Modularized Textual Grounding for Counterfactual Resilience

Zhiyuan Fang[1], Shu Kong[2], Charless Fowlkes[2], Yezhou Yang[1]

[1]{zy.fang, yz.yang}@asu.edu   Arizona State University, Tempe, USA
[2]{skong2, fowlkes}@ics.uci.edu   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 attribute, 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 [20, 22, 36] and semantic segmentation [43, 10], to more complex multimodal ones like visual question answering (VQA) [1] and image captioning [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 [67, 34, 32, 58, 65].

Current methods typically formulate the textual grounding problem as a search process or image-text matching. For example, [58] proposed textual-visual feature matching by reconstruction loss. [9] fulfills textual grounding with two steps: the generation of object proposals and match with the query. [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 [65].

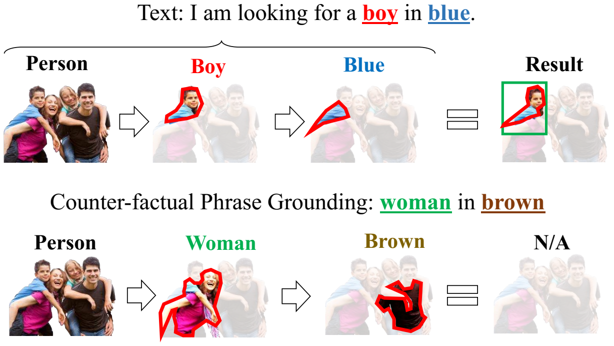

Text: I am looking for a boy in blue.
Person
Boy
Blue
Result
Counter-factual Phrase Grounding: woman in brown
Person
Woman
Brown
N/A

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/qeueries (bottom row).

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

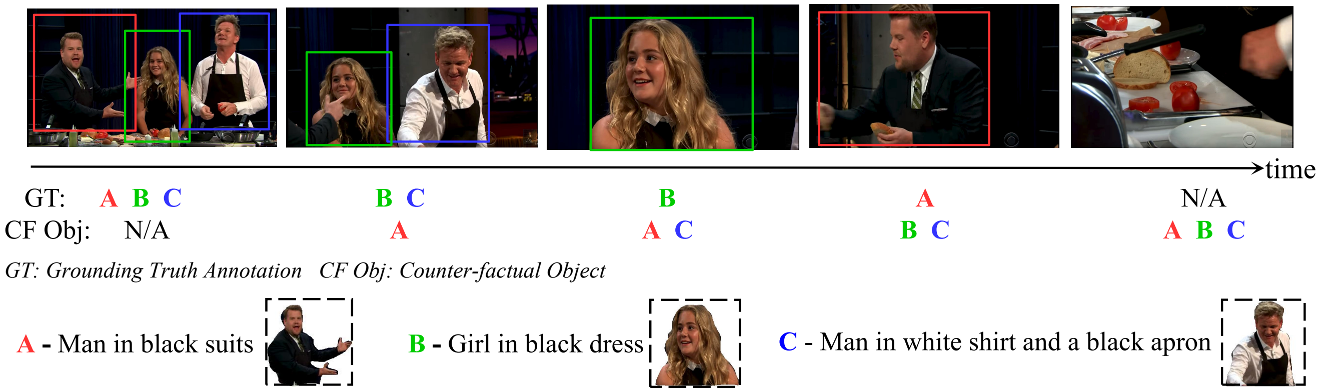

time
GT: A B C
CF Obj: N/A
B C
A
B
A C
A
B C
N/A
A B C
GT: Grounding Truth Annotation CF Obj: Counter-factual Object
A - Man in black suits
B - Girl in black dress
C - Man in white shirt and a black apron

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.

vulnerable in real-world applications [24, 62, 14, 15]. For example, as demonstrated by Figure 1, if one is asking "who is the woman in blue shirt in the image", a good model should return nothing instead of the closest person or someone with high matching score. Even more preferred, the model should explain why the decision is made in addition to the final grounding result. The interpretability and counterfactual resilience properties are also useful in literature and practical deployment. As demonstrated by another example about our application to correcting video auto-captioning, as shown in Figure 2 (details in Section 5). There exist three people in the first frame, while they may disappear in the following frames but the captioning are still not updated. Our counterfactually resilient grounding system is able to correct captioning mis-alignment issue.

In this work, we propose to modularize the textual grounding system by decomposing the textual description into multiple components, and perform grounding progressively through these components towards the final output. Recently, modular design is being advocated in the community [29, 27, 67], mainly focusing on visual-question-answering and referring expression visual matching. We show that such a modular design also increases the interpretability of our textual grounding system, that it explains along the way how the final decision is being made. It is worth noting that the modular design supports diverse training protocols to learn each component. Therefore, to alleviate the requirement for large-scale fine-grained manual annotations (e.g., bounding box), we propose to train our entity grounding module in a weakly supervised manner which only needs image level labels. We note that such data are easy to obtain, e.g., from internet search engine or social media with image tags [21, 3, 8].

To validate our system, we carry out extensive experiments on the COCO dataset [41] and Flickr30k Entities dataset [56]. We show that our system outperforms other weakly-supervised methods on textual grounding and even surpasses some strongly-supervised approaches. By introducing another dataset consisting of counterfactual cases, we emphasize that our system performs remarkably better than other methods w.r.t counterfactual resilience. To summarize our contributions:

1. We propose a textual grounding system with modular design. Together with the decomposition of textual descriptions, it allows for more diverse and specialized training protocols for each components.
2. We collect a counterfacutal textual grounding test set, and show that our system achieves better interpretability and resilience to counterfactual testing.
3. We demonstrate practical applications based on our system and expect future explorations based on our work.

In the rest of the paper, we first review related work, then describe our system in Section 3. We elaborate our training procedure and demonstrate the effectiveness of our system through experiment in Section 4 and broad application in Section 5, respectively, before concluding in Section 6.

2. Related Work

Multi-modal tasks, eg. assistive visual search [6, 38] and image captioning [66, 60], has been studied for decades in the community. While those tasks are classical topics in computer vision and natural language processing, current advancement has further energized it by interplaying vision (images) and language (high-level guide) for practical applications. Specific examples include referring expressing understanding [49, 29] and reasoning-aware visual-question-answering [28].

State-of-the-art textual grounding methods [67, 31, 58, 56, 64, 44] are based on deep neural networks and relying on large-scale training data with manual annotations for the object bounding box and relationship between phrases and 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 textual 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):

$$G = M_e \cdot (M_a + M_c). \tag{1}$$

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$_e$)

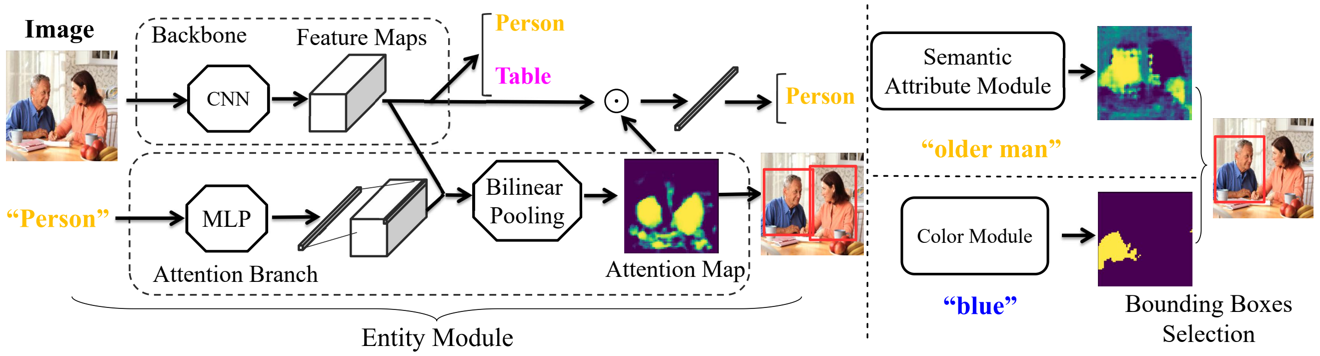
To overcome the limitation of current methods that require expensive manual annotations at fine-grained level, we propose to train the entity grounding module in a weakly supervised manner. This can help our system achieve better generalizability to other novel data domains which may just require fine-tuning 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] Ψ is applied on the outer product of the visual feature v_2 and an array repeating the word feature v_1 for dimensionality reduction: $\Psi(t) * \Psi(v)$. 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 Φ, the final 2D attentive map R is computed through several nonlinear 1×1 convolutional layers : $R = conv(\Psi)$, 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-



Caption: A lady discusses something with an older man in blue shirt at a table in a kitchen.

Target: older man in blue => {Entity: person Semantic Attribute: older man Color: blue}

Figure 3: Illustrative diagram for our entity grounding module (left) and the whole textual grounding system (right). The textual phrase is first decomposed into sub-elements, e.g., "older man in blue" can be parsed to "person" category with "older man" and "blue" to be it's attributes, and later fed into corresponding sub-module. The bounding boxes are generated and selected based upon the merged attention maps. We train the entity/semantic attribute grounding module in a weakly supervised fashion with a attention mechanism. The semantic attribute module also adopt similar architecture of entity module, however with a dictionary learning loss. (best viewed in color)

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:

$$y_i = \frac{1}{1 + \exp(-\mathbf{w}_i^T \mathbf{x})} \tag{2}$$

where $\mathbf{w}_i$ here represents the learning parameters, and $\mathbf{x}, y_i$ 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 $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_C]$. We can measure the (inverse) Euclidean distance between $\mathbf{x}$ and each dictionary atom for the similarity about which attribute $\mathbf{x}$ 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:

$$y_i = \frac{2}{1 + \exp(\|\mathbf{d}_i - \mathbf{x}\|_2^2)} \tag{3}$$

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 ϕ and ψ to transform the feature $\mathbf{x}$ and dictionary atoms, respectively. Then we have the following dictionary based scoring function for the i^{th} attribute:

$$y_i = \frac{2}{1 + \exp(\|\psi(\mathbf{D})_i - \phi(\mathbf{x})\|_2^2)} \tag{4}$$

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 ($\mathbf{M}_c$)

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 × 7 pooling layer and the last two 2 × 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 [1].

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-

[1] https://github.com/jacobswan1/MTG-pytorch

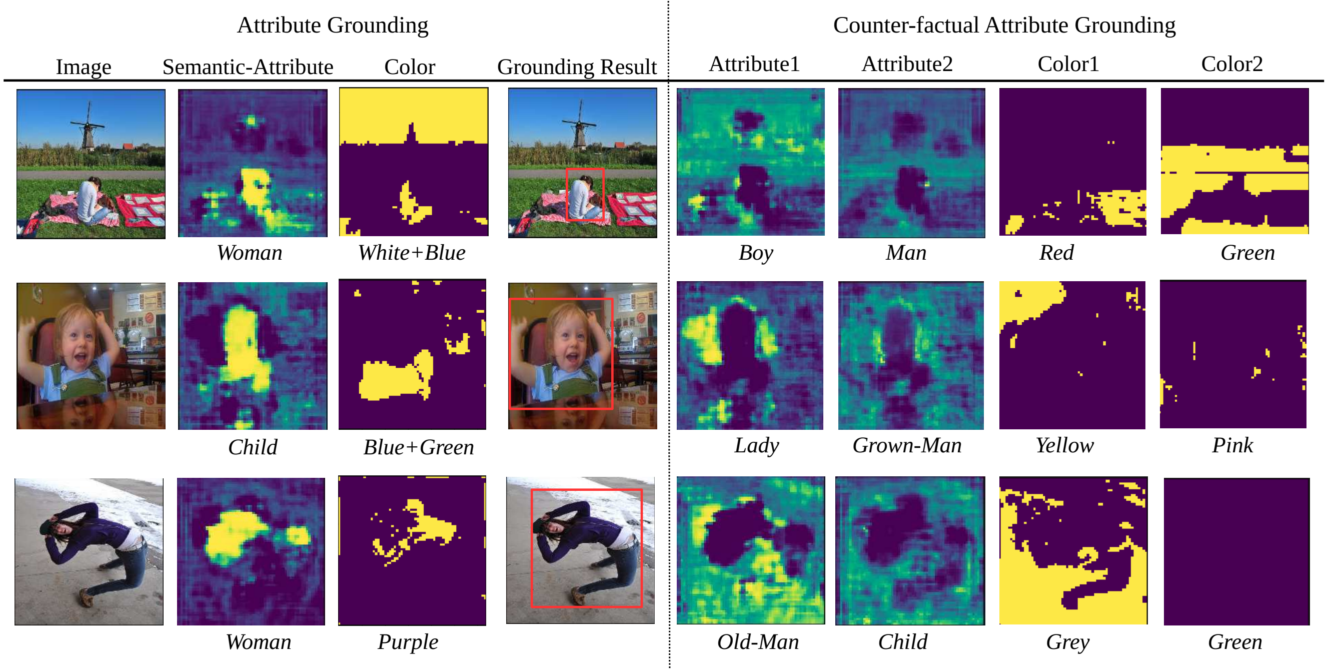


Figure 4: Examples of attribute grounding predictions (left) and counterfactual attribute grounding results (right). (best viewed in color)

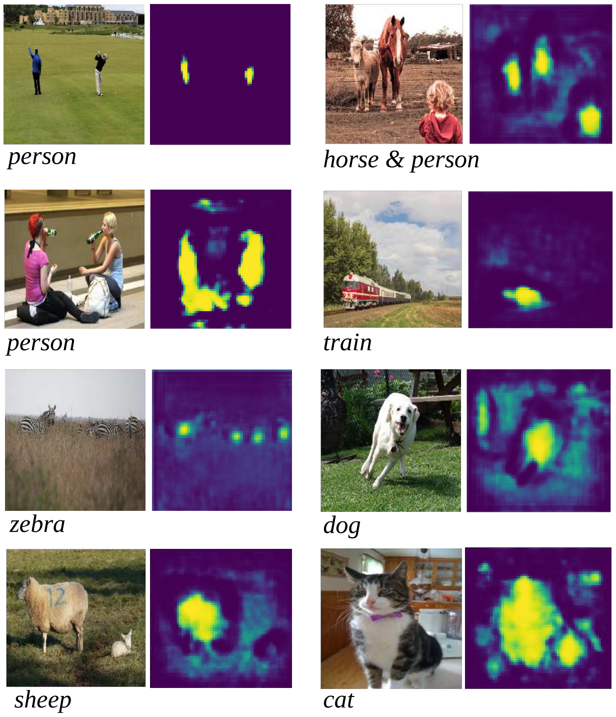


Figure 5: Qualitative examples of attention maps from the entity module.

COCO+ [34]. The images only contain persons and relevant attributes (e.g., gender, age, etc), so we call this dataset Person Attribute Counterfactual Grounding dataset (PACG). 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

Aprroach	Image Features	mAP (%)
Supervised		
SCRC [31]	VGG-cls	27.80
GroundeR$_s$ [58]	VGG-cls	47.81
CCA [56]	VGG-det	50.89
IGOP [64]	YOLO+DeepLab	**53.97**
Unsupervised		
Largest proposal	n/a	24.34
GroundeR$_u$ [58]	VGG-det	28.94
Mutual Info. [68]	VGG-det	31.19
UTG [65]	VGG-det	35.90
UTG [65]	YOLO-det	36.93
Weakly-Supervised		
Ours^1	Res101	29.01
Ours(Attr)^1	Res101	32.04
Ours(Attr+Col)^1	Res101	33.43
Faster-RCNN^2 [57]	Res101-det	35.35
Ours+Attr^2	Res101-det	47.46
Ours+Attr+Col^2	Res101-det	**48.66**

Table 1: Phrase localization performance on Flickr 30k Entities (accuracy in %).

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 GroundR$_u$ [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

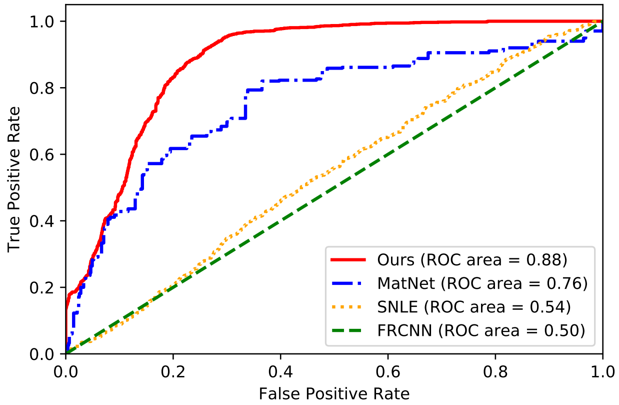


Figure 6: ROC of our modular network demonstrates high resolving ability on PACG dataset with an AUC of 0.88, comparing to other state of the art baseline models (best viewed in color).

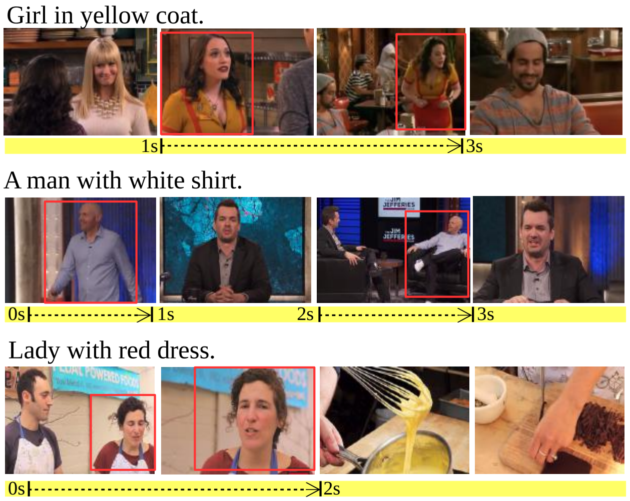


Figure 7: Temporal/Spatial grounding in video sequences. Time-segments contain phrases are selected to filter out irrelevant frmaes.

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.

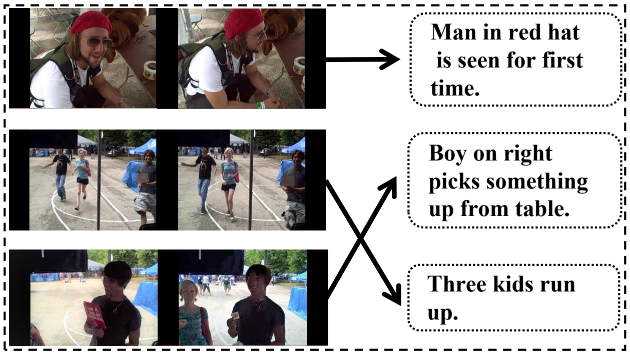

Man in red hat is seen for first time.
Boy on right picks something up from table.
Three kids run up.

Figure 8: Video captioning alignment. With unordered captions, our system links each sentence to it's corresponding frames. Examples took from DiDeMo [26].

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

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[3] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in neural information processing systems*, pages 181–189, 2010.

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[5] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *Proceedings of the IEEE international conference on computer vision*, pages 4462–4470, 2015.

[6] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 952–959. ACM, 2004.

[7] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[8] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of International Conference on Multimedia Retrieval*, page 1. ACM, 2014.

[9] K. Chen, R. Kovvuri, and R. Nevatia. Query-guided regression network with context policy for phrase grounding. *arXiv preprint arXiv:1708.01676*, 2017.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[11] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[13] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010.

[14] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623*, 2018.

[15] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, and A. Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[16] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[17] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[18] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018.

[19] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[20] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[21] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *European Conference on Computer Vision*, pages 198–208. Springer, 2012.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

[25] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017.

[26] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *International Conference on Computer Vision (ICCV)*, 2017.

[27] R. Hu, J. Andreas, T. Darrell, and K. Saenko. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[28] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017.

[29] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4418–4427. IEEE, 2017.

[30] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[31] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.

[32] D.-A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. Finding it: Weakly-supervised reference-aware visual grounding in instructional videos. CVPR, 2018.

[33] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[34] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[35] S. Kong and C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 7025–7034. IEEE, 2017.

[36] S. Kong and C. Fowlkes. Recurrent pixel embedding for instance grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9018–9028, 2018.

[37] S. Kong and C. C. Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 956–965, 2018.

[38] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *cbaivl*, page 24. IEEE, 1998.

[39] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2129, 2009.

[40] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[42] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[43] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[44] R. Luo and G. Shakhnarovich. Comprehension-guided referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[45] B. MacWhinney. Second language acquisition and the competition model. *Tutorials in bilingualism: Psycholinguistic perspectives*, pages 113–142, 1997.

[46] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[48] G. Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[49] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016.

[50] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. 2011.

[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[52] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

[53] S. Paul, S. Roy, and A. K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. *arXiv preprint arXiv:1807.10418*, 2018.

[54] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[55] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015.

[56] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[57] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[58] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[59] J. Van De Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[60] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2017.

[61] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy. Tracking emerges by colorizing videos. *arXiv preprint arXiv:1806.09594*, 2018.

[62] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. 2017.

[63] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3190–3197, 2014.

[64] R. Yeh, J. Xiong, W.-M. Hwu, M. Do, and A. Schwing. Interpretable and globally optimal prediction for textual grounding using image concepts. In *Advances in Neural Information Processing Systems*, pages 1912–1922, 2017.

[65] R. A. Yeh, M. N. Do, and A. G. Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proc. CVPR*, volume 8, 2018.

[66] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[67] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. *arXiv preprint arXiv:1801.08186*, 2018.

[68] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1681–1688, 2013.