Contrastive Learning for Weakly Supervised Phrase Grounding

Tanmay Gupta[1][*][†], Arash Vahdat[3], Gal Chechik[2,3], Xiaodong Yang[3], Jan Kautz[3], and Derek Hoiem[1]

[1] University of Illinois Urbana-Champaign
[2] Bar Ilan University
[3] NVIDIA

Abstract. Phrase grounding, the problem of associating image regions to caption words, is a crucial component of vision-language tasks. We show that phrase grounding can be learned by optimizing word-region attention to maximize a lower bound on mutual information between images and caption words. Given pairs of images and captions, we maximize compatibility of the attention-weighted regions and the words in the corresponding caption, compared to non-corresponding pairs of images and captions. A key idea is to construct effective negative captions for learning through language model guided word substitutions. Training with our negatives yields a $\sim 10\%$ absolute gain in accuracy over randomly-sampled negatives from the training data. Our weakly supervised phrase grounding model trained on COCO-Captions shows a healthy gain of 5.7% to achieve 76.7% accuracy on Flickr30K Entities benchmark. Our code and project material will be available at http://tanmaygupta.info/info-ground.

1 Introduction

Humans can learn from captioned images because of their ability to associate words to image regions. For instance, humans perform such word-region associations while acquiring facts from news photos, making a diagnosis from MRI scans and radiologist reports, or enjoying a movie with subtitles. This word-region association problem is called word or phrase *grounding* and is a crucial capability needed for downstream applications like visual question answering, image captioning, and text-image retrieval.

Existing object detectors can detect and represent object regions in an image, and language models can provide contextualized representations for noun phrases in the caption. However, learning a mapping between these continuous, independently trained visual and textual representations is challenging in the
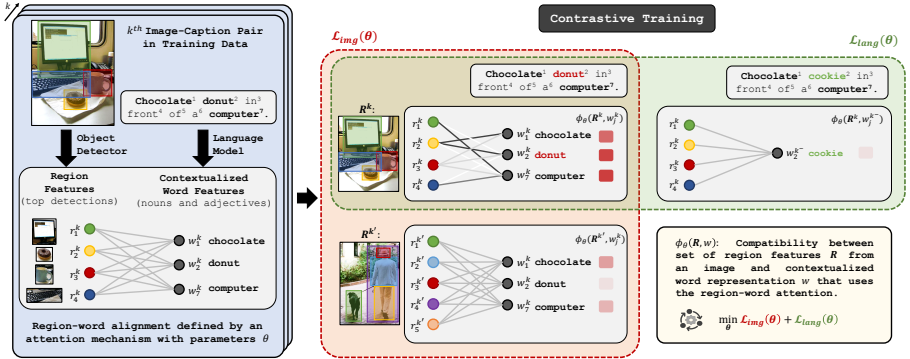

Fig. 1. Overview of our contrastive learning framework. We begin by extracting region and word features using an object detector and a language model respectively. Contrastive learning trains a word-region attention mechanism as part of a compatibility function ϕ_θ between the set of region features from an image and individual contextualized word representations. The compatibility function is trained to maximize a lower bound on mutual information with two losses. For a given caption word, $\mathcal{L}_{\mathrm{img}}$ learns to produce a higher compatibility for the true image than a negative image in the mini-batch. $\mathcal{L}_{\mathrm{lang}}$ learns to produce a higher compatibility of an image with a true caption-word than with a word in a negative caption. We construct negative captions by substituting a noun word like "donut" in the true caption with contextually plausible but untrue words like "cookie" using a language model.

absence of explicit region-word annotations. We focus on learning this mapping from weak supervision in the form of paired image-caption data without requiring laborious grounding annotations.

Current state-of-the-art approaches [11,1,33] formulate weakly supervised phrase grounding as a multiple instance learning (MIL) problem [25,18]. The image can be viewed as a bag of regions. For a given phrase, all images with captions containing the phrase are treated as positive bags while remaining images are treated as negatives. Models aggregate per region features or phrase scores to construct image-level predictions that can be supervised with image-level labels in the form of phrases or captions. Common aggregation approaches include max or mean pooling, noisy-OR [13], and attention [11,18]. Popular training objectives include binary classification loss [13] (whether the image contain the phrase) or caption reconstruction loss [33] (generalization of binary classification to caption prediction) or ranking objectives [1,11] (do true image-caption or image-phrase pairs score higher than negative pairs).

Fig. 1 provides an overview of our proposed contrastive training. We propose a novel formulation of the weakly supervised phrase grounding problem as that of maximizing a lower bound on mutual information between set of region features extracted from an image and contextualized word representations. We use pre-trained region and word representations from an object detector and a language

model and perform optimization over parameters of word-region attention instead of optimizing the region and word representations themselves. Intuitively, to compute mutual information with a word's representation, attention must discard nuisance regions in the word-conditional attended visual representation, thereby selecting regions that match the word. For any given word, the learned attention thus functions as a soft selection or grounding mechanism over regions.

Since computing MI is intractable, we maximize the recently introduced InfoNCE lower bound [30] on mutual information. The InfoNCE bound requires a compatibility score between each caption word and the image to contrast positive image and caption word pairs with negative pairs in a minibatch. We use two objectives. The first objective ($\mathcal{L}_{\texttt{img}}$ in Fig. 1) contrasts a positive pair with negative pairs with the same caption word but different image regions. The second objective ($\mathcal{L}_{\texttt{lang}}$ in Fig. 1) contrasts a positive pair with negative pairs with the same image but different captions. We show empirically that sampling negative captions randomly from the training data to optimize $\mathcal{L}_{\texttt{lang}}$ does not yield any gains over optimizing $\mathcal{L}_{\texttt{img}}$ only. Instead of random sampling, we propose to use a language model to construct context-preserving negative captions by substituting a single noun word in the caption.

We design the compatibility function using a `query-key-value` attention mechanism. The `queries` and `keys`, computed from words and regions respectively, are used to compute a word-specific attention over each region which acts as a soft alignment or grounding between words and regions. The compatibility score between regions and word is computed by comparing attended visual representation and the word representation.

Our key contributions are: (i) a novel MI based contrastive training framework for weakly supervised phrase grounding; (ii) an InfoNCE compatibility function between a set of regions and a caption word designed for phrase grounding; and (iii) a procedure for constructing context-preserving negative captions that provides $\approx 10\%$ absolute gain in grounding performance.

1.1 Related Work

Our work is closely related to three active areas of research. We now provide an overview of prior arts in each.

Weakly Supervised Phrase Grounding. Weakly supervised phrase localization is typically posed as a multiple instance learning (MIL) problem [25,18] where each image is considered as a bag of region proposals. Images whose captions mention a word or a phrase are treated as positive bags while rest of the images are treated as negatives for that word or phrase. Features or scores for a phrase or the entire caption are aggregated across all regions to make a prediction for the image. Common methods of aggregation are max or average pooling, noisy-OR [13], or attention [33,18]. With the ability to produce image-level scores for pairs of images and phrases or captions, the problem becomes an image-level fully-supervised phrase classification problem [13] or an image-caption retrieval problem [1,11]. An alternatives to the MIL formulations is the approach of Ye *et*

al. [44] which uses statistical hypothesis testing approach to link concepts detected in an image and words mentioned in the sentence. While all the above approaches assume paired image-caption data, Wang *et al.* [42] recently address the problem of phrase grounding without access to image-caption pairs. Instead they assume access to a set of scene and color classifiers, and object detectors to detect concepts in the scene and use word2vec [27] similarity between concept labels and caption words to achieve grounding.

MI-based Representation Learning. Recently MI-based approaches have shown promising results on a variety representation learning problems. Computing the MI between two representations is challenging as we often have access to samples but not the underlying joint distribution that generated the samples. Thus, recent efforts rely on variational estimation of MI [3,20,6,30]. An overview of such estimators is discussed in [31,40] while the statistical limitations are reviewed in [26,34].

In practice, MI-based representation learning models are often trained by maximizing an estimation of MI across different *transformations* of data. For example, deep InfoMax [17] maximizes MI between local and global representation using MINE [6]. Contrastive predictive coding [30,16] inspired by noise contrastive estimation [14,29] assumes an order in the features extracted from an image and uses summary features to predict future features. Contrastive multi-view coding [39] maximizes MI between different color channels or data modalities while augmented multiscale Deep InfoMax [5] and SimCLR [8] extract views using different augmentations of data points. Since the infoNCE loss is limited by the batch size, several previous work rely on memory banks [43,28,15] to increase the set of negative instances.

Joint Image-Text Representation Learning. With the advances in both visual analysis and natural language understanding, there has been a recent shift towards learning representation jointly from both visual and textual domains [23,35,24,37,38,45,22,9,2,36]. ViLBERT [24] and LXMERT [38] learn representation from both modalities using two-stream transformers, applied to image and text independently. In contrast, UNITER [9], VisualBERT [23], Unicoder-VL [22], VL-BERT [35] and B2T2 [2] propose a unified single architecture that learns representation jointly from both domains. Our method is similar to the first group, but differs in its fundamental goal. Instead of focusing on learning a task-agnostic representation for a range of downstream tasks, we are interested in the quality of region-phrase grounding emerged by maximizing mutual information. Moreover, we rely on the language modality as a weak training signal for grounding, and we perform phrase-grounding without any further finetuning.

2   Method

Consider the set of region features and contextualized word representation as two multivariate random variables. Intuitively, estimating MI between them requires

extracting the information content shared by these two variables. We model this MI estimation as maximizing a lower bound on MI with respect to parameters of a word-region attention model. This maximization forces the attention model to downweight regions from the image that do not match the word, and to attend to the image regions that contain the most shared information with the word representation.

Sec. 2.1 describes MI and the InfoNCE lower bound. Sec. 2.2 introduces notation and InfoNCE based objective for learning phrase grounding from paired image caption data. Sec. 2.3 presents the design of a word-region attention based compatibility function which is part of the InfoNCE objective.

2.1   InfoNCE Lower Bound on Mutual Information

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be random variables drawn from a joint distribution with density $p(x, y)$. The MI between x and y measures the amount of information that these two variables share:

$$\mathrm{MI}(x, y) = \mathbb{E}_{(x,y) \sim p(x,y)} \left[\log \frac{p(x,y)}{p(x)p(y)} \right], \tag{1}$$

which is also the KullbackLeibler Divergence from $p(x, y)$ to $p(x)p(y)$.

However, computing MI is intractable in general because it requires a complete knowledge of the joint and marginal distributions. Among the existing MI estimators, the InfoNCE [30] lower bound provides a low-variance estimation of MI for high dimensional data, albeit being biased [31]. The appealing variance properties of this estimator may explain its recent success in representation learning [8,30,16,36]. InfoNCE defines a lower bound on MI by:

$$\mathrm{MI}(x, y) \geq \log(k) - \mathcal{L}_k(\theta). \tag{2}$$

Here, $\mathcal{L}_k$ is the InfoNCE objective defined in terms of a compatibility function ϕ parametrized by θ: $\phi_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The lower bound is computed over a mini-batch $\mathcal{B}$ of size k, consisting of one positive pair $(x, y) \sim p(x, y)$ and $k-1$ negative pairs $\{(x'_i, y)\}_{i=1}^{k-1}$ where $x' \sim p(x)$:

$$\mathcal{L}_k(\theta) = \mathbb{E}_{\mathcal{B}} \left[-\log \left(\frac{e^{\phi_\theta(x,y)}}{e^{\phi_\theta(x,y)} + \sum_{i=1}^{k-1} e^{\phi_\theta(x'_i,y)}} \right) \right]. \tag{3}$$

Oord *et al.* [30] showed that maximizing the lower bound on MI by minimizing $\mathcal{L}_k$ with respect to θ leads to a compatibility function $\phi_{\theta*}$ that obeys

$$e^{\phi_{\theta*}(x,y)} \propto \frac{p(x|y)}{p(x)} = \frac{p(x,y)}{p(x)p(y)}, \tag{4}$$

where θ^* is the optimal θ obtained by minimizing $\mathcal{L}_k$.

2.2   InfoNCE for Phrase Grounding

Recent work [11] has shown that pre-trained object detectors such as Faster-RCNN [32] and language models such as BERT [12] provide rich representations in the visual and textual domains for the phrase grounding problem. Inspired by this, we aim to maximize mutual information between region features generated by an object detector and contextualized word representation extracted by a language model.

Let us denote image region features for an image by $\mathbf{R} = \{r_i\}_{i=1}^m$ where m is the number of regions in the image with each $r_i \in \mathbb{R}^{d_r}$. Similarly, caption word representations are denoted as $\mathbf{W} = \{w_j\}_{j=1}^n$ where n is the number of words in the caption with each word represented as $w_j \in \mathbb{R}^{d_w}$.

We maximize the InfoNCE lower bound on MI between image regions and each individual word representation denoted by $\mathrm{MI}(\mathbf{R}, w_j)$. Thus using Eq. 2 we maximize the following lower bound:

$$\sum_{j=1}^n \mathrm{MI}(\mathbf{R}, w_j) \geq n \log(k) - \sum_{j=1}^n \mathcal{L}_{kj}(\theta). \tag{5}$$

We empirically show that maximizing the lower bound in Eq. 5 with an appropriate choice of compatibility function ϕ_θ results in learning phrase grounding without strong grounding supervision. The following section details the design of the compatibility function.

2.3   Compatibility Function with Attention

The InfoNCE loss in our phrase grounding formulation requires a compatibility function between the *set* of region feature vectors $\mathbf{R}$ and the contextualized word representation w_j. To define the compatibility function, we propose to use a `query-key-value` attention mechanism [41]. Specifically, we define neural modules $k_r, v_r : \mathbb{R}^{d_r} \to \mathbb{R}^d$ to map each image region to `keys` and `values` and $q_w, v_w : \mathbb{R}^{d_w} \to \mathbb{R}^d$ to compute `query` and `values` for the words. The `query` vectors for each word are used to compute the attention score for every region given a word using

$$a(r_i, w_j) = \frac{e^{s(r_i, w_j)}}{\sum_{i'=1}^m e^{s(r_{i'}, w_j)}}, \tag{6}$$

where $s(r_i, w_j) = q_w(w_j)^T k_r(r_i)/\sqrt{d}$. The attention scores are used as a soft selection mechanism to compute a word-specific visual representation using a linear combination of region `values`

$$v_{att}(\mathbf{R}, w_j) = \sum_{i=1}^m a(r_i, w_j) v_r(r_i). \tag{7}$$

Finally, the compatibility function is defined as $\phi_\theta(\mathbf{R}, w_j) = v_w^T(w_j) v_{att}(\mathbf{R}, w_j)$, where θ refers to the parameters of neural modules $k_r, v_r, q_w,$ and v_w, implemented using simple feed-forward MLPs. Following Eqs. 3 & 5, the InfoNCE
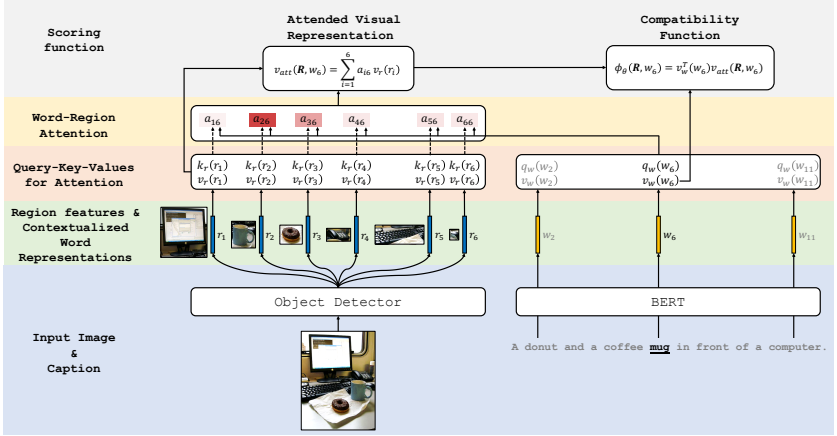
Fig. 2. Compatibility function ϕ_θ with word-region attention. The figure shows compatibility computation between the set of image regions and the word "mug" in the caption. The compatibility function consists of learnable `query-key-value` functions k_r, v_r, q_w, v_w. The `query` constructed from contextualized representation of the word "mug" is compared to `keys` created from region features to compute attention scores. The attention scores are used as weights to linearly combine `values` created from region features to construct an attended visual representation for "mug". The compatibility is defined by the dot product of the attended visual representation and `value` representation for "mug".

loss for phrase grounding is defined as

$$\mathcal{L}_{\texttt{img}}(\theta) = \mathbb{E}_\mathcal{B}\left[-\sum_{j=1}^n \log\left(\frac{e^{\phi_\theta(\mathbf{R}, w_j)}}{e^{\phi_\theta(\mathbf{R}, w_j)} + \sum_{i=1}^{k-1} e^{\phi_\theta(\mathbf{R}'_i, w_j)}} \right) \right]. \qquad (8)$$

which is marked using subscript `img` as negative pairs are created by replacing image regions from a positive pair with regions extracted from negative instance in the mini-batch.

Remark: We enforce compatibility between each word and all image regions using $\text{MI}(\mathbf{R}, w_j)$ in Eq. 5, but not between a region and all caption words ($\text{MI}(r_i, \mathbf{W})$). This is because the words only describe part of the image, so there will be regions with no corresponding word in the caption.

2.4   Context-Preserving Negative Captions

The objective in Eq. 8 trains the compatibility function by contrasting positive regions-word pairs against pairs with replaced image regions. We now propose a complementary objective function that contrasts the positive pairs against negative pairs whose captions are replaced with plausible negative captions. However,

Caption	Negatives Selected After Reranking	Candidates Rejected After Reranking
A man is seated at a counter with all types of delicious looking foods, yet is completely unaffected, casually reading his <u>newspaper</u>.	menu, books, phone, scripts, email, messages, bible, tablet	newspaper, paper, journal, article, magazine
A BMX bike rider in red clothing and a helmet is riding his bike next to a wooden <u>fence</u>.	bench, pole, statue, door, table, chair, sign, platform, piano	fence, gate, wall, railing, screen
A man in a blue jumpsuit stands next to a red <u>van</u> pulling a trailer.	bike, sedan, horse, jeep, cart, car, tractor, bull, engine, motorcycle	van, trailer, vehicle, light, truck
A <u>man</u> and a boy are playing with a dog in the evening.	girl, lady, mother, woman, teenager, child, teacher, mom	man, boy, guy, couple, youth
A <u>woman</u> in a brown sweater sits at a table covered with food.	boy, guy, gentleman, kid, nurse, soldier, waiter, priest, child	woman, female, person, face, lady
A man with <u>shorts</u> and a hat is holding onto a little boy and a dog.	gloves, glasses, coat, trousers, bags, apron, moustache, beard	shorts, ties, pants, stripes, jeans

Fig. 3. Context-preserving negative captions. We construct negative captions which share the same context as the true caption but substitute a noun word. We choose the substitute using a language model such that it is plausible in the context but we reject potential synonyms or hypernyms of the original word by a re-ranking procedure.

extracting negative captions that are related to a captions is challenging as it requires semantic understanding of words in a caption. Here, we leverage BERT as a pretrained bidirectional language model to extract such negative captions.

For a caption with a noun word s and context c, we define a context-preserving negative caption as one which has the same context c but a different noun s' with the following properties: (i) s' should be plausible in the context; and (ii) the new caption defined by the pair (s', c) should be untrue for the image. For example, consider the caption `"A man is walking on a beach"` where s is chosen as `"man"` and c is defined by `"A [MASK] is walking on a beach"` where `[MASK]` is the token that denotes a missing word. A potential candidate for a context-preserving negative caption might be `"A woman is walking on a beach"` where s' is `woman`. However, `"A car is walking on a beach"` and `"A person is walking on a beach"` are not negative captions because `car` is not plausible given the context, and the statement with `person` is still true given that the original caption is true for the image.

Constructing context-preserving negative captions. We propose to use a pre-trained BERT language model to construct context-preserving negative captions for a given true caption. Our approach for extracting such words consists of two steps: First, we feed the context c into the language model to extract 30 most likely candidates $\{s'_l\}_{l=1}^{30}$ for the masked word using probabilities $p(s'|c)$ predicted by BERT. Intuitively, these words correspond to those that fill in the masked word in caption according to BERT. However, the original masked word or its synonyms may be present in the set as well. Thus, in the second step, we pass the original caption into BERT to compute $q(s'_l|s, c)$ which we use as a proxy for how true (s'_l, c) is given that (s, c) is true. We re-rank the candidates using the score $\frac{p(s'|c)}{q(s'|s,c)}$ and we keep the top 25 captions $\{(s'_l, c)\}_{l=1}^{25}$ as negatives for the original caption (s, c).

We empirically find that the proposed approach is effective in extracting context-preserving negative captions. Fig. 3 shows a context-preserving negatives for a set of captions along with candidates that were rejected after re-ranking. Note that the selected candidates match the context and the rejected candidates are often synonyms or hypernyms of the true noun.

Training with context-preserving negative captions. Given the context-preserving negative captions, we can train our compatibility function by contrasting the positive pairs against negative pairs with plausible negative captions. We use a loss function similar to InfoNCE to encourage higher compatibility score of an image with the true caption than any negative caption. Let w and $\{w'_l\}_{l=1}^{25}$ denote the contextualized representation of the positive word s and the corresponding negative noun words $\{s'_l\}_{l=1}^{25}$. The language loss is defined as

$$\mathcal{L}_{\text{lang}}(\theta) = \mathbb{E}_{\mathcal{B}}\left[-\log\left(\frac{e^{\phi_\theta(\mathbf{R},w)}}{e^{\phi_\theta(\mathbf{R},w)} + \sum_{l=1}^{25} e^{\phi_\theta(\mathbf{R},w'_l)}}\right)\right]. \tag{9}$$

For captions with multiple noun words, we randomly select s from the noun words for simplicity.

2.5 Implementation Details

Regions and Visual Features. We use the Faster-RCNN object detector provided by Anderson *et al.* [4] and used for extracting visual features in the current state-of-the-art phrase grounding approach Align2Ground [11]. The detector is trained jointly on Visual Genome object and attribute annotations and we use a maximum of 30 top scoring bounding boxes per image with 2048 dimensional ROI-pooled region features.

Contextualized Word Representations. We use a pretrained BERT language model to extract 768 dimensional contextualized word representations for each caption word. Note that BERT is trained on a text corpora using masked language model training where words are randomly replaced by a `[MASK]` token in the input and the likelihood of the masked word is maximized in the distribution over vocabulary words predicted at the output. Thus, BERT is trained to model distribution over words given context and hence suitable for modeling $p(s|c)$ defined in Sec. 2.4 for constructing context-preserving negative captions.

Query-Key-Value Networks. We use an MLP with 1 hidden layer for each of k_r, v_r, q_w, v_w for all experiments except the ablation in Fig. 4. We use Batch-Norm [19] and ReLU activations after the first linear layer. The hidden layer has the same number of neurons as the input dimensions of these networks which are 2048 for (k_r, v_r), and 768 for (q_w, v_w). The output layer is 384 (= 768/2) for all networks.

Losses. Since we only care about grounding noun phrases, we compute $\mathcal{L}_{\text{img}}$ only for noun and adjective words in the captions as identified by a POS tagger instead of all caption words for computation efficiency.

Optimization. We optimize $\mathcal{L}_{img} + \mathcal{L}_{lang}$ computed over batches of 50 image-caption pairs using the ADAM optimizer [21] with a learning rate of 10^{-5}. We compute $\mathcal{L}_{\text{img}}$ for each image using other images in the batch as negatives.

Attention to phrase grounding. We use the BERT tokenizer to convert captions into individual word or sub-word tokens. Attention is computed per token. For evaluation, the phrase-level attention score for each region is computed as the maximum attention score assigned to the region by any of the tokens in the phrase. The regions are then ranked according to this phrase level score.

3   Experiments

Our experiments compare our approach to state-of-the-art on weakly supervised phrase localization (Sec. 3.2), ablate gains due to pretrained language representations and context-preserving negative sampling using a language model (Sec. 3.3), and analyse the relation between phrase grounding performance and the InfoNCE bound that we optimize as a proxy for phrase grounding (Sec. 3.4).

3.1   Datasets and Metrics

We train our models on image-caption pairs from COCO training set which consists of $\sim 83K$ training images. We use the validation set with $\sim 41K$ images for part of our analysis. Each image is accompanied with 5 captions. For evaluation, we use the Flickr30K Entities validation set for model selection (early stopping) and test set for reporting final performance. Both sets consist of 1K images with 5 captions each. We report two metrics:

Recall@k which is the fraction of phrases for which the ground truth bounding box has an IOU ≥ 0.5 with any of the top-k predicted boxes.

Pointing accuracy which requires the model to predict a single point location per phrase and the prediction is counted as correct if it falls within the ground truth bounding box for the phrase. Unlike recall@k, pointing accuracy does not require identifying the extent of the object. Since our model selects one of the detected regions in the image, we use use center of the selected bounding box as the prediction for each phrase for computing pointing accuracy.

3.2   Performance on Flickr30K Entities

Tab. 1 compares performance of our method to existing weakly supervised

Table 1. Grounding performance on Flickr30K Entities test set. We make our approach directly comparable to the current state-of-the-art, Align2Ground [11]. The performance of older methods are reported for completeness but the use of different visual features makes direct comparison difficult.

Method	Training Data	Visual Features	R@1	R@5	R@10	Accuracy
GroundeR (2015) [33]	Flickr30K Entities	VGG-det (VOC)	28.94	-	-	-
Yeh *et al.*(2018) [44]	Flickr30K Entities	VGG-cls (IN)	22.31	-	-	-
Yeh *et al.*(2018) [44]	Flickr30K Entities	VGG-det (VOC)	35.90	-	-	-
Yeh *et al.*(2018) [44]	Flickr30K Entities	YOLO (COCO)	36.93	-	-	-
KAC Net+Soft KBP (2018) [7]	Flickr30K Entities	VGG-det (VOC)	38.71	-	-	-
Fang *et al.*(2015) [13]	COCO	VGG-cls (IN)	-	-	-	29.00
Akbari *et al.*(2019) [1]	COCO	VGG-cls (IN)	-	-	-	61.66
Akbari *et al.*(2019) [1]	COCO	PNAS Net (IN)	-	-	-	69.19
Align2Ground (2019) [11]	COCO	Faster-RCNN (VG)	-	-	-	71.00
Ours	Flickr30K Entities	Faster-RCNN (VG)	**47.88**	**76.63**	**82.91**	**74.94**
Ours	COCO	Faster-RCNN (VG)	**51.67**	**77.69**	**83.25**	**76.74**

phrase grounding approaches on the Flickr30K Entities test set. A few existing approaches train on Flickr30K Entities train set and report recall@1 while recent methods use COCO train set and report pointing accuracy. Further, all approaches use different visual features making direct comparison difficult. For a fair comparison to state-of-the-art, we use Faster-RCNN trained on Visual Genome object and attribute annotations used in Align2Ground [11] and report performance for models trained on either datasets on both recall and pointing accuracy metrics.

Using the same training data and visual feature architecture, our model shows a 5.7% absolute gain in pointing accuracy over Align2Ground. Learning using our contrastive formulation is also quite sample efficient as can be seen by only a 2 to 3 points drop in performance when the model is trained on the much smaller Flickr30K Entities train set which has approximately one-third as many image-caption pairs as COCO.

3.3   Benefits of Language Modeling

Our approach benefits from language modeling in two ways: (i) using the pretrained language model to extract contextualized word representations, and (ii) using the language model to sample context-preserving negative captions. Tab. 2 evaluates along both of these dimensions.

Gains from pretrained word representations. In Tab. 2, BERT (Random) refers to the BERT architecture initialized with random weights and finetuned on COCO image-caption data along with parameters of the attention mechanism. BERT (Pretrained) refers to the off-the-shelf pretrained BERT model which is used as a contextualized word feature extractor during contrastive learning without finetuning. We observe a ~10% absolute gain in both recall@1 and pointing accuracy by using pretrained word representations from BERT.

Table 2. Benefits of language modeling. The first two rows show the gains due to pretrained language representations. The next three rows show gains from each step in our proposed context-preserving negative caption construction.

Negative Captions	Language Model	R@1	R@5	R@10	Accuracy
None	BERT (Random)	25.66	59.57	75.16	57.37
None	BERT (Pretrained)	35.74	72.91	82.07	66.89
Random	BERT (Pretrained)	36.32	72.42	81.81	66.92
Contextually plausible	BERT (Pretrained)	**48.05**	**76.78**	**82.97**	**74.91**
Excluding near-synonyms & hypernyms	BERT (Pretrained)	**51.67**	**77.69**	**83.25**	**76.74**

Gains from context-preserving negative caption sampling. Our context-preserving negative sampling has two steps. The first step is drawing negative noun candidates given the context provided by the true caption. The second step is re-ranking the candidates to filter out likely synonyms or hypernyms that are also true for the image.

First, note that randomly sampling negative captions from training data for computing $\mathcal{L}_{\text{lang}}$ performs similarly to only training using $\mathcal{L}_{\text{img}}$. Model trained with contextually plausible negatives significantly outperforms random sampling by $\geq 8\%$ gain in recall@1 and pointing accuracy. Excluding near-synonyms and hypernyms yields another ~ 3 points gain in recall@1 and accuracy.

3.4   Is InfoNCE a good proxy for learning phrase grounding?

The fact that optimizing our InfoNCE objective results in learning phrase grounding is intuitive but not trivial. Fig. 4 shows that maximizing the InfoNCE lower bound correlates well with phrase grounding performance on a heldout dataset. We make several interesting observations: **(i)** As training progresses (from left to right), InfoNCE lower bound (Eq. 5) mostly keeps increasing on the validation set. This indicates that there is no overfitting in terms of the InfoNCE bound. **(ii)** With the increase in InfoNCE lower bound, phrase grounding performance first increases until peak performance and then decreases. This shows that the InfoNCE bound is correlated with the grounding performance but maximizing it fully does not necessarily yield the best grounding. A similar observation has been made in [39] for representation learning. **(iii)** The peak performance and the number of iterations needed for the best performance depends on the choice of `key-value-query` modules. One and two layer MLPs hit the peak faster and perform better than linear functions. We refer the reader to Sec. A.1 in the appendix for a discussion of limitations of our approach.
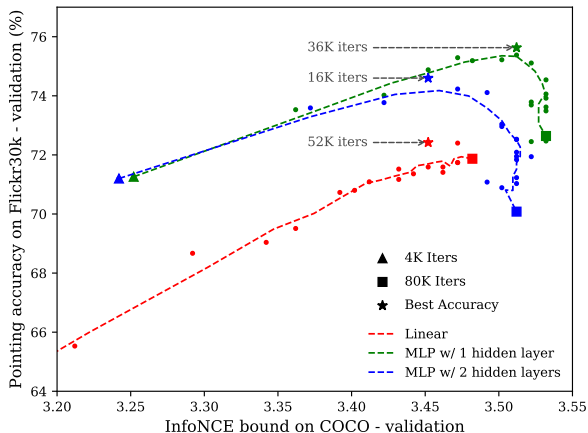
Fig. 4. Relation between `InfoNCE` lower bound and phrase grounding performance with training iterations for 3 different choices of `key-value` modules in the compatibility function ϕ_θ. Each epoch is $\sim$ 8K iterations. The scattered points visualize the measured quantities during training. The dashed lines are created by applying moving average to highlight the trend.

3.5 Qualitative Results

Fig. 5 visualizes the word-region attention learned by our model. The qualitative results demonstrate the following abilities: **(i)** localizing different objects mentioned in the same caption with varying degrees of semantic relatedness, e.g., `man` and `canine` in row 1 vs. `man` and `woman` in row 3; **(ii)** disambiguation between two instances of the same object category using caption context. For example, `boy` and `another` in row 4 and bride and groom from other men and women in row 3; **(iii)** localizing object parts such as toddler's `shirt` in row 2 and instrument's `mouthpiece` in row 5; **(iv)** handling occlusion, e.g., `table` covered with toys in row 6; **(v)** handling uncommon words or categories like `ponytail` and `mouthpiece` in row 5 and `hose` in row 7.

4 Conclusion

In this work, we offer a novel perspective on weakly supervised phrase grounding from paired image-caption data which has traditionally been cast as a multiple instance learning problem. We formulate the problem as that of estimating mutual information between image regions and caption words. We demonstrate that maximizing a lower bound on mutual information with respect to parameters of a region-word attention mechanism results in learning to ground words in images. We also show that language models can be used to generate context-preserving negative captions which greatly improve learning in comparison to randomly sampling negatives from training data.

(1)

A <u>man</u> and a <u>canine</u> both stand on a snowy plane looking out into the distance.

#boxes:27   man:0.26,0.17,0.09   canine:0.33,0.25,0.06

(2)

A <u>toddler</u> in a yellow <u>shirt</u> standing in front of a living complex next to a baby carriage.

#boxes:34   toddler:0.59,0.20,0.06   shirt:0.63,0.17,0.08

(3)

A <u>man</u> in a tuxedo and a <u>woman</u> in a bride's gown are leaving a church.

#boxes:50   man:0.17,0.09,0.09   woman:0.19,0.15,0.08

(4)

One <u>boy</u> follows <u>another</u> at the park.

#boxes:37   boy:0.25,0.18,0.18   another:0.12,0.10,0.08

(5)

A man with a <u>ponytail</u> wearing a blue collared shirt is playing an instrument's <u>mouthpiece</u>.

#boxes:12   ponytail:0.29,0.19,0.10   mouthpiece:0.32,0.16,0.09

(6)

Two <u>kids</u> sitting at a <u>table</u> full of toys.

#boxes:34   kids:0.22,0.17,0.16   table:0.33,0.15,0.10

(7)

A <u>curly-haired</u> little girl watering plants with a <u>hose</u>.

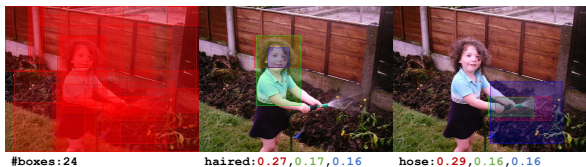#boxes:24   haired:0.27,0.17,0.16   hose:0.29,0.16,0.16

Fig. 5. Visualization of attention. We show all detected regions and top-3 attended regions with attention scores for two words highlighted in each caption. More qualitative results can be found on our project page http://tanmaygupta.info/info-ground/

References

1. Akbari, H., Karaman, S., Bhargava, S., Chen, B., Vondrick, C., Chang, S.F.: Multi-level multimodal common semantic space for image-phrase grounding. CVPR (2018)
2. Alberti, C., Ling, J., Collins, M., Reitter, D.: Fusion of detected objects in text for visual question answering. In: EMNLP (2019)
3. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: ICLR (2017)
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. CVPR (2017)
5. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910 (2019)
6. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: ICML (2018)
7. Chen, K., Gao, J., Nevatia, R.: Knowledge aided consistency for weakly supervised phrase grounding. In: CVPR (2018)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
9. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)
10. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. ArXiv (2020)
11. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. ICCV (2019)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT (2018)
13. Fang, H., Gupta, S., Iandola, F.N., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. CVPR (2014)
14. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: AISTATS (2010)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722 (2019)
16. Hénaff, O.J., Razavi, A., Doersch, C., Eslami, S., Oord, A.v.d.: Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272 (2019)
17. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019)
18. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: ICML (2018)
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML (2015)
20. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)

22. Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. arXiv preprint arXiv:1908.06066 (2019)
23. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
24. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019)
25. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: NeurIPS (1998)
26. McAllester, D., Stratos, K.: Formal limitations on the measurement of mutual information. arXiv preprint arXiv:1811.04251 (2018)
27. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
28. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. arXiv preprint arXiv:1912.01991 (2019)
29. Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. In: NeurIPS (2013)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv (2018)
31. Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., Tucker, G.: On variational bounds of mutual information. In: ICML (2019)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
33. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: ECCV (2016)
34. Song, J., Ermon, S.: Understanding the limitations of variational mutual information estimators. In: ICLR (2020)
35. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations (2020)
36. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Contrastive bidirectional transformer for temporal representation learning. arXiv preprint arXiv:1906.05743 (2019)
37. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
38. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP (2019)
39. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
40. Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M.: On mutual information maximization for representation learning. In: ICLR (2020)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
42. Wang, J., Specia, L.: Phrase localization without paired training examples. ICCV (2019)
43. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
44. Yeh, R.A., Do, M.N., Schwing, A.G.: Unsupervised textual grounding: Linking words to image concepts. In: CVPR (2018)
45. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: AAAI (2020)

A   Appendix

A.1   Limitations and Future Works

The empirical examination of our framework reveals the following limitations:

Pretrained representations. Like prior arts, our approach relies on pre-trained object detector and a language model to represent regions and caption-words. Ideally, we would expect to learn from scratch or improve existing region and word representations directly from image-caption data.

Need for fully-labeled validation set. In Fig. 4, we observe that an early stopping based on the validation performance is required to choose the best model for phrase grounding. While this is common practice for weakly supervised learning [10] and the Flickr30K Entities validation set we use is $80\times$ smaller than the COCO training set, this translates to using full supervision for a small set of images.

Bounds on MI. While $\log(K) - \mathcal{L}_{\texttt{img}}$ in Eq. 8 is a valid lower bound on MI, our $\log(K) - \mathcal{L}_{\texttt{lang}}$ in Eq. 9 is no longer a lower bound on MI as it oversamples negative words related to a caption. A valid bound would involve random sampling of captions from the training data however our context-preserving negative captions lead to much better performance.

A.2   Advantages of Context-Preserving Negative Sampling

Commonly used strategies for negative sampling for contrastive learning include randomly sampling captions from the training data as negatives or mining hard-negatives from a randomly sampled mini-batch. In our experiments (Tab. 2), random sampling showed no significant gains over a model trained without negative captions. This is because the sampled negatives often have an entirely different context as compared to the image and the positive caption which makes it too easy for the model to produce a low compatibility score for these negatives. In contrast, contrast-preserving negative sampling shows significant gains over random sampling (76.74% *vs.* 66.89% pointing accuracy). This is because we construct harder negative captions which yield a more informative training signal than random sampling. We construct negatives by substituting only a single word in the caption while preserving the context from the positive caption. The substitutions are further chosen to be plausible given the context while discarding likely synonyms and hypernyms. Unlike random sampling approaches whose success depends on the occurrence of informative negative captions in the training data and the likelihood of sampling such negatives for a positive caption in the same minibatch, our approach can construct effective negatives for any positive caption.

A.3   Relation between our query-key-value attention and self-attention in Transformers

Our query-key-value attention mechanism is related to the attention mechanism used in transformer-based [41] architectures like BERT [12]. Transformers use the mechanism for self-attention where queries, keys, and values are computed for each word in the input sentence and the attention scores are used for contextualization. In contrast, we use the attention mechanism for word-region alignment. Specifically, we compute queries for each contextualized word, keys for each region, and values for regions as well as words (using separate value networks for regions and words).

A.4   Comparison to Align2Ground

While we use the same visual features as the previous SOTA, Align2Ground [11], the two approaches use different textual features. While Align2Ground uses a bi-GRU, we chose BERT, a transformer-based language model which became more prevalent (as opposed to RNN-based) in the vision-language community. To estimate the gain due to pretrained language representations, Tab. 2 compares the grounding performance of randomly initialized BERT (57.37%) to that of pretrained BERT (66.89%). Negative sampling brings further gains (76.74%).