Modularized Textual Grounding for Counterfactual Resilience

Zhiyuan Fang\textsuperscript{1}, Shu Kong\textsuperscript{2}, Charless Fowlkes\textsuperscript{2}, Yezhou Yang\textsuperscript{1}

\textsuperscript{1}\{zy.fang, yz.yang\}@asu.edu
\textsuperscript{2}\{skong2, fowlkes\}@ics.uci.edu

Arizona State University, Tempe, USA
University of California, Irvine, USA

Abstract

Computer Vision applications often require a textual grounding module with precision, interpretability, and resilience to counterfactual inputs/queries. To achieve high grounding precision, current textual grounding methods heavily rely on large-scale training data with manual annotations at the pixel level. Such annotations are expensive to obtain and thus severely narrow the model’s scope of real-world applications. Moreover, most of these methods sacrifice interpretability, generalizability, and they neglect the importance of being resilient to counterfactual inputs. To address these issues, we propose a visual grounding system which is 1) end-to-end trainable in a weakly supervised fashion with only image-level annotations, and 2) counterfactually resilient owing to the modular design. Specifically, we decompose textual descriptions into three levels: entity, semantic attributes, and color information, and perform compositional grounding progressively. We validate our model through a series of experiments and demonstrate its improvement over the state-of-the-art methods. In particular, our model’s performance not only surpasses other weakly/un-supervised methods and even approaches the strongly supervised ones, but also is interpretable for decision making and performs much better in face of counterfactual classes than all the others.

1. introduction

Deep neural networks have spawned a flurry of successful work on various computer vision applications, from modular tasks like object instance detection \cite{20, 22, 36} and semantic segmentation \cite{43, 10}, to more complex multi-modal ones like visual question answering (VQA) \cite{1} and image captioning \cite{2, 33}. For complex vision applications (e.g., visual search engine and video auto-captioning), it is critical to build a reliable textual grounding system, which connects natural language descriptions and image regions \cite{67, 34, 32, 58, 65}.

Current methods typically formulate the textual grounding problem as a search process or image-text matching. For example, \cite{58} proposed textual-visual feature matching by reconstruction loss. \cite{9} fulfills textual grounding with two steps: the generation of object proposals and match with the query. \cite{67} utilizes pre-trained module to conduct searching and matching progressively. Given a novel image and queries, these models return the proposals which yield the highest matching score/probability as the final output. Although they achieve state-of-the-art performance in terms of grounding precision, they rely on a large-scale training sets with manually annotated bounding boxes on the objects of interest. This inevitably prevents them from generalizing to other data domains which have no such fine-grained manual annotations for model training or fine-tuning \cite{65}.

Moreover, these models lack the interpretability for decision making and the resilience to counterfactual queries, which often appear jointly to make these models even more

![Figure 1: Illustration of our textual grounding framework that decomposes textual descriptions into three levels: entity, semantic attributes and color information. As an example, for textual grounding from the sentence shown above, our system localizes the entity (person), semantic attributes (boy, woman), the color blue, and progressively produces the final textual grounding by combining results. Note that owing to the decomposable description and modular design, our system is highly interpretable and resilient to counterfactual inputs/queries (bottom row).](image-url)
Figure 2: Examples of counterfactual objects and applying our system to video captioning alignment. Although there are three persons in the beginning of the video, they may disappear later for some frames. This poses a challenge for video captioning, our system acts as a tool to ground the object temporally and correct mismatched description and frames.
figures/objects. This setup largely limits their broad applications as such strong supervision is expensive to obtain, and they also lack interpretability and resilience to counterfactual cases which do not appear in training.

Weakly supervised learning receives increasing attention [13, 50, 11, 46, 52, 55, 63]. It focuses on learning granular detectors given only coarse annotations. This is of practical significance as granular annotations (e.g., bounding boxes and pixel-level labels) are much more expensive to obtain compared to coarse image-level annotations. Recent study shows that weakly supervised methods can even outperform the strongly supervised method for image classification [46]. Unlike current work, we perform weakly-supervised learning for image grounding, including training for both entity grounding and textual-visual matching through a progressive modular procedure.

Modular design is also receiving more attention recently, mainly for complex systems like visual-question-answering or image captioning [29, 27, 67]. Such modular design is carried out by realizing some linguistic structures. In our work, we propose to decompose the query textual description into progressive levels, each of which is passed to a corresponding module, and then produce the final grounding result by progressively merging the intermediate results. In this way, our system enjoys high interpretability and resilience to counterfactual inputs.

3. Modularized Textual Grounding System

To obtain better interpretability and counterfactual resilience, we propose to modularize the our whole textual grounding system by decomposing the textual descriptions into multiple levels, each of which is passed to a specific module to process. We generate the final grounding result by progressively merging intermediate results from these modules.

Without losing generalization, in this work, we decompose the textual descriptions into three levels, and progressively process them with three different modules, respectively: entity grounding module $M_e$, semantic attribute grounding module $M_a$, and color grounding module $M_c$. We extracted phrases/words that belong to these three levels from text, and feed them into their corresponding submodules. We note that such a modular design allows for training different modules using different specialized protocols, e.g., fully supervised learning or weakly supervised learning, while also enabling end-to-end training. For the final grounding heat map $G$, we merge progressively the intermediate results from these modules (see Figure 3):

$$G = M_e \cdot (M_a + M_c).$$

In practice, we observe that such a merging approach achieves the best performance over a straightforward multiplicative or an additive fusion. This is because that the entity grounding defines the object constraints, and the summation over the attribute and color modules determines how the final results are generated interpretably, though they may partially cover some regions belonging to the object of interest. For the rest of Sec. 3, we elaborate the three modules with their adopted training protocols respectively.

3.1. Entity Grounding Module ($M_e$)

To overcome the limitation of current methods that require expensive manual annotations at fine-grained level, we propose to train the entity grounding module in a weakly supervised manner. This can help our system achieve better generalizability to other novel data domains which may just require fine-tuning over dataset annotated coarsely at image level. This weakly supervised learning can be expressed as selecting the best region $r$ in an image $I$ given an object of interest represented by a textual feature $t$, e.g., a word2vec feature. With well pre-trained feature extractor, we first extract visual feature maps $v$ over the image, based on which we train an attention branch $F$ that outputs a heatmap expected to highlight a matched region in the image.

Mathematically, we are interested in obtaining the region $R = F(t, v)$ in the format of heatmap and making sense of it. In practice, we find training a classification model at image level with the attention mechanism works well for entity grounding, which is the output through the attention maps, as illustrated by Figure 3 left. Moreover, rather than using a multiplicative gating layer to make use of the attention map, we find that it works better by using a bilinear pooling layer [42, 17, 35].

For bilinear pooling, we adopt the Multimodal Compact Bilinear (MCB) pooling introduced in [16] that effectively pools over visual and textual features. In MCB, the Count Sketch projection function [7] $\Psi$ is applied on the outer product of the visual feature $v_2$ and an array repeating the word feature $v_1$ for dimensionality reduction: $\Psi(t) \ast \Psi(v)$. If converted to frequency domain, the concatenated outer product can be written as: $\Phi = FFT^{-1}(FFT(\Psi(t)) \odot FFT(\Psi(v)))$. Based on $\Phi$, the final 2D attentive map $R$ is computed through several nonlinear $1 \times 1$ convolutional layers: $R = conv(\Phi)$, with the final one as sigmoid function to shrink all values into $[0, 1]$. Later we retrieve the regional representation $f$ by a global pooling over the element-wise product between entity attentive map and original visual feature maps: $f = pool(R \odot v)$, on which the weakly supervised classification loss is applied. Overall, to train the entity grounding module with the attention mechanism in a weakly supervised learning fashion, we train for image-level $K$-way classification using a cross-entropy loss.

3.2. Semantic Attribute Grounding Module ($M_a$)

The semantic attribute grounding module improves interpretability of the whole textual grounding system by ex-
plaining that it explains how the final decision is being made. For example, a model finding the “man in black suits” as shown in Figure 2 should not only output the final grounding mask, but also explain how the final result is being achieved by showing where “man” and “black suits” are localized in the image.

We also train this module with a weakly supervised learning protocol with similar architecture in the entity module. But instead of training with K-way classification over K predefined attributes as in training entity grounding module, we model this as a multi-label problem, since an image may deliver multiple attributes which are not exclusive to each other. Moreover, rather than classifying them, we propose to use regression for training, since attributes can become large in number while the features representing attribute names can lie in a manifold in the semantic space. This makes our module extensible to more novel attributes even trained with some pre-defined ones.

Note that we represent each attribute with the word2vec feature [47]. Although the word2vec model demonstrates very semantic grouping on words, we find that these features representing attributes do not deliver reasonable discriminativeness. For example, in word2vec features, “man” is more similar to “woman” than “boy” but we care more about the gender meaning in practice. Though retraining such a word2vec model solves the problem, we adopt an alternative method in this paper by proposing a dictionary based scoring function over the original word2vec features. We note that this method not only offers more discriminative scoring power, but also inherits the semantic manifolds in word2vec features, extensible to novel attributes without re-training whole model as done in K-way classification.

To introduce our dictionary based scoring function, we revisit the classic logistic normalization widely used in binary classification as below:

\[
y_i = \frac{1}{1 + \exp(-w_i^T x)}
\]

where \(w_i\) here represents the learning parameters, and \(x, y_i\) are the input vectors and predicted probability with respect to class \(i\). Note again that, although the logistic loss works well for binary classification or multi-label classification, it is not extensible to novel classes unless retraining the whole model. Our solution to this is based on the proposed dictionary based scoring function. Suppose there are \(C\) attributes, represented by word2vec and stacked as a dictionary \(D = [d_1, \ldots, d_C]\). We can measure the (inverse) Euclidean distance between \(x\) and each dictionary atom for the similarity about which attribute \(x\) is predicted.

So the dictionary acts as the parameter bank which can be fixed if we want to preserve the semantic manifold in the word2vec feature space, and we have the following modified sigmoid transformation:

\[
y_i = \frac{2}{1 + \exp(\|d_i - x\|^2_2)}
\]

However, as this may also be less discriminative, we opt to learn a new latent space. Concretely, we build new layers before the sigmoid transformation, and these layers form new function \(\phi\) and \(\psi\) to transform the feature \(x\) and dictionary atoms, respectively. Then we have the following dictionary based scoring function for the \(i^{th}\) attribute:

\[
y_i = \frac{2}{1 + \exp(\|\psi(D)_i - \phi(x))\|^2_2)}
\]
Furthermore, despite using the dictionary based scoring function as a modified sigmoid for logistic loss over the holistic feature globally pooled over the image, we also perform it at pixel levels. Concretely, during each iteration on each training image, we choose the $T$ pixels with the top scores to feed into the logistic loss. This practice is essentially a multi-instance learning at pixel level [53]. We find in our experiment that jointly using the two losses helps generate better attention maps.

### 3.3. Color Grounding Module ($M_c$)

When querying in natural languages, human beings typically rely on textual descriptions for low-level vision characteristics, e.g., color, texture, shape and locations. Recent work also demonstrates the feasibility of grounding low-level features in unsupervised learning [61]. In our work for the datasets we studied in our work, we notice that color is the most used one. In the Flickr30k Entities dataset [56] as studied in this paper, 70% attributes words are colors describing persons. Therefore, without loss of generalization, we develop a separate color grounding module to improve the interpretability of the overall textual grounding system.

Different from entity grounding and semantic attribute grounding modules, we train this color grounding module in a fully supervised way over a small-scale dataset, called Color Name Dataset [59], which contains 400 images with color name annotations at pixel level. We essentially perform pixel-level color segmentation over the input image to ground color reference. Moreover, we build this color grounding module over a ResNet50 model [23] pretrained on ImageNet dataset [12], and concatenate intermediate features at lower levels for pixel-level color segmentation. We find this works better than combining high-level features. We conjecture the reason is due to that color is a very low-level cue that does not require deep architectures and high-level feature abstraction. This is consistent with what reported in [40].

### 3.4. Architecture and Training

Our three modules are based on the ResNet architecture [23]. Similar to [10, 37], we increase the output resolution of ResNet by removing the top global $7 \times 7$ pooling layer and the last two $2 \times 2$ pooling layers, replacing them with atrous convolution with dilation rate 2 and 4, respectively to maintain a spatial sampling rate. Our model thus outputs predictions at $1/8$ the input resolution which are upsampled for benchmarking. For (multi-label or $K$-way) classification, we use a global pooling layer that produces a holistic image feature for classification. In addition, we also insert an $L_2$ regularization over the attention maps, and we observe that such a regularization term helps reduce noises effectively.

We use the standard stochastic gradient decent (SGD) for training in a stagewise fashion. Specifically, we first train a plain classification model for entity and semantic attribute grounding modules, then we build the attention branch for attentional learning.

Though our textual grounding system is end-to-end trainable, we train each module separately. And though joint training is straightforward to implement, we do not do this for practical reasons: 1) we can easily plug in a better trained module without retraining the whole system for better comparison; 2) we focus on the modular design, isolating the influence of the settings and parameters of each module.

### 4. Experiments

We now experimentally validate our system and compare it with the state-of-the-art methods. To highlight the generalizability of our system, we train it on COCO2017 dataset [41] while test it on another Flickr30K Entities dataset [56]. We first introduce the two datasets briefly before conducting thorough comparisons, then we carry out another experiment to show our (weakly supervised) model performs remarkably better than other (fully supervised) methods on a collected dataset consisting of counterfactual testing cases. We implement our algorithm using PyTorch toolbox [51] on a single GTX1080 Ti GPU.

#### 4.1. Datasets and Preprocessing

The two datasets we used in our experiments are: COCO2017 [41] for training our system and Flickr30k Entities Dataset [56] for testing it.

COCO2017 dataset contains 110k training images with 80 object categories at image level. These 80 object categories are used for training our entity grounding module as they can be seen exclusive to each other. The captioning task and the annotations provided in COCO2017 enables us to train our semantic attribute grounding module. Using [4, 48], we tokenize and mine out words related to semantic attributes (e.g., man, woman, boy, old and young) to form our corpus. To train the semantic attribute grounding module, we retrieve images from COCO2017 whose captions contain the attributes existing in our corpus. Eventually, 10,000 images and 34 attributes are collected from COCO2017 for weakly supervised training our modules. To alleviate imbalanced distribution of these attributes, we adopt inverse frequency reweighting during training.

The Flickr30k Entities dataset contains over 31k images with 275k bounding boxes with natural languages descriptions, and we only use this dataset for testing our system with the bounding boxes.

To carry out counterfactual testing experiment, we collect a new testing set with images from Flickr30k and Ref
By developing an easy-to-use interface, we are able to generate counterfactual captions for a given image with the good captions provided by the original dataset. Similar to work in [24], we generate counterfactual attributes by mining the negation of existing attributes. The overall PACG dataset consists 2,000 images, a half of which are with counterfactual attributes not existing in the image and the other half with “correct” attributes.

**Language Processing:** To deal with free-form textual queries, we use a language parser [4] to select the keywords according to the functionalities of the three modules. We first extract the entity words and pick the most similar object classes by word similarities. We then extract the semantic attribute words in the same way. Finally, we extract the the color keywords simply for the color grounding. To represent the textual attributes and color names, we adopt the word vectors from GloVe [54]. This enables meaningful similarity between the defined attributes/colors and novel ones when encountered at testing stage.

### 4.2. Textual Grounding Evaluation

We compare our modular textual grounding system with other supervised/unsupervised methods on the Flickr30k Entities dataset. We use the mean average precision (mAP) metric to measure the quantitative performance. The detailed comparison is listed in Table 1.

As the first baseline method similar to [65], we select the largest proposal as the final result. This method achieves 24.34% mAP. Then, we build another baseline model that

---

**Table 1:** Comparative analysis of textual grounding performance

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>24.34</td>
</tr>
<tr>
<td>Proposed System</td>
<td></td>
</tr>
</tbody>
</table>
Table 1: Phrase localization performance on Flickr 30k Entities (accuracy in %).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Image Features</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCRC [31]</td>
<td>VGG-cls</td>
<td>27.80</td>
</tr>
<tr>
<td>GroundeR\textsubscript{u} [58]</td>
<td>VGG-cls</td>
<td>47.81</td>
</tr>
<tr>
<td>CCA [56]</td>
<td>VGG-det</td>
<td>50.89</td>
</tr>
<tr>
<td>IGOP [64]</td>
<td>YOLO+DeepLab</td>
<td>53.97</td>
</tr>
<tr>
<td><strong>Unsupervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Largest proposal</td>
<td>n/a</td>
<td>24.34</td>
</tr>
<tr>
<td>GroundeR\textsubscript{u} [58]</td>
<td>VGG-det</td>
<td>28.94</td>
</tr>
<tr>
<td>Mutual Info. [68]</td>
<td>VGG-det</td>
<td>31.19</td>
</tr>
<tr>
<td>UTG [65]</td>
<td>VGG-det</td>
<td>35.90</td>
</tr>
<tr>
<td>UTG [65]</td>
<td>YOLO-det</td>
<td>36.93</td>
</tr>
<tr>
<td><strong>Weakly-Supervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours\textsuperscript{1}</td>
<td>Res101</td>
<td>29.01</td>
</tr>
<tr>
<td>Ours(Attr)\textsuperscript{1}</td>
<td>Res101</td>
<td>32.04</td>
</tr>
<tr>
<td>Ours(Attr+Col)\textsuperscript{1}</td>
<td>Res101</td>
<td>33.43</td>
</tr>
<tr>
<td>Faster-RCNN\textsuperscript{2} [57]</td>
<td>Res101-det</td>
<td>35.35</td>
</tr>
<tr>
<td>Ours+Attr\textsuperscript{2}</td>
<td>Res101-det</td>
<td>47.46</td>
</tr>
<tr>
<td>Ours+Attr+Col\textsuperscript{2}</td>
<td>Res101-det</td>
<td>48.66</td>
</tr>
</tbody>
</table>

we train the entity grounding module only through weakly supervised learning over a ResNet101 backbone, which is pretrained over ImageNet dataset. Then, over the entity grounding heatmaps, we generate bounding boxes candidates by sub-window search [39] together with contour detection results, followed by a Non-Maximum Suppression to further refine the proposal boxes. We select the box that encompasses largest ratio of object according to equation 1. We note that this simple baseline module (29.01% mAP) outperforms GroundeR\textsubscript{u} [58] (28.94% mAP) that learns grounding in an attentive way over large-scale training data. If we include our semantic attribute module, we improve the performance further (32.04% mPA), outperforming Mutual Info. [68]. If we further insert the color grounding module, we achieve comparable performance (33.43%) to UTG (36.93% mAP), which adopts an unsupervised method to link image concepts to query words [65]. We note that our models are trained on COCO dataset only, unlike all these methods which are trained on the same dataset (Flickr30k dataset). The effectiveness of our model is demonstrated by its good transferability, as it is trained and tested on different data domains.

It is also worth noting that, all the compared unsupervised methods unanimously adopt a well-trained object detector, even though they claim to be unsupervised learning. To gain an idea how the detector improves the performance, we fine-tune the faster-RCNN detector [19] on COCO and train our modules with weak supervision again. We report our results as the bottom two rows in Table 1. Now we can see our models perform significantly better, and even surpasses some fully supervised methods (SCRC [31] and GroundeR [58]). Although it seems unfair that our system adopts ResNet101 architecture while most compared methods uses shallower VGG networks, we note that IGOP which adopts both VGG and ResNet101 (denoted by DeepLab) achieves the best performance with fully supervised training. Even though our best model does not outperform IGOP, we believe the performance gap is small and reasonable as our training is carried out on a different dataset (COCO) rather than Flickr30k, and it does not rely on any strong supervision signals. We show output examples of entity grounding module in Figure 5 with various object categories as input, and attribute grounding outputs in Figure 4, with both existing attributes and counterfactual attributes as queries. These visualizations demonstrates how our system rejects in an explainable way the counterfactual queries through the modular output.

4.3. Counterfactual Grounding Evaluation

We now carry out in-depth study on how our system performs when facing of counterfactual textual queries over our collected PACG dataset, and compare with three baseline or state-of-the-art methods, Faster-RCNN [57], MattNet [67], SNLE [30]. We plot the ROC curves for these methods in Figure 6. Textual grounding system then selects the region with highest scores/probability. We compare the prediction scores/probabilities of the predicted regions between the counterfactual queries and normal queries and expecting to observe distinct difference between their numerical scores.

We clearly see from the figure that our system achieves the highest AUC among of these methods, meaning that modular design successfully increases the counterfactual resilience of the grounding system. Specifically, end-to-end models like SNLE [30] encode the textual query into a vector representation to extract spatial feature maps from the image as response map. However, such encoding do not consider the internal structure of sentences [45], also neglecting semantic nuances of near-synonyms. Note that MattNet [67] also adopts a modular design, but it is trained with fully supervised learning, also it is not easily extended to novel attributes and unable to reject counterfactual queries as effectively as our method. The AUC of Faster-RCNN is approximately 0.5 since the recognition ability is restricted to entity-level and not been able to discern among semantic attributes. We conclude that with the modular design and better scoring function in each modules, our model demonstrated highly resilient ability against counterfactual queries, even with only weakly-supervised training.

5. Extensive Applications

The counterfactual resilient design can be furthered applied to various tasks. In this section we showcase some
Figure 6: ROC of our modular network demonstrates high resolving ability on PAGC dataset with an AUC of 0.88, comparing to other state of the art baseline models (best viewed in color).

Girl in yellow coat.

A man with white shirt.

Lady with red dress.

Figure 7: Temporal/Spatial grounding in video sequences. Time-segments contain phrases are selected to filter out irrelevant frames.

practical applications.

Grounding Textual Phrase in Video To ground textual phrase in video, the system needs to first determine which temporal segment and moment to retrieve [25], then localize the region associated with the descriptions. In this case, textual information may be irrelevant to most of the video frames, thus requiring the system to be counterfactual resilient to query and discern whether it is existing or not in the current segment. Unlike an existing approach [18], which treats the problem as temporal localization, we score a set of frames and select out segments that are more likely to be relevant to sentence. We demonstrate this process in Figure 7 that modular network successfully conduct a temporal-spatial grounding task in video clips.

Video to Captioning alignment Our model can be used to correct misaligned captioning sentences like the work in [5]. Given mis-matched frames and captions, we examine the sentence-frame relevance and find the corresponding frame for each sentence. Figure 8 shows an example of the captioning alignment, the temporal linked sentences can be reordered based on video.

6. Conclusion

In this paper, we propose to modularize the complex textual grounding system by decomposing the textual description/query into three parts: entity, semantic attributes and color. Such a modular design largely improves the interpretability and counterfactual resilience of the system. Moreover, we propose to train the modules in a weakly supervised way, so we merely needs image-level labels which are easy to obtain. This largely helps alleviate the requirement of large-scale manual annotated images for training, and for fine-tuning if transferring the system to a new data domain. Through extensive experiments, we show our system not only surpasses all unsupervised textual grounding methods and some of fully supervised ones, but also delivers strong resilience when facing counterfactual queries.

Our modularized textual grounding system is of practical significance as it can be deployed in various problems. In this paper, we show how our system can be applied to video captioning correction and visual-textual search. We expect more applications can benefit from our modular design.

Acknowledgements: The support of the National Science Foundation under the Robust Intelligence Program (1816039 and 1750082), and a gift from Verisk AI are gratefully acknowledged. We also acknowledge NVIDIA for the donation of GPUs.
References


