Align2Ground: Weakly Supervised Phrase Grounding Guided by Image-Caption Alignment

Samyak Datta[⋆1,2] Karan Sikka[1] Anirban Roy[1] Karuna Ahuja[1]
Devi Parikh[2,3] Ajay Divakaran[1]
[1]SRI International, Princeton, NJ, [2]Georgia Institute of Technology, [3]Facebook AI Research
[2]{samyak, parikh}@gatech.edu
[1]{karan.sikka, anirban.roy, karuna.ahuja, ajay.divakaran}@sri.com

Abstract

We address the problem of grounding free-form textual phrases by using weak supervision from image-caption pairs. We propose a novel end-to-end model that uses caption-to-image retrieval as a "downstream" task to guide the process of phrase localization. Our method, as a first step, infers the latent correspondences between regions-of-interest (RoIs) and phrases in the caption and creates a discriminative image representation using these matched RoIs. In the subsequent step, this learned representation is aligned with the caption. Our key contribution lies in building this "caption-conditioned" image encoding which tightly couples both the tasks and allows the weak supervision to effectively guide visual grounding. We provide extensive empirical and qualitative analysis to investigate the different components of our proposed model and compare it with competitive baselines. For phrase localization, we report improvements of 4.9% *and* 1.3% *(absolute) over prior state-of-the-art on the VisualGenome and Flickr30k Entities datasets. We also report results that are at par with the state-of-the-art on the downstream caption-to-image retrieval task on COCO and Flickr30k datasets.*

1. Introduction

We focus on the problem of visual grounding which involves connecting natural language descriptions with image regions. Supervised learning approaches for this task entail significant manual efforts in collecting annotations for region-phrase correspondence [29, 39]. Therefore, in this work, we address the problem of grounding free-form textual phrases under weak supervision from only image-caption pairs [20, 37, 40, 45].

A key requirement in such a weakly supervised paradigm is a tight coupling between the task for which supervision is available (image-caption matching) and the task for which we do not have explicit labels (region-phrase matching). This joint reasoning ensures that the supervised loss from the former is able to effectively guide the learning of the latter.

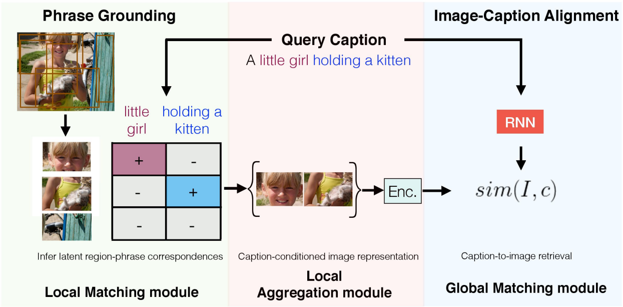


Figure 1: This figure[⋆⋆] shows a high-level overview of the proposed Align2Ground model which learns to ground phrases by using weak supervision from image-caption pairs. It first matches the phrases with local image region, aggregates these matched RoIs to generate a *caption-conditioned* image representation. It uses this encoding to perform image–caption matching.

Recent works [20, 21] have shown evidence that operating under such a paradigm helps boost performance for image-caption matching. Generally, these models consist of two stages: (1) a local matching module that infers the latent region-phrase correspondences to generate local matching information, and (2) a global matching module that uses this information to perform image-caption matching. This setup allows phrase grounding to act as an intermediate and a prerequisite task for image-caption matching. It is important to note that the primary objective of such works has been on image-caption matching and *not* phrase grounding.

An artifact of training under such a paradigm is the amplification of correlations between selective regions and phrases.

[⋆] A major part of this work was done by S. Datta when he was an intern at SRI International, Princeton, NJ

[⋆⋆] "Young girl holding a kitten" by Gennadiy Kolodkin is licensed under CC BY-NC-ND 2.0.

For example, a strong match for even a small subset of phrases in the first stage would translate to a high overall matching score for the image and the entire caption in the second stage. As a consequence, the model is able to get away with not learning to accurately ground *all* phrases in an image. Hence, such a strategy is not an effective solution if the primary aim is visual grounding. Such "cheating" tendencies, wherein models learn to do well at the downstream task without necessarily getting better at the intermediate task, has also been seen in prior works such as [11, 23, 31].

We argue that this "selective amplification" behavior is a result of how the local matching information from the first stage is transferred to the second stage – via average pooling of the RoI–phrase matching scores. We address this limitation by proposing a novel mechanism to relay this information about the latent, inferred correspondences in a manner that enables a much tighter coupling between the two stages. Our primary contribution is the introduction of a *Local Aggregation Module* that takes the subset of region proposals that match with phrases and encodes them to get a *caption-conditioned* image representation that is then used directly by the second stage for image-caption matching (Figure 1). We encode the matched proposal features using a permutation-invariant set encoder to get the image representation. Our novelty lies in designing this effective transfer of information between the supervised and unsupervised parts of the model such that the quality of image representations for the supervised matching task is a direct consequence of the correct localization of all phrases.

Our empirical results indicate that such an enforcement of the proper grounding of all phrases via caption-conditioned image representations (Figure 2) does indeed lead to a better phrase localization performance (Table 3, 4). Moreover, we also show that the proposed discriminative representation allows us to achieve results that are comparable to the state-of-the-art on the downstream image-caption matching task on both COCO and Flickr30k datasets (Table 2). This demonstrates that the proposed caption-conditioned representation not only serves as a mechanism for the supervised loss to be an effective learning signal for phrase grounding, but also does not compromise on the downstream task performance.

The contributions of our paper are summarized as follows.

- We propose a novel method to do phrase grounding by using weak supervision from the image-caption matching task. Specifically, we design a novel *Local Aggregation Module* that computes a *caption-conditioned* image representation, thus allowing a tight coupling between both the tasks.
- We achieve state-of-the-art performance for phrase localization. Our model reports absolute improvements of 4.9% and 1.3% over prior state-of-the-art on Visual Genome and Flickr30k Entities respectively.
- We also report state-of-the-art results on the (downstream) task of caption-to-image retrieval on the Flickr30k dataset and obtain performance which is comparable to the state-of-the-art on COCO.

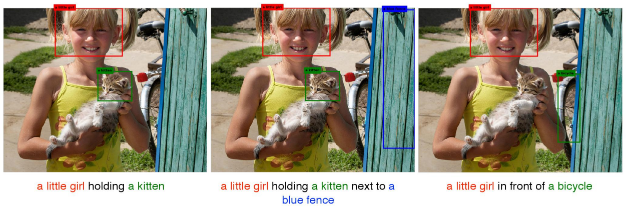

a little girl holding a kitten
a little girl holding a kitten next to a blue fence
a little girl in front of a bicycle

Figure 2: For a given image, we show the regions that match with phrases from three different query captions, as predicted by our model. Our proposed Local Aggregator module computes a caption-conditioned image representation by encoding the features of only the matched image regions. It is evident that in order for this representation to do well at image-caption matching, the grounding of caption-phrases should be proper.

2. Related Work

Visual-Semantic Embeddings [1, 4, 14, 16, 22, 38, 39] have been successfully applied to multimodal tasks such as image-caption retrieval. These methods embed an (entire) image and a caption in a shared semantic space, and employ triplet-ranking loss based objectives to fine-tune the metric space for image-caption matching. [14] further improves this learned, joint embedding space by using techniques such as hard negative sampling, data augmentation, and fine-tuning of the visual features. In addition to these joint embedding spaces, [39] also proposes a similarity network to directly fuse and compute a similarity score for an image-caption pair. In contrast to our proposed model, none of these approaches reason about local structures in the multimodal inputs i.e. words/phrases in sentences and regions in images.

The **Phrase Localization** task involves learning the correspondences between text phrases and image regions from a given training set of region–phrase mappings [29, 34]. A major challenge in these tasks is the requirement of ground-truth annotations which are expensive to collect and prone to human error. Thus, a specific focus of recent work [6, 37, 40] for phrase localization has been on learning with limited or no supervision. For example, [37] learns to leverage the bidirectional correspondence between regions and phrases by reconstructing the phrases from the predicted region proposals. However, their learning signal is only guided by the reconstruction loss in the text domain. [6] improves upon their work by adding consistency in both the visual and the text domains, while also adding external knowledge in the form of distribution of object labels predicted from a pre-trained CNN. As opposed to using hand-annotated phrases (as in the above methods), our model directly makes use of the readily available, aligned image-caption pairs for visual grounding.

Some prior works also use supervision from image-caption training pairs to perform phrase localization [9, 13, 20, 21, 40]. They either rely on using image–caption matching as a downstream task [13, 20, 21] or use the sentence parse-tree structure to guide phrase localization. [13] achieves phrase localization by first learning a joint embedding space, and then generating and aggregating top-k feature maps from the visual encoders (conditioned on the text encoding) to find the best matching spatial region for a query phrase. [9, 40] propose to use the parse tree structure of sentences to provide additional cues to guide the model for phrase localization. Among this family of approaches, our proposed model is conceptually most similar to Karpathy et al. [20]. These methods aggregate local region-phrase alignment scores to compute a global image-caption matching score. As noted in Section 1, such a strategy is able to correctly match an image-caption pair without actually learning to ground all phrases inside the caption leading to a suboptimal visual grounding. Our proposed model tackles this issue by by directly using the matched RoIs to build a discriminative image representation which is able to tightly couple the phrase localization task with the supervised downstream task.

Our work is also closely related to [33] where they not only map images and captions, but also phrases and regions in the same dense visual-semantic embeddings space. In contrast, our model provides a clean disentanglement between the region-phrase and the image-caption matching tasks, where the first stage of our model localizes the phrases and the second stage matches images and captions during training.

3. Approach

We work in a setting where we are provided with a dataset of image-caption pairs for training. We denote an image and a caption from a paired sample as I and c respectively. For each image, we extract a set of R region proposals, also referred to as Regions-of-Interest (RoIs), using a pre-trained object detector. We use a pre-trained deep CNN to compute features for these RoIs and denote them as $\{\boldsymbol{x}_j\}_{j=1}^{R}$, where $\boldsymbol{x}_j \in \mathbb{R}^{d_v}$. We perform a shallow parsing (or chunking) of the caption by using an off-the-shelf parser [10] to obtain P phrases. We encode each phrase using a Recurrent Neural Network (RNN) based encoder, denoted as Φ_{RNN}. We denote the encoded phrases as a sequence $(\boldsymbol{p}_k)_{k=1}^{P}$, where $\boldsymbol{p}_k \in \mathbb{R}^{d_s}$ and k is a positional index for the phrase in the caption. Note that we operate in a weakly supervised regime i.e. during training, we do not assume ground-truth correspondences between image regions and phrases in the caption.

During inference, our primary aim is to perform visual grounding. Given a query phrase p and an image I, the learned model identifies the RoI that best matches with the query. We now describe our proposed approach along with the loss function and the training procedure.

3.1. Align2Ground

We follow the general idea behind prior works that learn to match images and captions by inferring latent alignments between image regions and words/phrases in the caption [20, 21]. These methods operate under the assumption that optimizing for the downstream task of ranking images with respect to captions requires learning to accurately infer the latent alignment of phrases with regions i.e. phrase grounding. Specifically, these methods [20, 21] match image-caption pairs by first associating words in the caption to relevant image regions based on a scoring function. Thereafter, they average these local matching scores to compute the image-caption similarity score, which is used to optimize the loss. We shall refer to these methods as Pooling-based approaches due to their use of the average pooling operation. As discussed earlier, averaging can result in a model that performs well on the image–caption matching task without actually learning to accurately ground *all* phrases.

In contrast to such methods, our proposed model uses a novel technique that builds a discriminative image representation from the matched RoIs and uses this representation for the image-caption matching. Specifically, the image representation that is used to match an image with a caption is conditioned *only on the subset* of image regions that align semantically with all the phrases in that caption. We argue that such an architectural design primes the supervision available from image-caption pairs to be a stronger learning signal for visual grounding as compared to the standard Pooling-based methods. This is a consequence of the explicit aggregation of matched RoIs in our model which strongly couples both the local and global tasks leading to better phrase localization.

Conceptually, our model relies on three components (see Figure 3) to perform the phrase grounding and the matching tasks: (1) The Local Matching Module infers the latent RoI–phrase correspondences for all the phrases in the query caption, (2) The Local Aggregator Module takes the matched RoIs (as per the alignments inferred by the previous module) and computes a *caption-conditioned* representation for the image, and (3) The Global Matching Module takes the caption-conditioned representation of the image and learns to align it with the caption using a ranking loss. We now describe these modules in detail.

The Local Matching module is responsible for inferring the latent correspondences between regions in an image and phrases from a caption. We first embed both RoIs and phrases in a joint embedding space to measure their semantic

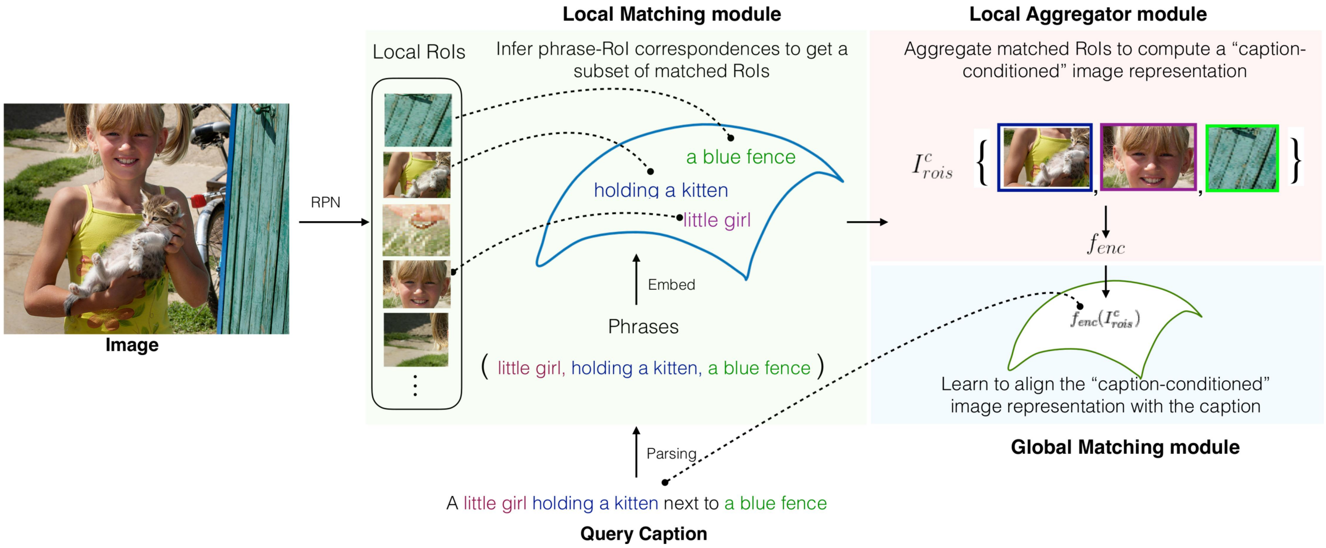

Local Matching module
Local Aggregator module
Local RoIs
Infer phrase-RoI correspondences to get a subset of matched RoIs
Aggregate matched RoIs to compute a "caption-conditioned" image representation
Image
RPN
a blue fence
holding a kitten
little girl
Embed
Phrases
(little girl, holding a kitten, a blue fence)
I^c_{rois}
f_{enc}
$f_{enc}(I^c_{rois})$
Learn to align the "caption-conditioned" image representation with the caption
Global Matching module
Parsing
A little girl holding a kitten next to a blue fence
Query Caption

Figure 3: This figure gives a detailed overview of our proposed architecture. The outputs from a Region Proposal Network (RoIs) and the shallow parser (phrases) are fed into the Local Matching module which infers the latent phrase-RoI correspondences. The Local Aggregator module then digests these matched RoIs to create a discriminative, caption-conditioned visual representation – which is then used to align the image-caption pairs in the Global Matching module.

similarity. To do so, we project the RoI $\boldsymbol{x}_j$ in the same space as the phrase embeddings, $\boldsymbol{p}_k$, via a linear projection. We then measure the semantic similarity, s_{jk}, between region $\boldsymbol{x}_j$ and phrase $\boldsymbol{p}_k$ using cosine similarity.

$$\hat{\boldsymbol{x}_j} = \boldsymbol{W_l}^T \boldsymbol{x}_j \tag{1}$$

$$s_{jk} = \frac{\hat{\boldsymbol{x}_j}^T \boldsymbol{p}_k}{\|\hat{\boldsymbol{x}_j}\|_2 \|\boldsymbol{p}_k\|_2} \tag{2}$$

where $\boldsymbol{W_l} \in \mathbb{R}^{d_v \times d_s}$ is the projection matrix.

A straightforward approach to infer the matched RoI for a phrase is to the select the top scoring box i.e. for a phrase p_k the matched RoI is $\boldsymbol{x}_{j^*}$, where $j^* = \arg\max_j s_{jk}$. However, it has been shown that such a strategy is prone to overfitting since the model often keeps on choosing the same erroneous boxes [5]. We also take inspiration from the recent advances in neural attention, and compute attention weights α_{jk} for each RoI based on a given phrase. We then generate an attended region vector as a linear combination of the RoI embeddings, weighted by the attention weights.

$$\alpha_{jk} = \text{softmax}(s_{jk})_{j=1}^R \quad \boldsymbol{x}_k^c = \sum_j \alpha_{jk} \boldsymbol{x}_j \tag{3}$$

Despite the success of this strategy for other multimodal tasks, we found that it is not an effective solution for the given problem. This is because the discriminativeness of the matched RoI seems to get compromised by the weighted averaging of multiple matched RoIs during training. We instead propose to add diversity to the training procedure by first selecting top-k ($k = 3$) scoring RoI candidates and then randomly selecting one of them as the matched RoI for the query phrase. We observe that this strategy adds more robustness to our model by allowing it to explore diverse options during training.

This module returns a list of caption-conditioned RoIs $I^c_{rois} = (\boldsymbol{x}^c_k)^P_{k=1}$, where $\boldsymbol{x}^c_k$ is the feature vector for the aligned RoI for phrase p_k in caption c.

The Local Aggregator module uses the RoIs matched in the previous step to generate a caption-conditioned representation of the image. In contrast to Pooling-based methods, we explicitly aggregate these RoIs to build a more discriminative encoding for the image. This idea takes inspiration from adaptive distance metric learning based approaches [41] where the learned distance metric (and equivalently, the embedding space) for computing similarities is conditioned on the input query. In our case, the image representation is conditioned on the query caption that we are trying to measure its similarity with.

We propose to use an order-invariant deep encoder to aggregate the RoIs [35, 43]. Our choice is motivated by the assumption of modeling a caption as an orderless collection of phrases. Such an assumption is justified because a match between a set of phrases and image regions should be invariant to the order in which those phrases appear in the caption. These different orders might be generated by say, swapping two noun phrases that are separated by a conjunction such as "and". We implement this encoder, denoted as f_{enc}, by using a two-layer Multilayer Perceptron (MLP) with a *mean* operation [43]. During experiments, we also compare our model with a order-dependent encoding by using a GRU encoder. The caption-conditioned image representation, which encodes of the set of matched RoIs, is then passed onto the next module. The primary contribution of this work is this module that build this caption-conditioned image representation and thus ensures a strong coupling between the (unsupervised) RoI-phrase matching and the supervised image-caption matching task.

The Global Matching module uses the caption-conditioned image encoding obtained from the Local Aggregator module and aligns it with the query caption. We measure similarity between the proposed image representation and the query caption by first embeddings the caption c, encoded by Φ_{RNN}, in the same output space as the image representation by using a two-layer MLP. We then compute cosine similarity between the two multimodal representations.

$$\hat{\boldsymbol{c}} = MLP(\Phi_{RNN}(c)) \quad \hat{\boldsymbol{r_c}} = f_{enc}(I^c_{rois}) \tag{4}$$

$$S_{Ic} = \frac{\hat{\boldsymbol{c}}^T \hat{\boldsymbol{r_c}}}{\|\hat{\boldsymbol{c}}\|_2 \|\hat{\boldsymbol{r_c}}\|_2} \tag{5}$$

S_{IC} is the similarity between image I and caption c.

Loss Function: We train our model with max-margin ranking loss that enforces the score between a caption and a paired image to be higher than a non-paired image and vice-versa. Similar to Faghri et al. [14], we sample the hardest negatives in the mini-batch while generating triplets for the ranking loss.

$$\mathcal{L} = \max_{c' \notin \mathbb{C}_I}(0, m - S_{Ic} + S_{Ic'}) + \max_{I' \notin \mathbb{I}_c}(0, m - S_{Ic} + S_{I'c}) \tag{6}$$

where m is the margin, $\mathbb{C}_I$ is the set of captions paired with image I, and $\mathbb{I}_c$ is the set of images paired with caption c.

4. Experiments

In this section, we discuss the experimental setup used to evaluate our model. We first outline the datasets and the evaluation metrics used to measure performance. We then provide implementation details for our method and a couple of relevant prior works that our model is conceptually related to. Next, we establish the benefits of our model by reporting quantitative results for the phrase localization and the caption-to-image retrieval tasks. We follow that with qualitative results to provide useful insight into the workings of our model. Finally, we compare our model with several state-of-the-art methods on both the tasks of phrase localization and caption-to-image retrieval

4.1. Dataset and Evaluation Metrics

COCO [27] dataset consists of $123,287$ images with 5 captions per image. The dataset is split into $82,783$ training, $5,000$ validation and $5,000$ test images. Following recent works [14, 20, 33], we use the standard splits [20] and augment the training set with $30,504$ images from the validation set, that were not included in the original $5,000$-image validation split.

Flick30k [34, 42] dataset consists of $31,783$ images with 5 captions per image. Additionally, the Flickr30k Entities dataset contains over $275k$ bounding boxes corresponding to phrases from the captions. Following prior work, we also use $1,000$ images each for validation and test set, and use the remaining images for training.

VisualGenome (VG) [24] dataset is used to evaluate phrase localization. We use a subset of images from VG that have bounding box annotations for textual phrases. This subset contains images that are present in both the validation set of COCO and VisualGenome and consists of $17,359$ images with $860,021$ phrases.

Metrics: Since the design of our model uses weak supervision from image–caption pairs to perform phrase localization, we evaluate our model on two tasks– (1) phrase localization, and (2) caption-to-image retrieval (C2I).

For **phrase localization**, we report the percentage of phrases that are correctly localized with respect to the ground-truth bounding box across all images, where correct localization means $IoU \geq 0.5$ [34]. We refer to this metric as phrase localization/detection accuracy ($Det.\%$). Prior works on visual grounding have also demonstrated localization using attention based heat maps [13, 40]. As such, they use a *pointing game* based evaluation metric, proposed in [44], which defines a hit if the center of the visual attention map lies anywhere inside the ground-truth box and reports the percentage accuracy of these hits. We compare our model with these prior works by reporting the same which we refer to as the $PointIt\%$ metric on the Visual Genome and Flickr30k Entities datasets.

For the **C2I** task, we report results using standard metrics– (i) Recall@K ($R@K$) for K = 1, 5 and 10 that measures the percentage of captions for which the ground truth image is among the top-K results retrieved by the model, and (ii) median rank of the ground truth image in the ranked list of images retrieved by the model. For C2I retrieval experiments, we train and evaluate our models using both COCO and Flickr30k datasets.

4.2. Implementation Details

Visual features: We extract region proposals for an image by using Faster-RCNN [36] trained on both objects and attributes from VG, as provided by Andreson et al.[1] [3]. For every image, we select the top 30 RoIs based on Faster-RCNN's class detection score (after non-maximal suppre-

[1] https://github.com/peteanderson80/bottom-up-attention

			COCO					Flickr30k				
			Caption-to-Image retrieval				Phrase	Caption-to-Image retrieval				Phrase
			$R@1$	$R@5$	$R@10$	Med r	$Det.\%$	$R@1$	$R@5$	$R@10$	Med r	$Det.\%$
Global			39.3	74.8	86.3	2	12.2	27.1	56.0	68.4	4	8.0
Pooling-based (words)			47.9	81.7	91.0	2	10.7	40.7	71.2	80.9	2	8.4
Pooling-based (phrases)			48.4	81.7	91.2	2	10.8	41.4	71.4	81.2	2	8.9
Align2Ground												
Proposed model	permInv	max	40.3	76.3	87.8	2	14.5	29.1	60.8	72.7	3	11.5
		topk	56.6	84.9	92.8	1	14.7	49.7	74.8	83.3	2	11.2
		attention	42.8	78.1	89.1	2	10.2	37.9	67.0	77.8	2	6.2
	sequence	max	39.4	75.0	87.1	2	14.5	29.9	60.9	72.7	3	11.5
		topk	58.4	86.1	93.5	1	14.5	47.9	75.6	83.5	2	11.3
		attention	41.9	77.1	88.4	2	9.8	38.2	68.4	78.2	2	5.6

Table 1: Phrase localization and Caption-to-Image retrieval results for models trained on COCO and Flickr30k datasets. Note that we report phrase localization numbers on VisualGenome in all the cases. We compare our proposed model (*permInv-topk*) with two prior methods and with different choices for the Local Matching module (max/topk/attention) and the Local Aggregator module (permInv/sequence) as discussed in Section 3.

sion and thresholding). [2] We then use RoIAlign [17] to extract features ($d_v = 2048$-d) for each of these RoIs using a ResNet-152 model pre-trained on ImageNet [18].

Text features: We perform shallow parsing (also known as chunking) using the SENNA parser [10] to parse a caption into its constituent phrases. Shallow parsing of sentences first identifies the constituent parts (such as nouns, verbs) of a sentence and then combines them into higher-order structures (such as noun-phrases and verb-phrases). In our current work, a phrase generally comprises noun(s)/verb(s) with modifiers such as adjective(s) and/or preposition(s). Additionally, we perform post-processing steps based on some handcrafted heuristics (refer to supplementary for more details) to get the final set of phrases from the captions.

Both phrases and sentences are encoded by using a 2-layer, bidirectional GRU [7] with a hidden layer of size 1024 and using inputs from a 300 dimensional word embeddings. We train the word-embeddings from scratch to allow for a fair comparison with prior work [10, 22]. We also experimented with a variant that uses pre-trained GloVe embeddings and found that the performance is worse than the former.

	COCO			Flickr30k		
	$R@$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
DVSA [20]	27.4	60.2	74.8	15.2	37.7	50.5
UVS [22]	31.0	66.7	79.9	22.0	47.9	59.3
m-RNN [30]	29.0	42.2	77.0	22.8	50.7	63.1
m-CNN [28]	32.6	68.6	82.8	26.2	56.3	69.6
HM-LSTM [33]	36.1	–	86.7	27.7	–	68.8
Order [38]	37.9	–	85.9	–	–	–
EmbeddingNet [39]	39.8	75.3	86.6	29.2	59.6	71.7
sm-LSTM [19]	40.7	75.8	87.4	30.2	60.4	72.3
Beans [13]	55.9	86.9	94.0	34.9	62.4	73.5
2WayNet [12]	39.7	63.3	–	36.0	55.6	–
DAN [32]	–	–	–	39.4	69.2	79.1
VSE++ [14]	52.0	–	92	39.6	–	79.5
SCAN [25]	**58.8**	**88.4**	**94.8**	48.6	**77.7**	**85.2**
Ours	56.6	84.9	92.8	**49.7**	74.8	83.3

Table 2: Comparison with the state-of-the-art on the downstream caption-to-image retrieval task.

Prior Works and Align2Ground: We compare our model with two competing works. The first method, refered to as *Global*, embeds both the image and caption in a joint embedding space and computes their matching score using cosine similarity [14]. We also implement the Pooling-based method [20], that computes similarity between image–caption pairs by summarizing the local region-phrase matching scores. We use our Local Matching module to infer phrase–region correspondences and then average these scores. Following the original implementation by Karpathy et al. [20], we first encode the entire caption using a GRU and then compute the embeddings for each word by using the hidden state at the corresponding word index (within that caption). We refer to this approach as *Pooling-based (words)*. We also implement a variant that uses phrases, as used in our method, instead of words (*Pooling-based (phrases)*). For a fair comparison we use the same image and text encoders for the baselines as well as our model.

To highlight the effectiveness of using the proposed *topk* scheme for the Local Matching module, we compare it against both *attention* and *max* based methods as discussed

[2] We also experimented with other region proposal methods such as EdgeBoxes and Faster-RCNN trained on COCO, but found this to be much better.

Table 3: Phrase Localization on Visual Genome

Random (baseline)	Center (baseline)	Linguistic Structure [40]	Beans In Burgers [13]	**Ours**
17.1	19.5	24.4	33.8	**38.7**

in Section 3. We also compare the orderless pooling scheme proposed for the Local Aggregator module with an order-dependent pooling scheme based on a bidirectional GRU (2-layer, hidden layer of size 256 units). For the orderless pooling scheme we use a 2-layer MLP with a hidden layer of size 256.

We train all models for 60 epochs with a batch size of 32 using the Adam optimizer and a learning rate of 0.0002. We use a margin of 0.1 for the triplet-ranking loss in all our experiments. We select the final checkpoints on the basis of the model's best performance on a small validation set for both localization and C2I tasks. We warm-start our model by initializing it with a model variant that is in spirit similar to the Pooling-based methods (during our experiments, we observed that it otherwise takes a long time for it to converge).

4.3. Quantitative Results

We now report the quantitative performance of prior methods as well as different design choices within the proposed model in Table 1. We start by comparing Pooling-based methods with the Global method. We then discuss the impact of using the proposed matching and aggregating strategies against other choices in our model on the phrase localization and the C2I tasks. We report all results in absolute percentage points.

We observe that the Pooling-based (phrases) model ($R@1 = 48.4$) performs better on the C2I task than the Global baseline ($R@1 = 39.3$) on COCO (with the difference being even higher for Flickr30k). We also note that the Pooling-based (phrases) outperforms its counterpart–Pooling-based (words) that uses words for matching. This shows that for the C2I task, it is better to represent a caption as phrases instead of individual words, as used in this work. We also notice improvements with the use of phrases on the phrase localization task ($Det\%$ $+0.1$ for COCO and $+0.5$ Flickr30k)

An interesting observation is that although the Pooling-based (phrases) method outperforms the Global baseline on the C2I task, its performance on phrase localization is not always better than the latter ($Det\%$ 10.8 vs. 12.2 for COCO and 8.9 vs. 8.0 for Flickr30k). As stated in Section 1, this trend could be explained by the fact that on account of averaging the local matching scores, the Pooling-based methods are able to achieve good results by selectively amplifying correspondences between phrases and image regions (e.g. by assigning high matching scores to visual noun-phrases e.g. "person") without learning to accurately ground all phrases (and ignoring less visual phrases e.g. "sitting") in the caption. Recall that this was one of the motivations that inspired the design of our proposed Align2Ground model.

Table 4: Phrase Localization on Flicr30K Entities

Akbari et al. [2] (prior SoTA)	Fang et al. [15]	Pooling-based (phrases)	**Ours**
69.7	29.0	65.7	**71.0**

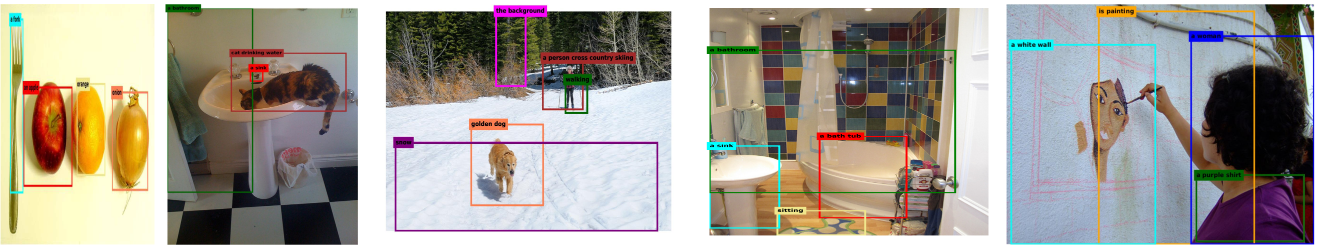
The Align2Ground model outperforms Global and Pooling-based baselines on both the datasets. Specifically, we see an improvement on the phrase localization performance, where our model yields better results than both the Global (by $+2.5$ on COCO and $+3.3$ on Flickr30k) and the Pooling-based (phrases) ($+3.9$ on COCO and $+2.8$ on Flickr30k) method. We believe that the superior performance of our model is due to the fact that our architectural design primes the supervised loss to be a stronger learning signal for the phrase grounding task as compared to the Pooling-based methods. We also observe improvements on the C2I task of 8.2 and 17.3 on COCO compared to the Pooling-based (phrases) and the Global methods respectively.

From our ablation studies on Align2Ground, we notice that the performance of our model is significantly influenced by the choice of our Local Matching module. The *topk* scheme consistently outperforms *max* and *attention* schemes for both the datasets. For example, when using topk for matching phrases with regions (i.e. permInv-topk), we see an increase (w.r.t. using permInv-max) of 16.3 and 20.6 on $R@1$ for COCO and Flickr30k respectively. We also observe similar trends when using the sequence encoder for encoding the matched RoIs. These results support our intuition that the introduction of randomness in the RoI selection step adds diversity to the model and prevents overfitting by prematurely selecting a specific subset of RoIs – a key issue in MIL [8, 26].

4.4. Qualitative Results

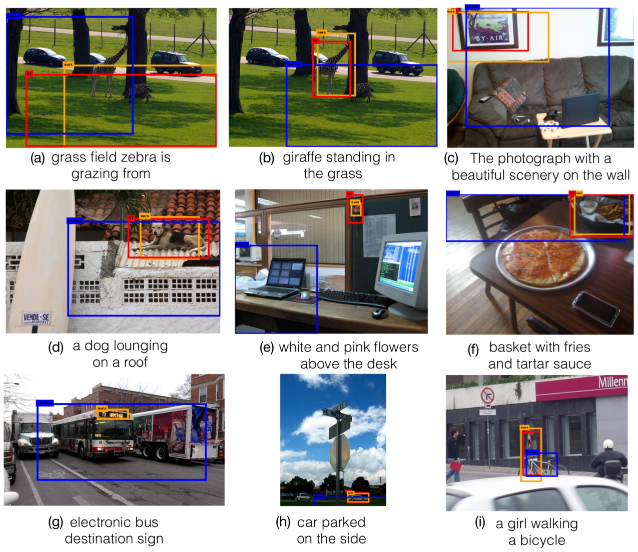
In Figure 4, we show qualitative results of visual grounding of phrases performed by our learned Local Matching module on a few test image–caption pairs (from COCO and Flickr30k). From Figure 4, it is evident that our model is able to localize noun phrases (e.g. "white wall", "large pizza") as well as verb phrases (e.g. "holding", "standing") from the query captions.

In Figure 5, we show qualitative examples of phrase localization from the VG dataset. We compare results of our model with those from Pooling-based methods. We observe that our model is able to correctly localize phrases even when they appear in the midst of visual clutter. For example, in image (f), our model is able to ground the phrase "a basket of fries with tartar sauce". Another interesting example is the grounding of the phrase "white and pink flowers above the desk" in (e) where the Pooling-based method gets confused



(a) a fork next to an apple, orange and onion (b) cat drinking water from a sink in a bathroom (c) golden dog walking in snow with a person cross country skiing the background (d) a bath tub sitting next to a sink in a bathroom (e) a person with a purple shirt is painting an image of a woman on a white wall

Figure 4: We show the image regions that are matched with the phrases in the query caption for five image-caption pairs. Our model is able to effectively learn these correspondences without any phrase level ground-truth annotations during training. Figure best viewed in color.



(a) grass field zebra is grazing from (b) giraffe standing in the grass (c) The photograph with a beautiful scenery on the wall (d) a dog lounging on a roof (e) white and pink flowers above the desk (f) basket with fries and tartar sauce (g) electronic bus destination sign (h) car parked on the side (i) a girl walking a bicycle

Figure 5: We show outputs of our method Align2Ground (in orange) and Pooling-based method (in blue) on the phrase localization task for a few test images. The ground-truth is shown in red. Align2Ground is able to clearly localize better than the Pooling-based model in grounding noun only phrases e.g. (e), (f) as well as phrases with verbs e.g. (d), (i). Figure best viewed in color.

and grounds the desk instead of the flowers. However, our model is able to correctly localize the main subject of the phrase.

4.5. Comparison with state-of-the-art

We compare Align2Ground with state-of-the-art methods from literature for both the tasks of phrase localization (Table 3, 4) and caption-to-image retrieval (Table 2). For phrase localization, we outperform the previous state-of-the-art [13], which uses a variant of the Global method with a novel spatial pooling step, by 4.9% based on the $PointIt\%$ metric on VG. On Flickr30k Entities, we out-perform prior state-of-the-art [2] with much simpler encoders (ResNet+bi-GRU v/s PNASNet+ElMo). For the caption-to-image retrieval task, we also achieve state-of-the-art performance ($R@1$ of 49.7 vs. 48.6 by [25]) on Flickr30k dataset and get competitive results relative to state-of-the-art ($R@1$ of 56.6 vs. 58.8 by [25]) on COCO dataset for the downstream C2I task. These performance gains demonstrate that our model not only effectively learns to ground phrases from the downstream C2I task, but the the tight coupling between these two also ends up helping the downstream task (C2I retrieval).

5. Conclusion

In this work, we have proposed a novel method for phrase grounding using weak supervision available from matching image–caption pairs. Our key contribution lies in designing the Local Aggregator module that is responsible for a tight coupling between phrase grounding and image-caption matching via caption–conditioned image representations. We show that such an interaction between the two tasks primes the loss to provide a stronger supervisory signal for phrase localization. We report improvements of 4.9% and 1.3% (absolute) for phrase localization on VG and Flickr30k Entities. We also show improvements of 3.9% and 2.8% on COCO and Flickr30k respectively compared to prior methods. This highlights the strength of the proposed representation in not allowiing the model to get away without learning to ground all phrases and also not compromising on the downstream task performance. Qualitative visualizations of phrase grounding shows that our model is able to effectively localize free-form phrases in images.

Acknowledgements

The SRI effort was supported in part by the USAMRAA under Contract No. W81XWH-17-C-0083. The Georgia Tech effort was supported in part by NSF, AFRL, DARPA, ONR YIPs. The views, opinions, findings and/or conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

[1] Karuna Ahuja, Karan Sikka, Anirban Roy, and Ajay Divakaran. Understanding visual ads by aligning symbols and objects using co-attention. *arXiv preprint arXiv:1807.01448*, 2018. 2

[2] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. *Conference on Computer Vision and Pattern Recognition*, 2019. 7, 8

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition*, page 6, 2018. 5

[4] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2

[5] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, volume 3, 2014. 4

[6] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. *arXiv preprint arXiv:1803.03879*, 2018. 2

[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075, 2015. 6

[8] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multifold multiple instance learning. *Transactions on Pattern Analysis and Machine Intelligence*, 39(1):189–203, 2017. 7

[9] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. *arXiv preprint arXiv:1805.10547*, 2018. 3

[10] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011. 3, 6

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision*, pages 1422–1430, 2015. 2

[12] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *Conference on Computer Vision and Pattern Recognition*, 2017. 6

[13] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018. 3, 5, 6, 7, 8

[14] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017. 2, 5, 6

[15] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 7

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems*, pages 2121–2129, 2013. 2

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, pages 2980–2988. IEEE, 2017. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 6

[19] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Conference on Computer Vision and Pattern Recognition*, page 7, 2017. 6

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 1, 3, 5, 6

[21] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Neural Information Processing Systems*, pages 1889–1897, 2014. 1, 3

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2, 6

[23] Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge'naturally'in multi-agent dialog. *arXiv preprint arXiv:1706.08502*, 2017. 2

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

[25] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018. 6, 8

[26] Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In *Conference on Computer Vision and Pattern Recognition*, pages 4277–4285, 2015. 7

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[28] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *International Conference on Computer Vision*, pages 2623–2631, 2015. 6

[29] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016. 1, 2

[30] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 6

[31] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle

and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2

[32] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016. 6

[33] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *International Conference on Computer Vision*, pages 1899–1907. IEEE, 2017. 3, 5, 6

[34] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision*, pages 2641–2649, 2015. 2, 5

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 4

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems*, pages 91–99, 2015. 5

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

[38] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 2, 6

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

[40] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. *arXiv preprint arXiv:1705.01371*, 2017. 1, 2, 3, 5, 7

[41] Jieping Ye, Zheng Zhao, and Huan Liu. Adaptive distance metric learning for clustering. In *Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007. 4

[42] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[43] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Neural Information Processing Systems*, pages 3391–3401, 2017. 4

[44] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 5

[45] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016. 1