Grounding of Textual Phrases in Images by Reconstruction

Anna Rohrbach[1] Marcus Rohrbach[2,3] Ronghang Hu[2]
Trevor Darrell[2] Bernt Schiele[1]

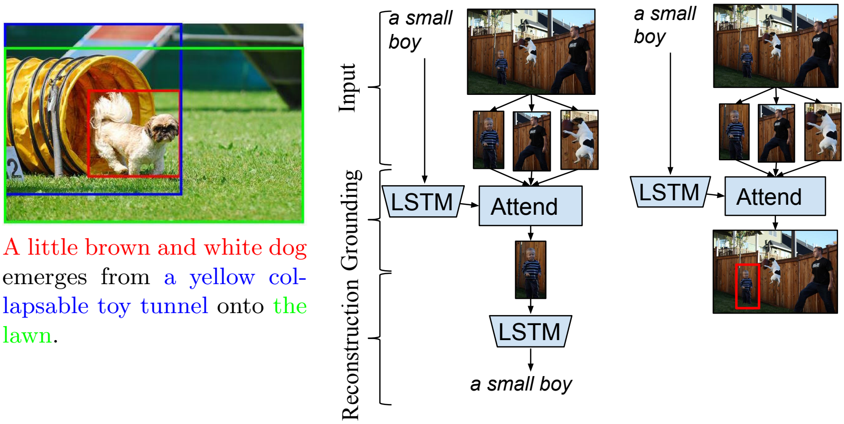
[1]Max Planck Institute for Informatics, Saarbrücken, Germany
[2]UC Berkeley EECS, CA, United States
[3]ICSI, Berkeley, CA, United States

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 model 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 Flickr 30k Entities [1] and ReferItGame [2] 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 [3–5, 1, 6] and natural language processing communities [7, 8]. 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 [9, 10], to fine, e.g. localizing a word or phrase in a given image [1, 6]. 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 [11, 5]. 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 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



(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 right box. (c) At test time, the attention model infers the correct grounding for each phrase. For semi-supervised and fully supervised variants see Fig. 2.

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.

The paper makes the following contributions. Our approach to grounding of textual phrases in images is novel in that it can operate in all supervision modes: with no, a few, or all grounding annotations available. We evaluate the performance of our method on the Flickr 30k Entities [1] and ReferItGame [2] datasets and show that: a) our unsupervised approach is competitive or better than the state-of-the-art (both unsupervised and supervised) b) our semi-supervised approach can effectively exploit small amounts of labeled data; c) our supervised approach significantly outperforms state-of-the-art on both datasets.

2 Related work

Grounding natural language in images and video. For grounding language in images, the approach of [5] 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. [12] uses a Conditional Random Field to ground the specifically designed scene graph query in the image. [4] grounds dependency-tree relations to image regions using Multiple Instance Learning and a ranking objective. [3] 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 Section 5. Recently, [1] presented a new dataset, Flickr30k Entities, which augments the Flickr30k dataset [13] with bounding boxes for all noun phrases present in textual descriptions. [1] report the localization performance of their proposed CCA embedding [14] approach. The Spatial Context Recurrent ConvNet (SCRC) [6] and the approach of [15] use a caption generation framework to score the phrase on the set of proposal boxes, to select the box with highest probability. We compare favourably to [6] in our experimental evaluation. [16] attempts to localize relation phases of type Subject-Verb-Object at a rather large scale in order to verify their correctness, while relying on detectors from [17].

In the video domain some of the representative works on spatial-temporal language grounding are [11] and [18]. 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 [19–25]. In [19] a structural SVM model is learned to detect objects from image-level supervision on the presence or absence of an object class. [21] performs co-localization on videos using temporal consistency terms and turns the problem into an integer program using the Frank-Wolfe algorithm. [24] solves the co-localization problem with a model that jointly incorporates similarity, prior and discriminability, and addresses noisy images. 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 [17] and [26]. 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 of the tree” which is the focus of our work.

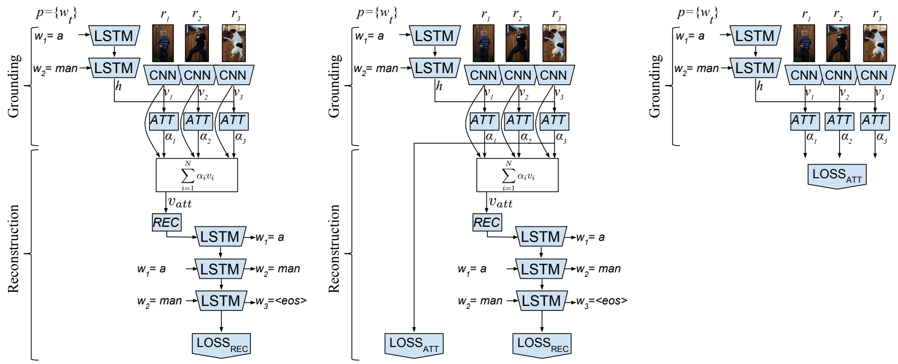
Attention in vision tasks. Recently, different attention mechanisms have been applied to a range of computer vision tasks. The general idea is that given a visual input, e.g. set of features, at any given moment we might want to focus only on part of it, e.g. attend to a specific subset of features [27]. [28] 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. [29] adapts the soft-attention mechanism and attends to bounding box proposals, one word at a time, while generating an image captioning. [30] relies on a similar mechanism to perform temporal attention for selecting frames in video description task. [31] uses attention mechanism to densely label actions in a video sequence. Our approach relies on soft-attention mechanism, similar to the one of [28]. 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 image region(s) through attention, and then the image region(s) are 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 [32], input data is first mapped to a compressed vector during encoding, and then reconstructed during decoding. [33] uses a bi-directional mapping from visual features to words and words to visual features in a recurrent neural network model to generate descriptions based on visual feature, and to reconstruct visual descriptions given a description. Similar to [33], 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, [34] 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, 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. Fig. 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. Fig. 2a visualizes the network structure. At test time, we 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.



(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.

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 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 [29, 28]. 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 [28] 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) [35] as our phrase encoder, as it has been shown effective in various language modeling tasks, e.g. translation [36]. 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_i 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:

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

Based on the encoded phrase and feature representation of each proposal, we use a two layer perceptron to compute the attention $f_{ATT}(h, v_i)$ on the proposal:

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

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 [27].

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

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

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.

During test time, we calculate the IOU (intersection over union) value between the selected box r_j and the ground truth box.

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 $P(\hat{p}|v'_{att})$ is a distribution over the phrases conditioned on the input visual feature. Our approach for phrase generation is inspired by [37, 38] 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 Experimental Setup

We evaluate our approach on the Flickr 30k Entities [1] and ReferItGame [2] datasets as they provide grounded natural language phrases. The Flickr 30k

Entities dataset 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". Similar to [1], 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, as in [1]. The ReferItGame dataset contains over 99K bounding boxes from 20K images. Boxes are associated with natural language phrases (referring expressions), constructed to disambiguate the described objects. We use the same test split as in [6], namely 10K images for testing; additionally we split the rest in 9K training and 1K validation images.

For the language encoding we rely on the LSTM variant implemented in Caffe [39]. Our LSTM is not pre-trained for image description.

We obtain 100 bounding box proposals for each image using Selective Search [40] for Flickr 30k Entities and Edge Boxes [41] 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.

On the Flickr 30k Entities for the visual representation we rely on the VGG16 network [42] trained on ImageNet [43]. For each box we extract a 4,096 dimensional features from the fully connected fc7 layer (last before the classifier). We also consider a VGG16 network fine-tuned for the object detection task on PASCAL dataset [44], trained using Fast R-CNN [45]. In the following we refer to both features as VGG-CLS and VGG-DET respectively, and compare performance of the two. 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 [6]. We concatenate both and refer to the obtained feature as VGG+SPAT.

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 50% IOU.

5 Experiments

We first discuss design choices we made on the validation set of Flickr 30k Entities and then present quantitative results on the test sets of Flickr 30k Entities (Table 1) 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.

5.1 Design choices and findings

In this section we report our results for optimizing hyperparmeters on the validation set of Flickr 30k Entities while using the VGG-CLS features.
Learning. In all experiments we use the Adam learning algorithm [46], which adaptively changes the learning rate during training.

Parameter regularization. We found that 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%.
Layer initialization. We experiment with different ways to initialize the layer parameters. The configuration which works best for us is using Xavier initialization [47] for all layers except LSTM, where we use a uniform initialization as in [37], and convolutional layers, where we use MSRA [48]. This initialization is better adapted to the ReLU non-linearity. Switching from Xavier to MSRA initialization for the convolutional layers improves the accuracy of the unsupervised model from 21.04% to 22.96% on the validation set.
Batch normalization. We introduce batch normalization [49] after the phrase encoding LSTM and visual feature, which leads to a significant improvement, in particular from 37.42% to 40.93% in the supervised scenario.

5.2 Experiments on Flickr 30k Entities dataset

We report the performance of our approach with multiple levels of supervision in Table 1. We compare to the unsupervised Deep Fragments approach of [4]. Note, that [4] does not report the grounding performance and does not allow for direct comparison with our work. [4] focuses on aligning proposal boxes to dependency tree fragments extracted from sentences. We train the Deep Fragments model on the the Flickr30k dataset and evaluate with the Flickr30k Entities ground truth phrases and boxes. We note that the model we trained achieves slightly better retrieval performance than reported in [4][1], but the focus here is the grounding performance which was not evaluated previously. 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 [4]. 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. We also compare to [1] whose approach is based on a Canonical Correlation Analysis (CCA) embedding, learned from phrases and associated visual features. Thus their approach uses the bounding box supervision for training. We compare to another supervised approach, SCRC [6], which scores phrases in a caption generation framework and exploits spatial configuration of boxes as well as global context to score the proposals.

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 VGG

[1] Our trained Deep Fragments model achieves 11.2%/16.5% recall@1 for image annotation/search compared to 10.3%/16.4% reported in [4].

Approach	Accuracy
Unsupervised training	
Deep Fragments [4]	27.11
GroundeR (VGG-CLS)	24.66
GroundeR (VGG-DET)	32.42
Semi-supervised training	
GroundeR (VGG-CLS) 3.12% annot.	33.02
GroundeR (VGG-CLS) 6.25% annot.	37.10
GroundeR (VGG-CLS) 12.5% annot.	38.67
Supervised training	
CCA embedding [1]	25.30
SCRC [6]	27.80
GroundeR (VGG-CLS)	41.56
GroundeR (VGG-DET)	47.70
Proposal upperbound	77.90

Table 1: Phrase localization performance on Flickr 30k Entities with different levels of bounding box supervision, accuracy in %.

network trained on ImageNet has not seen any bounding box annotations at training time. Naturally, we expect that the network fine-tuned for detection would perform better. With the VGG-DET we achieve the improved accuracy of 32.42%. We can further improve the performance by taking into account a sentence constraint. 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 32.77% 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. With our best case evaluation on Deep Fragments [4], which also relies on detection boxes and features, we achieve an accuracy of 27.11%. Overall, the ranking objective in [4] 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.

Semi-supervised training. Now we move to the semi-supervised scenario. Notation "x% annot." means that x% of annotated data (where ground-truth attention is available[2]) is used. As described in Section 3.2 we have a parameter λ which controls the weight of the attention loss vs. the reconstruction loss. We estimate the value of λ on the validation set. Importantly, we found that we have to put much higher weight on the attention loss, and it should be increased as

[2] Note, that less than x% of all the available data is used, as the proposal upperbound (recall) is 77.90%

Phrase type	people	clothing	body-parts	animals	vehicles	instruments	scene	other	all	novel
Unsupervised training										
GroundeR (VGG-CLS)	36.01	9.54	0.76	24.13	32.50	15.43	37.00	13.43	24.66	22.49
Semi-supervised training										
GroundeR (VGG-CLS) 3.12% annot.	45.53	22.77	7.65	39.77	40.50	19.14	37.00	19.38	33.02	27.73
Supervised training										
CCA embedding [1]	29.58	24.20	10.52	33.40	34.75	35.80	20.20	20.75	25.30	n/a
GroundeR (VGG-CLS)	53.80	34.04	7.27	49.23	58.75	22.84	52.07	24.13	41.56	34.28
Proposal upperbound	85.93	66.70	41.30	84.94	89.00	70.99	91.17	69.29	77.90	79.90

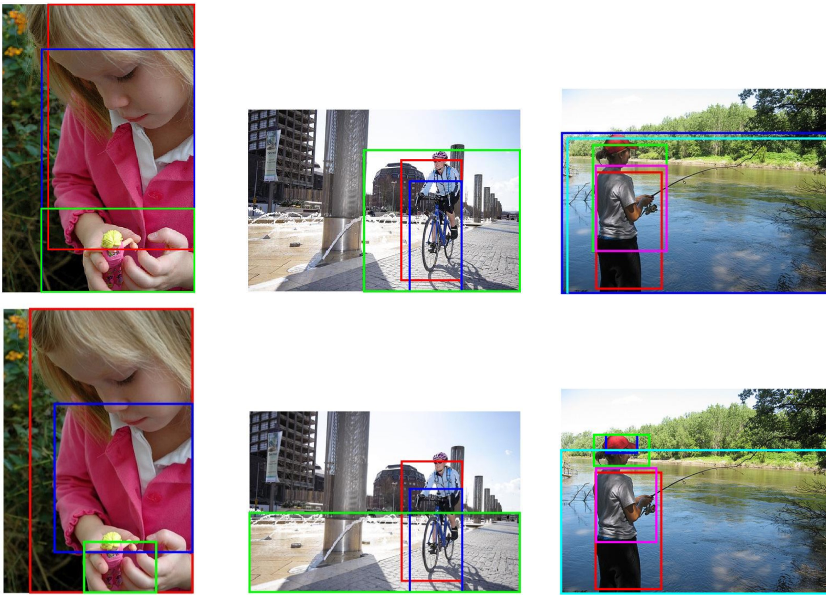
Table 2: Detailed phrase localization, Flickr30k Entities, accuracy in %.

the ratio of annotated data decreases. When integrating 3.12% of the available annotated data into the model we improve the accuracy from 24.66% to 33.02%. The number further increases to 37.10% and 38.68% when adding 6.25% and 12.5% of the annotated data respectively. As comparision, we also evaluate the performance of our fully supervised model when only using the reduced amount of annotated data. It reaches 31.30%, 33.94%, 36.30% in respective regimes, indicating that our semi-supervised model effectively exploits the supervision as well as the unlabeled data.

Supervised training. Finally we come to the fully supervised scenario. The accuracy achieved by [1] is 25.30%[3]. SCRC [6] improves this result to 27.80%. Our approach, when using VGG-CLS features achieves an impressive accuracy of 41.56%, improving over both prior works. We further improve our result when using VGG-DET features, which brings us to 47.70%.

Results per phrase type. Flickr 30k Entities dataset provides a “type of phrase” annotation for each phrase. These types are analyzed in Table 2. Our unsupervised approach achieves rather high performance on phrases like “people”, “animals”, “vehicles” and “scenes”, worse results on “instruments” and “other” and rather low on “clothing” and “body parts”. The reason for the latter is likely a confusion between people and their clothing or body parts. The potential solution for this could be joint modeling of the phrases and incorporating spatial relations between them in the model. Body parts are the most challenging type to detect, as the upper-bound accuracy of the proposals here is the lowest, only 41.3%. Incorporating a small amount of supervision significantly improves performance for all phrase types, among others for “clothing” and “body parts”. The fully supervised model further improves the performance, outperforming [1] in all types except “body parts” and “instruments”. In the last column of Table 2 we report the accuracy for novel phrases, i.e. the ones which did not appear in the training data. As we see, about 15% of test phrases have not been seen during training. On these phrases our approach maintains high performance, although the performance does decrease compared to the overall

[3] [1] states that a parser is used at test time, however, the authors of [1] reported to us that the result on the ground truth phrases (which we use) is 24.78%, where the difference is within the normal variation due to randomness in their approach.



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 Flickr 30k Entities. Top : GroundeR (VGG-DET) unsupervised, bottom: GroundeR (VGG-DET) supervised.

accuracy. This means that learned language representations are effective and allow transfer to unseen phrases.

Summary Flickr30k Entities. Our unsupervised approach performs similar (VGG-CLS) or better (VGG-DET) than the fully supervised methods of [1] and [6] (Table 1). Incorporating a small amount of supervision (e.g. 3.12% of annotated data) allows us to outperform [1] and [6] also when VGG-CLS features are used. Our semi-supervised model with 12.5% of annotated data is performing close to the fully supervised one (38.67% vs. 41.56%). Our best supervised model achieves 47.70%, surpassing the best previously reported results (27.80%) on this dataset by a large margin of 20%.

5.3 Experiments on ReferItGame dataset

Table 3 summarizes our results on the ReferItGame dataset. We compare our approach to the previously introduced fully supervised method SCRC [6], as well as provide the reference numbers for two other baselines: LRCN [50] and CAFFE-7K [51] reported in [6]. The LRCN baseline of [6] is using the image captioning model LRCN [50] trained on MSCOCO [52] to score how likely the query phrase is to be generated for the proposal box. CAFFE-7K is a large scale

Approach	Accuracy
Unsupervised training	
LRCN [50] (reported in [6])	8.59
CAFFE-7K [51] (reported in [6])	10.38
GroundeR (VGG+SPAT)	10.44
Semi-supervised training	
GroundeR (VGG+SPAT) 3.12% annot.	15.03
GroundeR (VGG+SPAT) 6.25% annot.	19.53
GroundeR (VGG+SPAT) 12.5% annot.	21.65
Supervised training	
SCRC [6] (VGG+SPAT)	17.93
GroundeR (VGG+SPAT)	26.93
Proposal upperbound	59.38

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

object classifier trained on ImageNet [43] to distinguish 7K classes. For each proposal box [51] predicts a class and constructs a word bag with all the synonyms of the class-name based on WordNet [53]. 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.

Unsupervised training. In the unsupervised scenario our proposed approach improves over the LRCN baseline, achieving 10.44% accuracy and matches CAFFE-7K, which achieves 10.38%. Overall the performance on the ReferItGame dataset is significantly lower than on the Flickr 30k Entities. We attribute this to three reasons. First, the training set of ReferItGame is rather small compared to Flickr30k (9k vs. 29k images). Second, the phrases in ReferItGame are not only about objects but also about "stuff" and different image regions. Third, the proposal upperbound on the ReferItGame is significantly lower than on Flickr 30k Entities (59.38% vs 77.90%), which is due to the complex nature of the described objects.

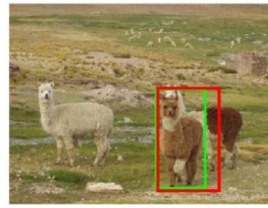
Semi-supervised training. Moving to the semi-supervised scenario again improves the results, even when small amounts of supervision are used, e.g. 15.03% accuracy with 3.12% of the annotated data.

Supervised training. In the supervised scenario we compare to the state-of-the-art method SCRC [6], which reaches 17.93% accuracy. Our supervised approach significantly improves this performance bringing us to 26.93%.

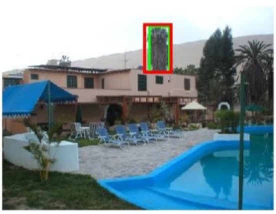
Summary ReferItGame dataset. While the unsupervised only matches prior work, the semi-supervised version can effectively learn from few labeled training instances, and the supervised version achieves 26.93%, improving over [6] by a large margin of 9% points.

two people on right

picture of a bird flying above sand

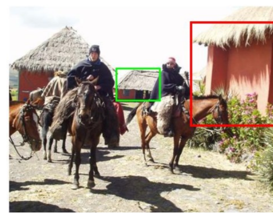


dat alpaca up in front, total coffeelate swag



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.

5.4 Qualitative results

We provide qualitative results on the Flickr 30K Entities dataset in Figure 3. Here we compare our unsupervised and supervised approaches, while both rely on 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 shows qualitative results on the ReferItGame dataset. Here we only showcase the supervised approach, while also showing the ground-truth boxes. One can see the difficulty of the task from the presented examples, including two failure cases in the second row. One clearly requires good language understanding in order to correctly localize such complex phrases.

6 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. Finally, our supervised model outperforms state-of-the-art, both on Flickr 30K Entities and ReferItGame datasets by 20% and 9%, 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 Vision and Learning Center.

References

1. 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)
2. 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)
3. 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)
4. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems (NIPS). (2014)
5. 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), IEEE (2014) 3558–3565
6. 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) arXiv:1511.04164v2.
7. Krishnamurthy, J., Kollar, T.: Jointly learning to parse and perceive: connecting natural language to the physical world. Transactions of the Association for Computational Linguistics (TACL) (2013)
8. 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)
9. 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). (2015) 1827–1835
10. 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. arXiv:1506.06724 (2015)

11. 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), IEEE (2014) 2657–2664
12. 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). (2015) 3668–3678
13. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2** (2014) 67–78
14. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer (2014) 529–545
15. 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)
16. 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)
17. 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)
18. 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). (2013) 53–63
19. Blaschko, M., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: Advances in Neural Information Processing Systems (NIPS). (2010) 235–243
20. 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)
21. Joulin, A., Tang, K., Fei-Fei, L.: Efficient image and video co-localization with frank-wolfe algorithm. In: Proceedings of the European Conference on Computer Vision (ECCV). (2014)
22. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. arXiv:1505.03825 (2015)
23. 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)
24. 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)
25. Yu, H., Siskind, J.M.: Sentence directed video object codetection. arXiv:1506.02059 (2015)
26. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
27. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (ICLR). (2015)

28. 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. arXiv:1502.03044 (2015)
29. 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)
30. 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)
31. 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)
32. 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)
33. 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)
34. Ammar, W., Dyer, C., Smith, N.A.: Conditional random field autoencoders for unsupervised structured prediction. In: Advances in Neural Information Processing Systems (NIPS). (2014)
35. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
36. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2014) 3104–3112
37. 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)
38. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
39. 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, ACM (2014) 675–678
40. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International Journal of Computer Vision (IJCV) **104**(2) (2013)
41. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014, Springer (2014) 391–405
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
43. 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)
44. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision (IJCV) **88**(2) (2010) 303–338

45. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
46. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
47. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International conference on artificial intelligence and statistics. (2010) 249–256
48. 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)
49. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 (2015)
50. 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. arXiv:1411.4389 (2014)
51. Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., Darrell, T.: Open-vocabulary object retrieval. In: Robotics: science and systems. (2014)
52. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). (2014)
53. Fellbaum, C.: WordNet: An Electronical Lexical Database. The MIT Press (1998)