the model (Fig. 1b, bottom) is able to predict the correct phrase only if the first part of the model attended correctly (Fig. 1b, top), this can be learned without additional bounding box supervision. Our method is based on *Ground*ing with a *R*econstruction loss and hence named *GroundeR*. Additional supervision is integrated in our model by adding a loss function which directly penalizes incorrect attention before the reconstruction step. At test time we evaluate whether the model attends to the correct bounding box.

We propose a novel approach to grounding of textual phrases in images which can operate in all supervision modes: with no, a few, or all grounding annotations available. We evaluate our GroundeR approach on the Flickr30k Entities [35] and ReferItGame [26] datasets and show that our unsupervised variant is better than prior work and our supervised approach significantly outperforms state-of-the-art on both datasets. Interestingly, our semi-supervised approach can effectively exploit small amounts of labeled data and surpasses the supervised variant by exploiting multiple losses.

2 Related Work

Grounding Natural Language in Images and Video. For grounding language in images, the approach of [28] is based on a Markov Random Field which aligns 3D cuboids to words. However it is limited to nouns of 21 object classes relevant to indoor scenes. [22] uses a Conditional Random Field to ground the specifically designed scene graph query in the image. [25] grounds dependency-tree relations to image regions using Multiple Instance Learning and a ranking objective. [24] simplifies this objective to just the maximum score and replaces the dependency tree with a learned recurrent network. Both works have not been evaluated for grounding, but we discuss a quantitative comparison in Sect. 4. Recently, [35] presented a new dataset, Flickr30k Entities, which augments the Flickr30k dataset [49] with bounding boxes for all noun phrases present in textual descriptions. [35] report the localization performance of their proposed CCA embedding [14] approach. [45] proposes Deep Structure-Preserving Embedding for image-sentence retrieval and also applies it to phrase localization, formulated as ranking problem. The Spatial Context Recurrent ConvNet (SCRC) [18] and the approach of [33] use a caption generation framework to score the phrase on the set of proposal boxes, to select the box with highest probability. One advantage of our approach over [18,33] is its applicability to un- and semi-supervised training regimes. We believe that our approach of encoding the phrase optimizes the better objective for grounding than scoring the phrase with a text generation pipeline as in [18,33]. As for the fully-supervised regime we empirically show our advantage over [18]. [36] attempts to localize relation phases of type Subject-Verb-Object at a large scale in order to verify their correctness, while relying on detectors from [8].

In the video domain some of the representative works on spatial-temporal language grounding are [31] and [50]. These are limited to small set of nouns.

Object co-localization focuses on discovering and detecting an object in images or videos without any bounding box annotation, but only from image/video level labels [3,6,23,30,38,40,51]. These works are similar to ours with respect to the amount of supervision, but they focus on a few discrete classes, while our approach can handle arbitrary phrases and allows for localization of novel phrases. There are also works that propose to train detectors for a wide range of concepts using image-level annotated data from web image search, e.g. [4,8]. These approaches are complementary to ours in the sense of obtaining large scale concept detectors with little supervision, however they do not tackle complex phrases e.g. "a blond boy on the left" which is the focus of our work.

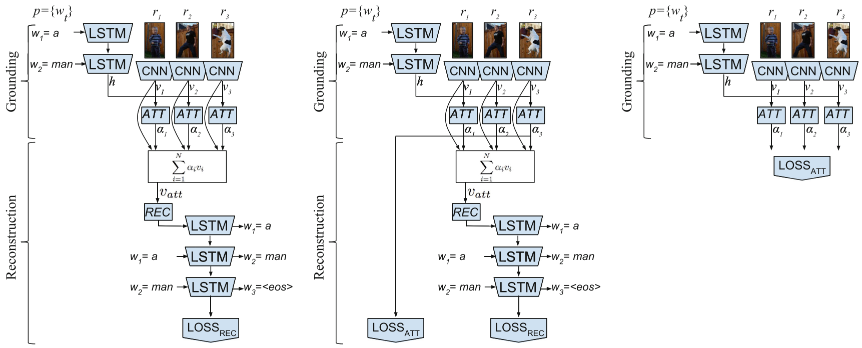
Attention in Vision Tasks. Recently, different attention mechanisms have been applied to a range of computer vision tasks. The general idea is that given a visual input, e.g. set of features, at any given moment we might want to focus only on part of it, e.g. attend to a specific subset of features [2]. [46] integrates spatial attention into their image captioning pipeline. They consider two variants: "soft" and "hard" attention, meaning that in the latter case the model is only allowed to pick a single location, while in the first one the attention "weights" can be distributed over multiple locations. [21] adapts the soft-attention mechanism and attends to bounding box proposals, one word at a time, while generating an image captioning. [47] relies on a similar mechanism to perform temporal attention for selecting frames in video description task. [48] uses attention mechanism to densely label actions in a video sequence. Our approach relies on soft-attention mechanism, similar to the one of [46]. We apply it to the language grounding task where attention helps us to select a bounding box proposal for a given phrase.

Bi-directional Mapping. In our model, a phrase is first mapped to a image region through attention, and then the image region is mapped back to phrase during reconstruction. There is conceptual similarity between previous work and ours on the idea of bi-directional mapping from one domain to another. In autoencoders [43], input data is first mapped to a compressed vector during encoding, and then reconstructed during decoding. [5] uses a bi-directional mapping from visual features to words and from words to visual features in a recurrent neural network model. The idea is to generate descriptions from visual features and then to reconstruct visual features given a description. Similar to [5], our model can also learn to associate input text with visual features, but through attending to an image region rather than reconstructing directly from words. In the linguistic community, [1] proposed a CRF Autoencoder, which generates latent structures for the given language input and then reconstructs the input from these latent structures, with the application to e.g. part-of-speech tagging.

3 GroundeR: ***Ground***ing by ***R***econstruction

The goal of our approach is to ground natural language phrases in images. More specifically, to ground a phrase p in an image I means to find a region r_j in the image which corresponds to this phrase. r_j can be any subset of I, e.g. a segment or a bounding box. The core insight of our method is that there is a bi-directional correspondence between an image region and the phrase describing it. As a correct grounding of a textual phrase should result in an image region which a human would describe using this phrase, i.e. it is possible to reconstruct the phrase based on the grounded image region. Thus, the key idea of our approach is to learn to ground a phrase by reconstructing this phrase from an automatically localized region. Figure 1 gives an overview of our approach.

In this work, we utilize a set of automatically generated bounding box proposals $\{r_i\}_{i \in N}$ for the image I. Given a phrase p, during training our model works in two parts: the first part aims to attend to the most relevant region r_j (or potentially also multiple regions) based on the phrase p, and then the second part tries to reconstruct the same phrase p from region(s) r_j it attended to in the first phase. Therefore, by training to reconstruct the text phrase, the model learns to first ground the phrase in the image, and then generate the phrase from that region. Figure 2a visualizes the network structure. At test time, we



(a) Unsupervised (b) Semi-supervised (c) Fully supervised

Fig. 2. Our model learns grounding of textual phrases in images with (a) no, (b) little (c) or full supervision of localization, through a grounding part and a reconstruction part. During training, the model distributes its attention to a single or several boxes, and learns to reconstruct the input phrase based on the boxes it attends to. At test time, only the grounding part is used.

remove the phrase reconstruction part, and use the first part for phrase grounding. The described pipeline can be extended to accommodate partial supervision, i.e. ground-truth phrase localization. For that we integrate an additional loss into the model, which directly optimizes for correct attention prediction, see Fig. 2b. Finally, we can adapt our model to the fully supervised scenario by removing the reconstruction phase, see Fig. 2c.

In the following we present the details of the two parts in our approach: learning to attend to the correct region for a given phrase and learning to reconstruct the phrase from the attended region. For simplicity, but without loss of generality, we will refer to r_j as a single bounding box.

3.1 Learning to Ground

We frame the problem of grounding a phrase p in image I as selecting a bounding box r_j from a set of image region proposals $\{r_i\}_{i=1,\cdots,N}$. To select the correct bounding box, we define an attention function f_{ATT} and select the box j which receives the maximum attention:

$$j = \arg\max_{i} f_{ATT}(p, r_i) \tag{1}$$

In the following we describe the details of how we model the attention in f_{ATT}. The attention mechanism used in our model is inspired by and similar to the soft attention formulations of [21,46]. However, our inputs to the attention predictor are not single words but rather multi-word phrases, and consequently we also do not have a "doubly stochastic attention" which is used in [46] to normalize the attention across words.

The phrases that we are dealing with might be very complex thus we require a good language model to represent them. We choose a Long Short-Term Memory network (LSTM) [17] as our phrase encoder, as it has been shown effective in various language modeling tasks, e.g. translation [39]. We encode our query phrase word by word with an LSTM and obtain a representation of the phrase using the hidden state h at the final time step as:

$$h = f_{LSTM}(p) \tag{2}$$

Each word w_t in the phrase p is first encoded with a one-hot-vector. Then it is embedded in the lower dimensional space and given to LSTM.

Next, each bounding box r_i is encoded using a convolutional neural network (CNN) to compute the visual feature vector v_i:

$$v_i = f_{CNN}(r_i) \tag{3}$$

Based on the encoded phrase and feature representation of each proposal, we use a two layer perceptron to compute the attention on the proposal r_i:

$$\bar{\alpha}_i = f_{ATT}(p, r_i) = W_2\phi(W_h h + W_v v_i + b_1) + b_2 \tag{4}$$

where ϕ is the rectified linear unit (ReLU): $\phi(x) = max(0, x)$. We found that this architecture performs better than e.g. a single layer perceptron with a hyperbolic tangent nonlinearity used in [2].

We get normalized attention weights α_i by using softmax, which can be interpreted as probability of region r_i being the correct region $r_{\hat{j}}$:

$$\alpha_i = P(i = \hat{j}|\bar{\alpha}) = \frac{\exp(\bar{\alpha}_i)}{\sum_{k=1}^{N}\exp(\bar{\alpha}_k)} \tag{5}$$

If at training time we have ground truth information, i.e. that $r_{\hat{j}}$ is the correct proposal box, then we can compute the loss L_{att} based on our prediction as:

$$L_{att} = -\frac{1}{B}\sum_{b=1}^{B}\log(P(\hat{j}|\bar{\alpha})), \tag{6}$$

where B is the number of phrases per batch. This loss activates only if the training sample has the ground-truth attention value, otherwise, it is zero. If we do not have ground truth annotations then we have to define a loss function to learn the parameters of f_{ATT} in a weakly supervised manner. In the next section we describe how we define this loss by aiming to reconstruct the phrase based on the boxes that are attended to. At test time, we calculate the IOU (intersection over union) value between the selected box r_j and the ground truth box $r_{\hat{j}}$.

3.2 Learning to Reconstruct

The key idea of our phrase reconstruction model is to learn to reconstruct the phrase only from the attended boxes. Given an attention distribution over the boxes, we compute a weighted sum over the visual features and the attention weights α_i:

$$v_{att} = \sum_{i=1}^{N} \alpha_i v_i, \tag{7}$$

which aggregates the visual features from the attended boxes. Then, the visual features v_{att} are further encoded into v'_{att} using a non-linear encoding layer:

$$v'_{att} = f_{REC}(v_{att}) = \phi(W_a v_{att} + b_a) \tag{8}$$

We reconstruct the input phrase based on this encoded visual feature v'_{att} over attended regions. During reconstruction, we use an image description LSTM that takes v'_{att} as input to generate a distribution over phrases p:

$$P(p|v'_{att}) = f_{LSTM}(v'_{att}) \tag{9}$$

where $P(p|v'_{att})$ is a distribution over the phrases conditioned on the input visual feature. Our approach for phrase generation is inspired by [9,44] who have effectively used LSTM for generating image descriptions based on visual features. Given a visual feature, it learns to predict a word sequence $\{w_t\}$. At each time step t, the model predicts a distribution over the next word w_{t+1} conditioned on the input visual feature v'_{att} and all the previous words. We use a single LSTM layer and we feed the visual input only at the first time step. We use LSTM as our phrase encoder as well as decoder. Although one could potentially use other approaches to map phrases into a lower dimensional semantic space, it is not clear how one would do the reconstruction without the recurrent network, given that we have to train encoding and decoding end-to-end.

Importantly, the entire grounding+reconstruction model is trained as a single deep network through back-propagation by maximizing the likelihood of the ground truth phrase $\hat{p}$ generated during reconstruction, where we define the training loss for batch size B:

$$L_{rec} = -\frac{1}{B} \sum_{b=1}^{B} \log(P(\hat{p}|v'_{att})) \tag{10}$$

Finally, in the semi-supervised model we have both losses L_{att} and L_{rec}, which are combined as follows:

$$L = \lambda L_{att} + L_{rec} \tag{11}$$

where parameter λ regulates the importance of the attention loss.

4 Experiments

We first discuss the experimental setup and design choices of our implementation and then present quantitative results on the test sets of Flickr30k Entities (Tables 1 and 2) and ReferItGame (Table 3) datasets. We find our best results to outperform state-of-the-art on both datasets by a significant margin. Figures 3 and 4 show qualitatively how well we can ground phrases in images.

4.1 Experimental Setup

We evaluate GroundeR on the datasets Flickr30k Entities [35] and ReferItGame [26]. Flickr30k Entities [35] contains over 275K bounding boxes from 31K images associated with natural language phrases. Some phrases in the dataset correspond to multiple boxes, e.g. "two men". For consistency with [35], in such cases we consider the union of the boxes as ground truth. We use 1,000 images for validation, 1,000 for testing and 29,783 for training. The ReferItGame [26] dataset contains over 99K regions from 20K images. Regions are associated with natural language expressions, constructed to disambiguate the described objects. We use the bounding boxes provided by [18] and the same test split, namely 10K images for testing; the rest we split in 9K training and 1K validation images.

We obtain 100 bounding box proposals for each image using Selective Search [42] for Flickr30k Entities and Edge Boxes [53] for ReferItGame dataset. For our semi-supervised and fully supervised models we obtain the ground-truth attention by selecting the proposal box which overlaps most with the ground-truth box, while the overlap IOU (intersection over union) is above 0.5. Thus, our fully supervised model is not trained with all available training phrase-box pairs, but only with those where such proposal boxes exist.

On the Flickr30k Entities for the visual representation we rely on the VGG16 network [37] trained on ImageNet [7]. For each box we extract a 4,096 dimensional feature from the fully connected fc7 layer. We also consider a VGG16 network fine-tuned for object detection on PASCAL [10], trained using Fast R-CNN [12]. In the following we refer to both features as VGG-CLS and VGG-DET, respectively. We do not fine-tune the VGG representation for our task to reduce computational and memory load, however, our model trivially allows back-propagation into the image representation which likely would lead to further improvements. For the ReferItGame dataset we use the VGG-CLS features and additional spatial features provided by [18]. We concatenate both and refer to the obtained feature as VGG+SPAT. For the language encoding and decoding we rely on the LSTM variant implemented in Caffe [20] which we initialize randomly and jointly train with the grounding task.

At test time we compute the accuracy as the ratio of phrases for which the attended box overlaps with the ground-truth box by more than 0.5 IOU.

4.2 Design Choices and Findings

In all experiments we use the Adam solver [27], which adaptively changes the learning rate during training. We train our models for about 20/50 epochs for the Flickr30k Entities/ReferItGame dataset, respectively, and pick the best iteration on the validation set.

Next, we report our results for optimizing hyperparameters on the validation set of Flickr30k Entities while using the VGG-CLS features.

Regularization. Applying L2 regularization to parameters (weight decay) is important for the best performance of our unsupervised model. By introducing the weight decay of 0.0005 we improve the accuracy from 20.33 % to 22.96 %. In contrast, when supervision is available, we introduce batch normalization [19] for the phrase encoding LSTM and visual feature, which leads to a performance improvement, in particular from 37.42 % to 40.93 % in the supervised scenario.

Layer Initialization. We experiment with different ways to initialize the layer parameters. The configuration which works best for us is using uniform initialization for LSTM, MSRA [16] for convolutional layers, and Xavier [13] for all other layers. Switching from Xavier to MSRA initialization for the convolutional layers improves the accuracy of the unsupervised model from 21.04 % to 22.96 %.

4.3 Experiments on Flickr30k Entities Dataset

We report the performance of our approach with multiple levels of supervision in Table 1. In the last line of the table we report the proposal upper-bound accuracy, namely the presence of the correct box among the proposals (which overlaps with the ground-truth box with $IOU > 0.5$).

Unsupervised Training. We start with the unsupervised scenario, i.e. no phrase localization ground-truth is used at training time. Our approach, which relies on VGG-CLS features, is able to achieve 24.66 % accuracy. Note that the

Table 1. Phrase localization performance on Flickr30k Entities with different levels of bounding box supervision, accuracy in %.

Approach	Accuracy		
	Other	VGG-CLS	VGG-DET
Unsupervised training			
Deep fragments [6]	21.78	–	–
GroundeR	–	24.66	28.94
Supervised training			
CCA [35]	–	27.42	–
SCRC [18]	–	27.80	–
DSPE [45]	–	–	43.89
GroundeR	–	41.56	47.81
Semi-supervised training			
GroundeR 3.12 % annot.	–	33.02	42.32
GroundeR 6.25 % annot.	–	37.10	44.02
GroundeR 12.5 % annot.	–	38.67	44.96
GroundeR 25.0 % annot.	–	39.31	45.32
GroundeR 50.0 % annot.	–	40.72	46.65
GroundeR 100.0 % annot.	–	42.43	48.38
Proposal upperbound	77.90	77.90	77.90

VGG network trained on ImageNet has not seen any bounding box annotations at training time. VGG-DET, which was fine-tuned for detection, performs better and achieves 28.94 % accuracy. We can further improve this by taking a sentence constraint into account. Namely, it is unlikely that two different phrases from one sentence are grounded to the same box. Thus we post-process the attended boxes: we jointly process the phrases from one sentence and greedily select the highest scoring box for each phrase, while the same box cannot be selected twice. This allows us to reach the accuracy of 25.01 % for VGG-CLS and 29.02 % for VGG-DET. While we currently only use a sentence constraint as a simple post processing step at test time, it would be interesting to include a sentence level constraint during training as part of future work. We compare to the unsupervised Deep Fragments approach of [25]. Note, that [25] does not report the grounding performance and does not allow for direct comparison with our work. With our best case evaluation[1] of Deep Fragments [25], which also relies on detection boxes and features, we achieve an accuracy of 21.78 %. Overall, the ranking objective in [25] can be seen complimentary to our reconstruction objective. It might be possible, as part of future work, to combine both objectives to learn even better models without grounding supervision.

Supervised Training. Next we look at the fully supervised scenario. The accuracy achieved by [35] is 27.42 %[2] and by SCRC [18] is 27.80 %. Recent approach of [45] achieves 43.89 % with VGG-DET features. Our approach, when using VGG-CLS features achieves an accuracy of 41.56 %, significantly improving over prior works that use VGG-CLS. We further improve our result to impressive 47.81 % when using VGG-DET features.

Semi-supervised Training. Finally, we move to the semi-supervised scenario. The notation "x % annot." means that x % of the annotated data (where ground-truth attention is available) is used. As described in Sect. 3.2 we have a parameter λ which controls the weight of the attention loss L_{att} vs. the reconstruction loss L_{rec}. We estimate the value of λ on validation set and fix it for all iterations. We found that we need higher weight on L_{att} when little supervision is available. E.g. for 3.12 % of supervision $\lambda = 200$ and for 12.5 % supervision $\lambda = 50$. This is due to the fact that in these cases only 3.12 %/12.5 % of labeled instances contribute to L_{att}, while all instances contribute to L_{rec}.

[1] We train the Deep Fragments model [25] on the Flickr30k dataset and evaluate with the Flickr30k Entities ground truth phrases and boxes. Our trained Deep Fragments model achieves 11.2 %/16.5 % recall@1 for image annotation/search compared to 10.3 %/16.4 % reported in [25]. As there is a large number of dependency tree fragments per sentence (on average 9.5) which are matched to proposal boxes, rather than on average 3.0 noun phrases per sentence in Flickr30k Entities, we make a best case study in favor of [25]. For each ground-truth phrase we take the maximum overlapping dependency tree fragments (w.r.t. word overlap), compute the IOU between their matched boxes and the ground truth, and take the highest IOU.

[2] The number was provided by the authors of [35], while in [35] they report 25.30 % for phrases automatically extracted with a parser.

Table 2. Detailed phrase localization, Flickr30k Entities, accuracy in %.

Phrase type	People	Clothing	Body parts	Animals	Vehicles	Instruments	Scene	Other	Novel
Number of instances	5,656	2,306	523	518	400	162	1,619	3,374	2,214
Unsupervised training									
GroundeR (VGG-DET)	44.32	9.02	0.96	46.91	46.00	19.14	28.23	16.98	25.43
Supervised training									
CCA embedding [35]	29.58	24.20	10.52	33.40	34.75	35.80	20.20	20.75	n/a
GroundeR (VGG-CLS)	53.80	34.04	7.27	49.23	58.75	22.84	52.07	24.13	34.28
GroundeR (VGG-DET)	61.00	38.12	10.33	62.55	68.75	36.42	58.18	29.08	40.83
Semi-supervised training									
GroundeR (VGG-DET) 3.12 % annot.	56.51	29.84	9.18	57.34	59.75	28.40	50.71	24.48	34.28
GroundeR (VGG-DET) 100.0 % annot.	60.24	39.16	14.34	64.48	67.50	38.27	59.17	30.56	42.37
Proposal upperbound	85.93	66.70	41.30	84.94	89.00	70.99	91.17	69.29	79.90

When integrating 3.12 % of the available annotated data into the model we significantly improve the accuracy from 24.66 % to 33.02 % (VGG-CLS) and from 28.94 % to 42.32 % (VGG-DET). The accuracy further increases when providing more annotations, reaching 42.43 % for VGG-CLS and 48.38 % for VGG-DET when using all annotations. As ablation of our semi-supervised model we evaluated the supervised model while only using the respective x % of annotated data. We observed consistent improvement of our semi-supervised model over the supervised model. Interestingly, when using all available supervision, L_{rec} still helps to improve performance over the supervised model (42.43 % vs. 41.56 %, 48.38 % vs. 47.81 %). Our intuition for this is that L_{att} only has a single correct bounding box (which overlaps most with the ground truth), while L_{rec} can also learn from overlapping boxes with high but not best overlap.

Results per Phrase Type. Flickr30k Entities dataset provides a "type of phrase" annotation for each phrase, which we analyze in Table 2. Our unsupervised approach does well on phrases like "people", "animals", "vehicles" and worse on "clothing" and "body parts". This could be due to confusion between people and their clothing or body parts. To address this, one could jointly model the phrases and add spatial relations between them in the model. Body parts are also the most challenging type to detect, with the proposal upper-bound of only 41.3 %. The supervised model with VGG-CLS features outperforms [35] in all types except "body parts" and "instruments", while with VGG-DET it is better or similar in all types. Semi-supervised model brings further significant performance improvements, in particular for "body parts". In the last column we report the accuracy for novel phrases, i.e. the ones which did not appear in the training data. On these phrases our approach maintains high performance, although it is lower than the overall accuracy. This shows that learned language representation is effective and allows transfer to unseen phrases.

Summary Flickr30k Entities. Our unsupervised approach performs similar (VGG-CLS) or better (VGG-DET) than the fully supervised methods of [18,35] (Table 1). Incorporating a small amount of supervision (e.g. 3.12 % of annotated data) allows us to outperform [18,35] also when VGG-CLS features are used. Our best supervised model achieves 47.81 %, surpassing all the previously reported results, including [45]. Our semi-supervised model efficiently exploits the reconstruction loss L_{rec} which allows it to outperform the supervised model.

4.4 Experiments on ReferItGame Dataset

Table 3 summarizes results on the ReferItGame dataset. We compare our approach to the previously introduced fully supervised method SCRC [18], as well as provide reference numbers for two other baselines: LRCN [9] and CAFFE-7K [15] reported in [18]. The LRCN baseline of [18] is using the image captioning model LRCN [9] trained on MSCOCO [32] to score how likely the query phrase is to be generated for the proposal box. CAFFE-7K is a large scale object classifier trained on ImageNet [7] to distinguish 7K classes. [15] predicts a class for each proposal box and constructs a word bag with all the synonyms of the class-name based on WordNet [11]. The obtained word bag is then compared to the query phrase after both are projected to a joint vector space. Both approaches are unsupervised w.r.t. the phrase bounding box annotations. Table 3 reports the results of our approach with VGG, as well as VGG+SPAT features of [18].

Table 3. Phrase localization performance on ReferItGame with different levels of bounding box supervision, accuracy in %.

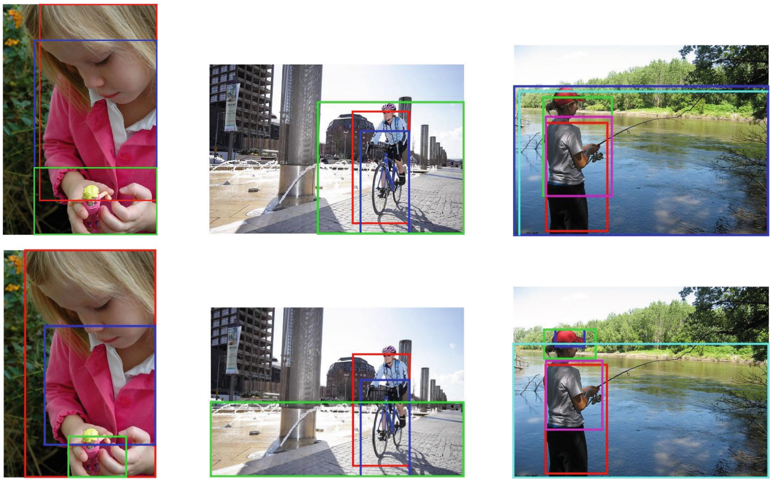
Approach	Accuracy		
	Other	VGG	VGG+SPAT
Unsupervised training			
LRCN [9] (reported in [18])	8.59	–	–
CAFFE-7K [15] (reported in [18])	10.38	–	–
GroundeR	–	10.69	10.70
Supervised training			
SCRC [18]	–	–	17.93
GroundeR	–	23.44	26.93
Semi-supervised training			
GroundeR 3.12 % annot.	–	13.70	15.03
GroundeR 6.25 % annot.	–	16.19	19.53
GroundeR 12.5 % annot.	–	19.02	21.65
GroundeR 25.0 % annot.	–	21.43	24.55
GroundeR 50.0 % annot.	–	22.67	25.51
GroundeR 100.0 % annot.	–	24.18	28.51
Proposal upperbound	59.38	59.38	59.38

Unsupervised Training. In the unsupervised scenario our GroundeR performs competitive with the LRCN and CAFFE-7K baselines, achieving 10.7 % accuracy. We note that in this case VGG and VGG+SPAT perform similarly.

Supervised Training. In the supervised scenario we compare to the best prior work on this dataset, SCRC [18], which reaches 17.93 % accuracy. Our supervised approach, which uses identical visual features, significantly improves this performance to 26.93 %.

Semi-supervised Training. Moving to the semi-supervised scenario again demonstrates performance improvements, similar to the ones observed on Flickr30k Entities dataset. Even the small amount of supervision (3.12 %) significantly improves performance to 15.03 % (VGG+SPAT), while with 100 % of annotations we achieve 28.51 %, outperforming the supervised model.

Summary ReferItGame Dsataset. While the unsupervised model only slightly improves over prior work, the semi-supervised version can effectively learn from few labeled training instances, and with all supervision it achieves 28.51 %, improving over [18] by a large margin of 10.6 %. Overall the performance on ReferItGame dataset is significantly lower than on Flickr30k Entities. We attribute this to two facts. First, the training set of ReferItGame is rather small compared to Flickr30k (9k vs. 29k images). Second, the proposal upperbound on ReferItGame is significantly lower than on Flickr30k Entities (59.38 % vs 77.90 %) due to the complex nature of the described objects and "stuff" image regions.



A little girl in a pink shirt is looking at a toy doll.

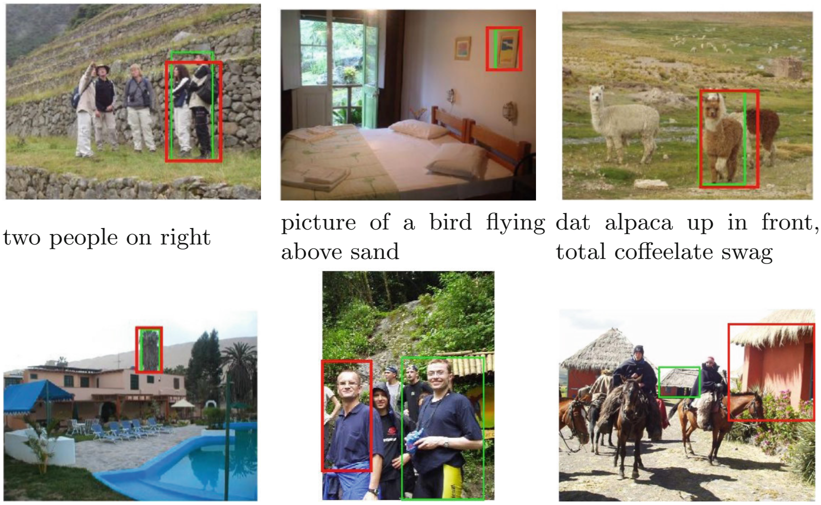
A woman is riding a bicycle on the pavement.

A girl with a red cap, hair tied up and a gray shirt is fishing in a calm lake.

Fig. 3. Qualitative results on the test set of Flickr30k Entities. Top: GroundeR (VGG-DET) unsupervised, bottom: GroundeR (VGG-DET) supervised.

4.5 Qualitative Results

We provide qualitative results on Flickr30K Entities dataset in Fig. 3. We compare our unsupervised and supervised approaches, both with VGG-DET features. The supervised approach visibly improves the localization quality over the unsupervised approach, which nevertheless is able to localize many phrases correctly. Figure 4 presents qualitative results on ReferItGame dataset. We show the predictions of our supervised approach, as well as the ground-truth boxes. One can see the difficulty of the task from the presented examples, including two failures in the bottom row. One requires good language understanding in order to correctly ground such complex phrases. In order to ground expressions like "hut to the nearest left of the person on the right" we would need to additionally model relations between objects, an interesting direction for future work.



two people on right

picture of a bird flying above sand

dat alpaca up in front, total coffeelate swag

palm tree coming out of the top of the building

guy with blue shirt and yellow shorts

hut to the nearest left of the person on the right

Fig. 4. Qualitative results on the test set of ReferItGame: GroundeR (VGG+SPAT) supervised. Green: ground-truth box, red: predicted box. (Color figure online)

5 Conclusion

In this work we address the challenging task of grounding unconstrained natural phrases in images. We consider different scenarios of available bounding box supervision at training time, namely none, little, and full supervision. We propose a novel approach, GroundeR, which learns to localize phrases in images by attending to the correct box proposal and reconstructing the phrase and is able to operate in all of these supervision scenarios. In the unsupervised scenario we are competitive or better than related work. Our semi-supervised approach works well with a small portion of available annotated data and takes advantage of the unsupervised data to outperform purely supervised training using the same amount of labeled data. It outperforms state-of-the-art, both on Flickr30k Entities and ReferItGame dataset, by 4.5 % and 10.6 %, respectively.

Our approach is rather general and it could be applied to other regions such as segmentation proposals instead of bounding box proposals. Possible extensions are to include constraints within sentences at training time, jointly reason about multiple phrases, and to take into account spatial relations between them.

Acknowledgements. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD). This work was supported by DARPA, AFRL, DoD MURI award N000141110688, NSF awards IIS-1427425 and IIS-1212798, and the Berkeley Artificial Intelligence Research (BAIR) Lab.

References

1. Ammar, W., Dyer, C., Smith, N.A.: Conditional random field autoencoders for unsupervised structured prediction. In: Advances in Neural Information Processing Systems (NIPS) (2014)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (ICLR) (2015)
3. Blaschko, M., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: Advances in Neural Information Processing Systems (NIPS), pp. 235–243 (2010)
4. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
5. Chen, X., Zitnick, C.L.: Mind's eye: a recurrent visual representation for image caption generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
6. Cinbis, R.G., Verbeek, J., Schmid, C.: Multi-fold MIL training for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
8. Divvala, S., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
9. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) challenge. Int. J. Comput. Vis. (IJCV) **88**(2), 303–338 (2010)

11. Fellbaum, C.: WordNet: An Electronical Lexical Database. The MIT Press, Cambridge (1998)
12. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
14. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 529–545. Springer, Switzerland (2014)
15. Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., Darrell, T.: Open-vocabulary object retrieval. In: Robotics: Science and Systems (2014)
16. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
18. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
19. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 (2015)
20. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
21. Jin, J., Fu, K., Cui, R., Sha, F., Zhang, C.: Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv:1506.06272 (2015)
22. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3668–3678 (2015)
23. Joulin, A., Tang, K., Fei-Fei, L.: Efficient image and video co-localization with Frank-Wolfe algorithm. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 253–268. Springer, Heidelberg (2014)
24. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
25. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems (NIPS) (2014)
26. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referit game: referring to objects in photographs of natural scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
27. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
28. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? Text-to-image coreference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3558–3565. IEEE (2014)

29. Krishnamurthy, J., Kollar, T.: Jointly learning to parse and perceive: connecting natural language to the physical world. Trans. Assoc. Comput. Linguist. (TACL) **1**, 193–206 (2013)
30. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
31. Lin, D., Fidler, S., Kong, C., Urtasun, R.: Visual semantic search: retrieving videos via complex textual queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2657–2664. IEEE (2014)
32. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Switzerland (2014)
33. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
34. Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., Fox, D.: A joint model of language and perception for grounded attribute learning. In: Proceedings of the International Conference on Machine Learning (ICML) (2012)
35. Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
36. Sadeghi, F., Divvala, S.K., Farhadi, A.: Viske: visual knowledge extraction and question answering by visual verification of relation phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
38. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. arXiv:1403.1024 (2014)
39. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 3104–3112 (2014)
40. Tang, K., Joulin, A., Li, L.J., Fei-Fei, L.: Co-localization in real-world images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2014)
41. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: Book2movie: aligning video scenes with book chapters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1827–1835 (2015)
42. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. (IJCV) **104**(2), 154–171 (2013)
43. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the International Conference on Machine Learning (ICML) (2008)
44. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
45. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

46. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (ICML) (2015)
47. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
48. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: dense detailed labeling of actions in complex videos. arXiv:1507.05738 (2015)
49. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. **2**, 67–78 (2014)
50. Yu, H., Siskind, J.M.: Grounded language learning from video described with sentences. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 53–63 (2013)
51. Yu, H., Siskind, J.M.: Sentence directed video object codetection. arXiv:1506.02059 (2015)
52. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
53. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 391–405. Springer, Switzerland (2014)