MAF: Multimodal Alignment Framework for Weakly-Supervised Phrase Grounding

Qinxin Wang[1], Hao Tan[2], Sheng Shen[3], Michael W. Mahoney[3], Zhewei Yao[3]
[1]Shanghai Jiao Tong University [2]UNC Chapel Hill [3]University of California, Berkeley
qinzzz@sjtu.edu.cn, haotan@cs.unc.edu, {sheng.s, mahoneymw, zheweiy}@berkeley.edu

Abstract

Phrase localization is a task that studies the mapping from textual phrases to regions of an image. Given difficulties in annotating phrase-to-object datasets at scale, we develop a Multimodal Alignment Framework (MAF) to leverage more widely-available caption-image datasets, which can then be used as a form of weak supervision. We first present algorithms to model phrase-object relevance by leveraging fine-grained visual representations and visually-aware language representations. By adopting a contrastive objective, our method uses information in caption-image pairs to boost the performance in weakly-supervised scenarios. Experiments conducted on the widely-adopted Flickr30k dataset show a significant improvement over existing weakly-supervised methods. With the help of the visually-aware language representations, we can also improve the previous best unsupervised result by 5.56%. We conduct ablation studies to show that both our novel model and our weakly-supervised strategies significantly contribute to our strong results.[1]

1 Introduction

Language grounding involves mapping language to real objects or data. Among language grounding tasks, phrase localization—which maps phrases to regions of an image—is a fundamental building block for other tasks. In the *phrase localization task*, each data point consists of one image and its corresponding caption, i.e., $d = \langle I, S \rangle$, where I denotes an image and S denotes a caption. Typically, the caption S contains several query phrases $\mathcal{P} = \{p_n\}_{n=1}^{N}$, where each phrase is grounded to a particular object in the image. The goal is to find the correct relationship between (query) phrases in the caption and particular objects in the image. Existing work (Rohrbach et al., 2016; Kim et al., 2018; Li et al., 2019; Yu et al., 2018; Liu et al., 2020) mainly focuses on the supervised phrase localization setting. This requires a large-scale annotated dataset of phrase-object pairs for model training. However, given difficulties associated with manual annotation of objects, the size of grounding datasets is often limited. For example, the widely-adopted Flickr30k (Plummer et al., 2015) dataset has 31k images, while the caption dataset MS COCO (Lin et al., 2014) contains 330k images.

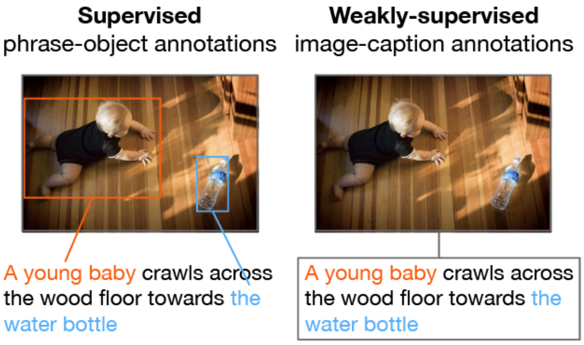

Supervised
phrase-object annotations
Weakly-supervised
image-caption annotations
A young baby crawls across the wood floor towards the water bottle
A young baby crawls across the wood floor towards the water bottle

Figure 1: Comparison of phrase localization task under supervision (left) and weak supervision (right).

To address this limited data challenge, two different approaches have been proposed. First, a weakly-supervised setting—which requires only *caption-image annotations*, i.e., no *phrase-object annotations*—was proposed by Rohrbach et al. (2016). This is illustrated in Figure 1. Second, an unsupervised setting—which does not need any training data, i.e., neither caption-image and phrase-object annotation—was proposed by Wang and Specia (2019). To bring more semantic information in such a setting, previous work (Yeh et al., 2018; Wang and Specia, 2019) used the detected object labels from an off-the-shelf object detector (which we will generically denote by PreDet) and achieved promising results. In more detail, for a given im-

[1] Code is available at https://github.com/qinzzz/Multimodal-Alignment-Framework.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2030–2038,
November 16–20, 2020. ©2020 Association for Computational Linguistics

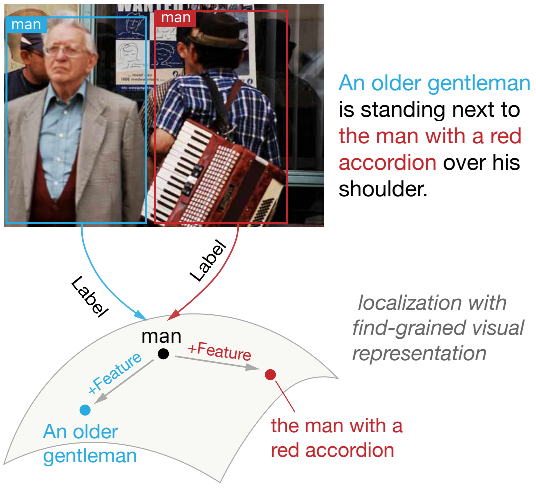

man
man
An older gentleman is standing next to the man with a red accordion over his shoulder.
Label
Label
localization with find-grained visual representation
man
+Feature
+Feature
An older gentleman
the man with a red accordion

Figure 2: Example of the ambiguity caused by label-based localization (top); and our fine-grained visual representation disambiguate labels (bottom).

age I, the PreDet first generates a set of objects $\mathcal{O} = \{o_m\}_{m=1}^{M}$. Afterward, all the query phrases $\mathcal{P}$ and the detected objects $\mathcal{O}$ are fed into an alignment model to predict the final phrase-object pairs. However, purely relying on the object labels causes ambiguity. For example, in Figure 2, the grounded objects of phrases “an older man” and “the man with a red accordion” are both labeled as “man,” and thus they are hard to differentiate.

Given these observations, we propose a Multimodal Alignment Framework (MAF), which is illustrated in Figure 3. Instead of using only the label features from the PreDet (in our case, a Faster R-CNN (Ren et al., 2015; Anderson et al., 2018a)), we also enhance the visual representations by integrating visual features from the Faster R-CNN into object labels. (This is shown in Figure 2.) Next, we build visually-aware language representations for phrases, which thus could be better aligned with the visual representations. Based on these representations, we develop a multimodal similarity function to measure the caption-image relevance with phrase-object matching scores. Furthermore, we use a training objective to score relevant caption-image pairs higher than irrelevant caption-image pairs, which guides the alignment between visual and textual representations.

We evaluate MAF on the public phrase localization dataset, Flickr30k Entities (Plummer et al., 2015). Under the weakly-supervised setting (i.e., using only caption-image annotations without the more detailed phrase-object annotations), our method achieves an accuracy of 61.43%, outperforming the previous weakly-supervised results by 22.72%. In addition, in the unsupervised setting, our visually-aware phrase representation improves the performance from the previous 50.49% by 5.56% up to 56.05%. Finally, we validate the effectiveness of model components, learning methods, and training techniques by showing their contributions to our final results.

2 Related Work

With the recent advancement in research in computer vision and computational linguistics, multimodal learning, which aims to explore the explicit relationship across vision and language, has drawn significant attention. Multimodal learning involves diverse tasks such as Captioning (Vinyals et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2015; Venugopalan et al., 2015), Visual Question Answering (Anderson et al., 2018a; Kim et al., 2018; Tan and Bansal, 2019), and Vision-and-Language Navigation (Anderson et al., 2018b; Chen et al., 2019; Thomason et al., 2020). Most of these tasks would benefit from better phrase-to-object localization, a task which attempts to learn a mapping between phrases in the caption and objects in the image by measuring their similarity. Existing works consider the phrase-to-object localization problem under various training scenarios, including supervised learning (Rohrbach et al., 2016; Yu et al., 2018; Liu et al., 2020; Plummer et al., 2015; Li et al., 2019) and weakly-supervised learning (Rohrbach et al., 2016; Yeh et al., 2018; Chen et al., 2018). Besides the standard phrase-object matching setup, previous works (Xiao et al., 2017; Akbari et al., 2019; Datta et al., 2019) have also explored a pixel-level “pointing-game” setting, which is easier to model and evaluate but less realistic. Unsupervised learning was studied by Wang and Specia (2019), who directly use word similarities between object labels and query phrases to tackle phrase localization without paired examples. Similar to the phrase-localization task, Hessel et al. (2019) leverages document-level supervision to discover image-sentence relationships over the web.

3 Methodology

3.1 Fine-grained Visual/Textual Features

Visual Feature Representations. Previous works usually use only one specific output of the PreDet as the *visual feature representation* (VFR). For example, Kim et al. (2018) uses the

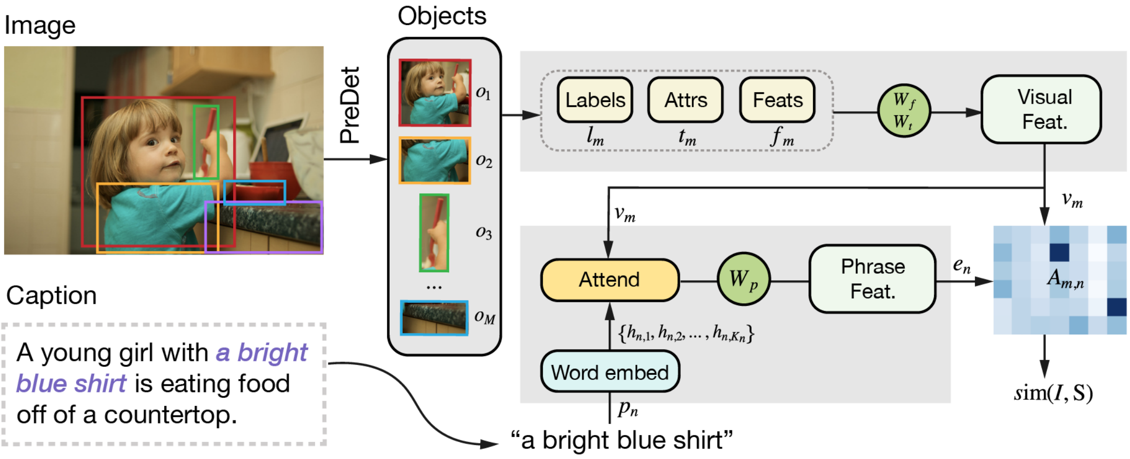


Figure 3: Overview of our proposed Multimodal Alignment Framework (MAF). A dataset of images and their captions is the input to our model. PreDet predicts bounding boxes for objects in the image and their labels, attributes, and features, which are then integrated into visual feature representations. Attention is applied between word embedding and visual representations to compute the visually-aware language representations for phrases. Finally, a multi-modal similarity function is used to measure the caption-image relevance based on the phrase-object similarity matrix.

final output feature of PreDet (denoted as $\boldsymbol{f_m}$) as the VFR, and Wang and Specia (2019) uses the label embedding (denoted as $\boldsymbol{l_m}$) of the predicted label from PreDet as the VFR. This unitary VFR usually lacks the counter-side information. Hence, we exploit different aspects of features extracted from PreDet for each object o_m in the image. In particular, we consider the output feature $\boldsymbol{f_m}$, the label embedding $\boldsymbol{l_m}$, and the attribute embedding $\boldsymbol{t_m}$ of the object o_m as the VFR,

$$\boldsymbol{v_m} = \boldsymbol{l_m} + W_t \boldsymbol{t_m} + W_f \boldsymbol{f_m}, \tag{1}$$

where W_t and W_f are two projection matrices. Naively initializing W_t and W_f will lead the model to a sub-optimal solution. In Section 4, we discuss the effectiveness of different initializations.

Textual Feature Representations. Existing works for *textual feature representation* (TFR) (Kim et al., 2018; Yu et al., 2018; Wang and Specia, 2019) commonly treat it independently of the VFR. From a different angle, we use the attention between the textual feature and the VFR $\boldsymbol{v}_m$ to integrate the visual information from the object into TFR. In more detail, we first use the GloVe embedding (Pennington et al., 2014) to encode the K_n words in the phrase p_n to $\{\boldsymbol{h}_{n,k}\}_{k=1}^{K_n}$, where $\boldsymbol{h}_{n,k} \in \mathbb{R}^d$. Here, the dimension of $\boldsymbol{h}_{n,k}$ is the same as $\boldsymbol{v_m}$. We then define a word-object matching score $a_{n,k}^m$ for each $\boldsymbol{h_{n,k}}$ in the phrase to all object features $\boldsymbol{v_m}$. In particular, for each word $\boldsymbol{h}_{n,k}$ in the phrase, we select the object with the highest matching score,

$$\begin{aligned} \boldsymbol{a}_{n,k}^m &= \operatorname*{soft\,max}_m \left\{ \frac{\boldsymbol{h}_{n,k}^T \boldsymbol{v}_m}{\sqrt{d}} \right\}, \\ \alpha_{n,k} &= \max_m \{\boldsymbol{a}_{n,k}^m\}. \end{aligned} \tag{2}$$

Finally, we normalize the attention weights for each word in the phrase p_n to obtain the final TFR, $\boldsymbol{e_n}$:

$$\begin{aligned} \beta_{n,k} &= \operatorname*{soft\,max}_k \{\alpha_{n,k}\}, \\ \boldsymbol{e_n} &= W_p \left(\sum_k \beta_{n,k} \boldsymbol{h}_{n,k} \right). \end{aligned} \tag{3}$$

where W_p is a projection matrix. In Section 4, we study the (superb) performance of the weight $\beta_{n,k}$ over simply the average $\boldsymbol{h}_{n,k}$ as well as the importance of the initialization of W_p.

3.2 Training Objective and Learning Settings

Contrastive loss. For the weakly-supervised setting, we use a contrastive loss to train our model, due to the lack of phrase-object annotations. The contrastive objective $\mathcal{L}$ aims to learn the visual and textual features by maximizing the similarity score between paired image-caption elements and minimizing the score between the negative samples (i.e., other irrelevant images). Inspired by the previous work in caption ranking (Fang et al., 2015), we use the following loss,

$$\mathcal{L} = -\log \frac{e^{\text{sim}(I,S)}}{\sum_{I' \in batch} e^{\text{sim}(I',S)}}. \tag{4}$$

Here, $\mathrm{sim}(I, S)$ is the similarity function defined below. Particularly, for each caption sentence, we use all the images $I^{'}$ in the current batch as candidate examples.

Multimodal Similarity Functions. Following the document-level dense correspondence function in Hessel et al. (2019), our multimodal similarity function is defined as:

$$\mathrm{sim}(I, S) = \frac{1}{N} \sum_{n} \max_{m} A_{n,m}. \tag{5}$$

Here, $A \in \mathbb{R}^{N \times M}$ is the phrase-object similarity matrix, and its component is computed as

$$A_{n,m} = \boldsymbol{e}_n^T \boldsymbol{v}_m, \tag{6}$$

and $\mathrm{sim}(I, S)$ measures the image-caption similarity. It is calculated based on the similarity score between each phrase in the caption and each object in the image. Note that the maximum function $\max_m A_{n,m}$ directly connects our training objective and inference target, which alleviates the discrepancy between training and inference.

Weakly-supervised setting. During training, our PreDet model is frozen. The word embeddings, W_t, W_f, and W_p are trainable parameters. Here, the word embedding is initialized with GloVe (Pennington et al., 2014). We study the different initialization methods for the rest in Section 4. During inference, for the n-th phrase p_n in an image-caption pair, we choose the localized object by

$$m_n^{\mathrm{pred}} = \arg\max_{m} A_{n,m} = \arg\max_{m} \boldsymbol{e}_n^T \boldsymbol{v}_m. \tag{7}$$

Unsupervised setting. In the unsupervised setting, the localized object is determined by

$$m_n^{\mathrm{pred}} = \arg\max_{m} \left(\sum_{k} \beta_{n,k} \boldsymbol{h}_{n,k}^T \right) \boldsymbol{l}_m. \tag{8}$$

We drop the parameters W_t, W_f, and W_p here, because there is no training in the unsupervised setting. $\beta_{n,k}$ is only calculated based on $\boldsymbol{l}_m$ (instead of $\boldsymbol{v}_m$).

4 Empirical Results

Dataset details. The Flickr30k Entities dataset contains 224k phrases and 31k images in total, where each image will be associated with 5 captions and multiple localized bounding boxes. We use 30k images from the training set for training and 1k images for validation. The test set consists of 1k images with 14,481 phrases. Our evaluation metric is the same as Plummer et al. (2015).[2] We consider a prediction to be correct if the IoU (Intersection of Union) score between our predicted bounding box and the ground-truth box is larger than 0.5. Following Rohrbach et al. (2016), if there are multiple ground-truth boxes, we use their union regions as a single ground-truth bounding box for evaluation.

Weakly-supervised Results. We report our weakly-supervised results on the test split in Table 1. We include here upper bounds (UB), which are determined by the correct objects detected by the object detectors (if available). Our MAF with ResNet-101-based Faster R-CNN detector pretrained on Visual Genome (VG) (Krishna et al., 2017) can achieve an accuracy of 61.43%. This outperforms previous weakly-supervised methods by 22.71%, and it narrows the gap between weakly-supervised and supervised methods to 15%. We also implement MAF with a VGG-based Faster R-CNN feature extractor pretrained on PASCAL VOC 2007 (Everingham et al., 2010), following the setting in KAC (Chen et al., 2018), and we use the same bounding box proposals as our ResNet-based detector. We achieve an accuracy of 44.39%, which is 5.68% higher than existing methods, showing a solid improvement under the same backbone model.

Table 1: Weakly-supervised experiment results on Flick30k Entities. (We abbreviate backbone visual feature model as “Vis. Feature,” and upper bound as “UB.”)

Method	Vis. Features	Acc. (%)	UB
Supervised			
GroundeR (Rohrbach et al., 2016)	$\mathrm{VGG_{det}}$	47.81	77.90
CCA (Plummer et al., 2015)	$\mathrm{VGG_{det}}$	50.89	85.12
BAN (Kim et al., 2018)	ResNet-101	69.69	87.45
visualBERT (Li et al., 2019)	ResNet-101	71.33	87.45
DDPN (Yu et al., 2018)	ResNet-101	73.30	-
CGN (Liu et al., 2020)	ResNet-101	76.74	-
Weakly-Supervised			
GroundeR (Rohrbach et al., 2016)	$\mathrm{VGG_{det}}$	28.93	77.90
Link (Yeh et al., 2018)	$\mathrm{YOLO_{det}}$	36.93	-
KAC (Chen et al., 2018)	$\mathrm{VGG_{det}}$	38.71	-
MAF (Ours)	$\mathrm{VGG_{det}}$	44.39	86.29
MAF (Ours)	ResNet-101	**61.43**	86.29

[2] To be specific, we use the evaluation code provided by Wang and Specia (2019) at `https://github.com/josiahwang/phraseloceval`.

Unsupervised Results.[3] We report our unsupervised results for the phrase localization method (described in Section 3.2) in Table 2. For a fair comparison, we re-implemented Wang and Specia (2019) with a Faster R-CNN model trained on Visual Genome (Krishna et al., 2017). This achieves 49.72% accuracy (similar to 50.49% as reported in their paper). Overall, our result (with VG detector) significantly outperforms the previous best result by 5.56%, which demonstrates the effectiveness of our visually-aware language representations.

Table 2: Unsupervised experiment results on Flick30k Entities. w2v-max refers to the similarity algorithm proposed in (Wang and Specia, 2019); Glove-att refers to our unsupervised inference strategy in Section 3.2; CC, OI, and PL stand for detectors trained on MS COCO (Lin et al., 2014), Open Image (Krasin et al., 2017), and Places (Zhou et al., 2017).

Method	TFR	Detector	Acc. (UB) (%)
Whole Image	None	None	21.99
(Wang and Specia, 2019)	w2v-max	Faster R-CNN	49.72 (86.29)
(Wang and Specia, 2019)	w2v-max	CC+OI+PL	50.49 (57.81)
MAF (Ours)	Glove-att	Faster R-CNN	56.05 (86.29)

Ablation Experiments. In this section, we study the effectiveness of each component and learning strategy in MAF. The comparison of different feature representations is shown in Table 3. Replacing the visual attention based TFR with an average pooling based one decreases the result from 61.43% to lower than 60%. For the VFR, using only object label $\boldsymbol{l}_m$ or visual feature $\boldsymbol{f}_m$ decreases the accuracy by 4.20% and 2.94%, respectively. One interesting finding here is that the performance with all visual features (last row) is worse than the model with only $\boldsymbol{l}_m$ and $\boldsymbol{f}_m$. Actually, we can infer that attributes cannot provide much information in localization (24.08% accuracy if used alone), partly because attributes are not frequently used to differentiate objects in Flickr30k captions.

We then investigate the effects of different initialization methods for the two weight matrices, W_f and W_p. The results are presented in Table 4. Here ZR means zero initialization, RD means random initialization with Xavier (Glorot and Bengio, 2010), and ID+RD means identity with small random noise initialization. We run each experiment for five times with different random seeds and compute the variance. According to Table 4, the best combination is zero initialization for W_f and identity+random initialization for W_p. The intuitions here are: (i) For W_f, the original label feature $\boldsymbol{l}_m$ can have a non-trivial accuracy 57.23% (see Table 3), thus using RD on initializing W_f will disturb the feature from $\boldsymbol{l}_m$; (ii) For W_p, an RD initialization will disrupt the information from the attention mechanism, while ID+RD can both ensure basic text/visual feature matching and introduce a small random noise for training.

Table 3: Ablation experiment results of different visual and textual features. TFR and VFR denotes textual and visual feature representation respectively.

TFR	VFR			Accuracy(%)
	l_m	f_m	t_m	
Average	✓			55.73
Average		✓		56.18
Average	✓	✓		59.51
Attention	✓			57.23
Attention		✓		58.49
Attention			✓	24.08
Attention	✓		✓	53.20
Attention		✓	✓	57.98
Attention	✓	✓		61.43
Attention	✓	✓	✓	60.86

Table 4: Ablation results of different initialization. (ZR: zero initialization; RD: random initialization; ID+RD: noisy identity initialization.)

W_f		W_p		Accuracy ± Var.(%)
ZR	RD	ID+RD	RD	
	✓		✓	58.54 ± 0.26
✓			✓	60.05 ± 0.31
	✓	✓		59.68 ± 0.35
✓		✓		61.28 ± 0.32

5 Conclusions

We present a Multimodal Alignment Framework, a novel method with fine-grained visual and textual representations for phrase localization, and we train it under a weakly-supervised setting, using a contrastive objective to guide the alignment between visual and textual representations. We evaluate our model on Flickr30k Entities and achieve substantial improvements over the previous state-of-the-art methods with both weakly-supervised and unsupervised training strategies. Detailed analysis is also provided to help future works investigate other critical feature enrichment and alignment methods for this task.

[3] More unsupervised results are available in Appendix B.

Acknowledgments

We thank reviewers and the area chair for their helpsuggestions. We are grateful for a gracious fund from the Amazon AWS. MWM would also like to acknowledge DARPA, NSF, and ONR for providing partial support of this work. HT is supported by Bloomberg Data Science Ph.D. Fellowship. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12476–12486.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2Ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2601–2610.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338.

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

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2034–2045.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligenc*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

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

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Josiah Wang and Lucia Specia. 2019. Phrase localization without paired training examples. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4663–4672.

Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Raymond A Yeh, Minh N Do, and Alexander G Schwing. 2018. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6125–6134.

Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking diversified and discriminative proposal generation for visual grounding. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

A Implementation Details

For GloVE word embeddings, we use the one with the hidden dimension 300. Phrases are split into words by space. We replace all out-of-vocabulary words with the introduced ⟨UNK⟩ token. For object proposals, we apply an off-the-shelf Faster R-CNN model (Ren et al., 2015) as the object detector[4] for object pseudo-labels. The backbone of the detector is ResNet-101 (He et al., 2016), and it is pre-trained on Visual Genome with mAP=10.1. We keep all bounding boxes with a confidence score larger than 0.1. For ResNet-based visual features, we use the 2048-dimensional feature from Bottom-up attention (Anderson et al., 2018a), which is pre-trained with 1600 object labels and 400 attributes.

The extracted visual features are frozen during training, and we use a batch size of 64 during training. Our optimizer is Adam with learning rate $lr = 1e^{-5}$. Except for word embeddings, trainable parameters include $W_t \in \mathbb{R}^{d_T \times d_T}$, $W_f \in \mathbb{R}^{d_V \times d_T}$, and $W_p \in \mathbb{R}^{d_T \times d_T}$, where $d_T = 300$, $d_V = 2048$ for ResNet-101 backbone and $d_V = 4096$ for VGG backbone. During training, it takes around 350 seconds to train an epoch using a single Tesla K80. We train our model for 25 epochs and report the results at the last epoch.

[4] https://github.com/jwyang/faster-rcnn.pytorch

B Baselines

In Table 5, we report the results of different unsupervised methods:

- Random: Randomly localize to a detected object.
- Center-obj: Localize to the object which is closest to the center of image, where we use an L_1 distance $D = |x - x_{\text{center}}| + |y - y_{\text{center}}|$.
- Max-obj: Localize to the object with the maximal area.
- Whole Image: Always localize to the whole image.
- Direct Match: Localize with the direct match between object labels and words in the phrase, e.g., localize “a red apple” to the object with the label “apple.” If multiple labels are matched, we choose the one with the largest bounding box.
- Glove-max: Consider every word-label similarity independently and select the object label with the highest semantic similarity with any word.
- Glove-avg: Represent a phrase using an average pooling over Glove word embeddings and select the object label with highest the semantic similarity with the phrase representation.
- Glove-att: Use our visual attention based phrase representation, as is described in the Methodology 3.1.

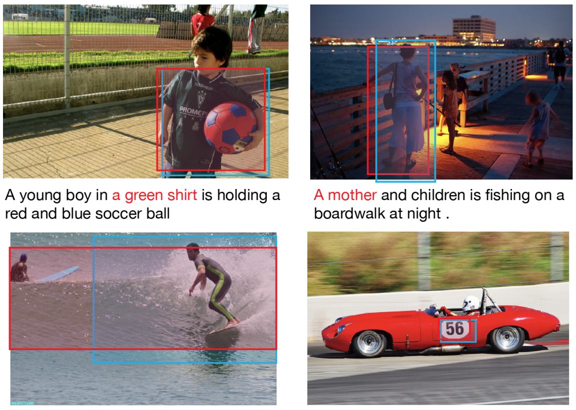
Note that in all label-based methods (Direct Match (Wang and Specia, 2019), and our unsupervised method), if multiple bounding boxes share the same label, we choose the largest one as the predicted box.

C Qualitative Analysis

To analyze our model qualitatively, we show some visualization results in Figure 4 and Figure 5. Figure 4 shows examples with consistent predictions between supervised and unsupervised models. In these cases, both methods can successfully learn to localize various objects, including persons (“mother”), clothes (“shirt”), landscapes (“wave”), and numbers (“56”). Figure 5 shows examples where supervised and unsupervised methods localize to different objects. In the first image, they both localize the phrase “entrance” incorrectly. In the remaining three images, the supervised method

Table 5: Baseline results of unsupervised methods on Flick30k Entities. Abbreviations are explained above.

Method	Detector	Acc. (%)
Random	Faster R-CNN	7.19
Center-obj	Faster R-CNN	18.24
Whole Image	None	21.99
Max-obj	Faster R-CNN	24.51
Direct match	Faster R-CNN	26.42
Glove-max	Faster R-CNN	26.28
Glove-avg	Faster R-CNN	54.51
Glove-att	Faster R-CNN	56.05



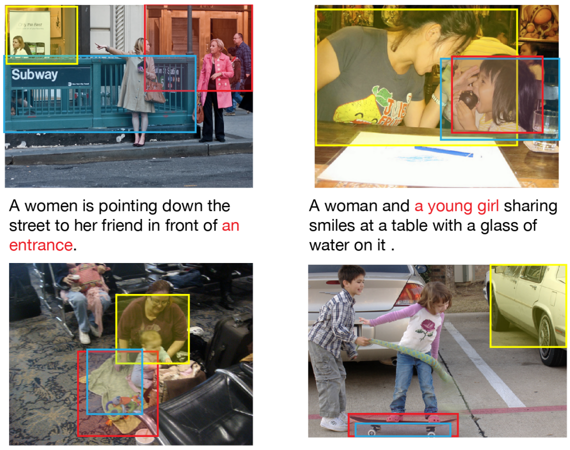
A young boy in a green shirt is holding a red and blue soccer ball

A mother and children is fishing on a boardwalk at night .

A surfer wearing a black and green wetsuit riding a wave.

A number 56 red racing car is speeding left past the frame .

Figure 4: Example of predictions on Flickr30k. (Red box: ground truth, blue box: our prediction).



A women is pointing down the street to her friend in front of an entrance.

A woman and a young girl sharing smiles at a table with a glass of water on it .

Near seats , woman has spread out green blanket to sit on floor with baby .

A boy hitting a girl on a skateboard with a plushie snake .

Figure 5: Example of predictions on Flickr30k. (Red box: ground truth, blue box: supervised prediction, yellow box: unsupervised prediction)

learns to predict a tight bounding box on the correct object, while the unsupervised method localizes to other irrelevant objects. For example (bottom left figure for Figure 5), if the object detector fails to detect the "blanket," then the unsupervised method can never localize "green blanket" to the right object. Still, the supervised method can learn from negative examples and obtain more information.