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

Samyak Datta*,1,2 Karan Sikka1 Anirban Roy1 Karuna Ahuja1 Devi Parikh2,3 Ajay Divakaran1
1SRI International, Princeton, NJ, 2Georgia Institute of Technology, 3Facebook AI Research

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.

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

**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.

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 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 $\{x_j\}_{j=1}^R$, where $x_j \in \mathbb{R}^{d_i}$. 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 $\Phi_{RNN}$. We denote the encoded phrases as a sequence $\{p_k\}_{k=1}^P$, where $p_k \in \mathbb{R}^{d_p}$ 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
similarity. To do so, we project the RoI $x_j$ in the same space as the phrase embeddings, $p_k$, via a linear projection. We then measure the semantic similarity, $s_{jk}$, between region $x_j$ and phrase $p_k$ using cosine similarity.

$$\hat{x}_j = W_l^T x_j$$

$$s_{jk} = \frac{\hat{x}_j^T p_k}{\|\hat{x}_j\|_2 \|p_k\|_2}$$

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

A straightforward approach to infer the matched RoI for a phrase is to select the top scoring box i.e. for a phrase $p_k$ the matched RoI is $x_{j^*}$, 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 $\alpha_{jk}$ 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.

$$\alpha_{jk} = \text{softmax}(s_{jk})_{j=1}^R \quad x_k^c = \sum_j \alpha_{jk} x_j$$

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 $I_{c\text{rois}}^c = (x_k^c)_{k=1}^R$, where $x_k^c$ 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 repre-
sentation, 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 \( \Phi_{RNN} \), 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.

\[
\hat{c} = MLP(\Phi_{RNN}(c)) \quad \hat{r}_c = f_{enc}(I^r_{rois}) \\
S_{Ic} = \frac{\hat{c}^T \hat{r}_c}{\|\hat{c}\|_2 \|\hat{r}_c\|_2} 
\]

\( S_{Ic} \) 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.

\[
L = \max_{c' \in \mathcal{C}_I}(0, m - S_{Ic} + S_{Ic'}) + \max_{I' \notin \mathcal{I}_c}(0, m - S_{Ic} + S_{I'c}) 
\]

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

**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 Point11% 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 Anderson et al.\(^1\) [3]. For every image, we select the top 30 RoIs based on Faster-RCNN’s class detection score (after non-maximal suppr-
We then use RoIAlign with two competing works. The first method, referred to as Global, embeds both the image and caption in a joint embedding space and computes their matching score using cosine similarity. We also implement the Pooling-based method, that computes similarity between image–caption pairs by summarizing the local region–phrase matching scores. We use our Local Matching module to improve 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 in Section 3.
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—Poolin-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 ($Det@1 +0.1$ for COCO and $+0.5$ for 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 ($Det@1 10.8$ 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.

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.
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.

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 allowing 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


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


