Learning Unsupervised Visual Grounding Through Semantic Self-Supervision

Syed Ashar Javed[*1], **Shreyas Saxena** and **Vineet Gandhi**[2]
[1]The Robotics Institute, Carnegie Mellon University
[2]CVIT, Kohli Center of Intelligent Systems (KCIS), IIIT Hyderabad
sajaved@andrew.cmu.edu, shreyas.saxena2@gmail.com, vgandhi@iiit.ac.in

Abstract

Localizing natural language phrases in images is a challenging problem that requires joint understanding of both the textual and visual modalities. In the unsupervised setting, lack of supervisory signals exacerbate this difficulty. In this paper, we propose a novel framework for unsupervised visual grounding which uses concept learning as a proxy task to obtain self-supervision. The intuition behind this idea is to encourage the model to localize to regions which can explain some semantic property in the data, in our case, the property being the presence of a concept in a set of images. We present thorough quantitative and qualitative experiments to demonstrate the efficacy of our approach and show a 5.6% improvement over the current state of the art on Visual Genome dataset, a 5.8% improvement on the ReferItGame dataset and comparable to state-of-art performance on the Flickr30k dataset.

1 Introduction

The recent advancements in computer vision have seen the problem of visual localization evolve from using pre-defined object vocabularies, to arbitrary nouns and attributes, to the more general problem of grounding arbitrary length phrases. Utilizing phrases for visual grounding overcomes the limitation of using a restricted set of categories and provides a more detailed description of the region of interest as compared to single-word nouns or attributes. Recent works have used supervised learning for the task of visual grounding (i.e localizing) [Fukui *et al.*, 2016; Plummer *et al.*, 2017; Chen *et al.*, 2017; Rohrbach *et al.*, 2016; Deng *et al.*, 2018]. However, these approaches require expensive bounding box annotations for the phrase, which are difficult to scale since they are a function of scene context and grow exponentially with the number of entities present in the scene. Furthermore, bounding box annotations for phrases are subjective in nature and might contain non-relevant regions with respect to the phrase. This brings us to our main motivation, which is to explore new ways in which models can directly harness unlabelled data and its regularities to learn visual grounding of phrases. Given the lack of supervision, we develop a self-supervised proxy task which can be used for guiding the learning. The general idea behind self-supervision is to design a proxy task which involves explaining some regularity about the input data. Since there are no ground truth annotations, the model is trained with a surrogate loss which tries to optimize for a proxy task, instead of directly optimizing for the final task. A good proxy task improves performance on the final task when the surrogate loss is minimized. In this work we propose concept-learning as a substitute task for visual grounding. During training, we create *concept batches* of size k, consisting of k different phrase-image pairs, all containing a common concept (as illustrated in Figure 1). The proxy task for the model is to decode the common concept present within each concept batch. We induce a parameterization which, given the input text and image, can generate an attention map to localize a region. These localized regions are then used to predict the common concept. Adopting concept-learning as our substitute task, we align our proxy and empirical task, and by introducing concept batches, we constrain the model to learn concept representations across multiple contexts in an unsupervised way.

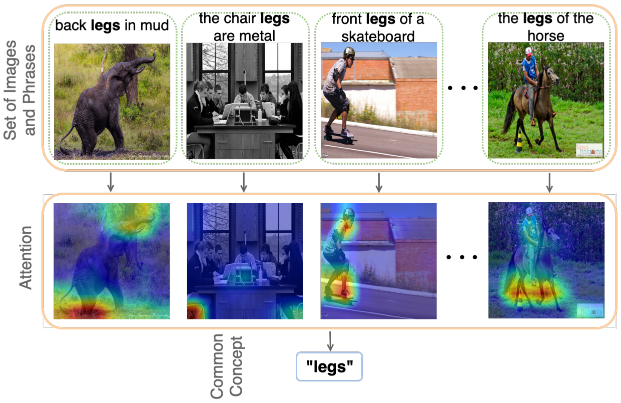
Set of Images and Phrases
back **legs** in mud
the chair **legs** are metal
front **legs** of a skateboard
the **legs** of the horse
Attention
Common Concept
"legs"

Figure 1: We exploit the presence of semantic commonalities within a set of image-phrase pairs to generate supervisory signals. We hypothesize that to predict these commonalities, the model must localize them correctly within each image of the set.

Previous work on unsupervised visual grounding can also

[*] Work done while at IIIT-Hyderabad

be interpreted as having proxy losses to guide the localization. [Rohrbach *et al.*, 2016] use reconstruction of the whole phrase as a substitute task for grounding. However, the objective of reconstructing the entire phrase can also be optimized by learning co-occurrence statistics of words and may not always be a result of attending to the correct bounding box. Moreover, precise reconstruction of certain uninformative parts of the phrase might not necessarily correlate well with the correct grounding. This limitation is also evident in other methods like that of [Xiao *et al.*, 2017] which uses a discriminative loss on the whole phrase instead of generating discrimination for the object to be localized. Many other works like [Ramanishka *et al.*, 2017] and [Zhang *et al.*, 2016] only allow for word-level grounding, thus making them average over the heatmaps to get a phrase-level output. In contrast, our formulation does not suffer from these limitations. Our proxy task deals with the full phrase and forces the model to limit the attention to areas which can explain the concept to be grounded, thus aligning the objective better with the task of visual grounding.

To evaluate the generality of our approach, we test our approach on three diverse datasets. Our ablations and analysis identify certain trends which highlight the benefits of our approach. In summary, the main contributions of our work are as follows:

- We propose a novel framework for visual grounding of phrases through semantic self-supervision where the proxy task is formulated as concept learning. We introduce the idea of a concept batch to aid learning.
- We evaluate our approach on the Visual Genome and ReferIt dataset and achieve state-of-art performance with a gain of 5.6% and 5.8% respectively. We also get performance comparable to the state-of-art on Flickr30k dataset.
- We analyze the behavior of our surrogate loss and the concept batch through thorough ablations which gives an insight into the functioning of our approach. We also analyze the correlation of performance for visual grounding with respect to size of the bounding box and possible bias induced due to the similarity of the grounded concepts to the ImageNet labels.

2 Related Work

The problem of image-text alignment has received much attention in the vision community in the recent years. Early work like DeViSE [Frome *et al.*, 2013] focus on learning semantic visual embeddings which have a high similarity score with single-word labels. Similar to DeViSE, [Ren *et al.*, 2017] learn a multi-modal alignment by constructing a semantic embedding space, but instead of image-label correspondences, they learn region-label correspondences through a multiple-instance learning approach. [Kiros *et al.*, 2014] learn a joint embedding space for a complete sentence and an image using a CNN-LSTM based encoder and a neural language model based decoder. Since the release of the Flickr30k Entities dataset [Plummer *et al.*, 2015] and subsequently the Visual Genome dataset [Krishna *et al.*, 2017], availability of bounding box annotations of phrases has allowed many new attempts at the problem of visual grounding of phrases. [Plummer *et al.*, 2015] provide a baseline for Flickr30k Entities dataset using Canonical Correlation Analysis (CCA) to compute the region-phrase similarity. [Wang *et al.*, 2016] construct a two-branch architecture that enforces a structure and bi-directional ranking constraint to improve upon the CCA baseline. Another recent work from [Chen *et al.*, 2017] departs from the standard usage of bounding box proposals and uses the primary entity of the phrase along with its context to regress localization coordinates. They use a combination of a regression, a classification and a reinforcement learning based loss to train multiple networks in their framework. Prior to our work, there are two papers which take up the problem of unsupervised visual grounding of phrases. [Rohrbach *et al.*, 2016] use reconstruction of the original phrase as a substitute objective function to improve visual attention. But the output predictions in their work is in the form of bounding boxes which, as noted by [Chen *et al.*, 2017], puts an upper bound on the performance. In a more recent work, [Xiao *et al.*, 2017] use the parent-child-sibling structure in the dependency tree of the phrase along with a discriminative loss to generate weak supervision and produce heatmap based outputs for localization. Following [Xiao *et al.*, 2017], we too generate heatmap based localizations, but use an objective which is better aligned with the grounding task. Apart from these papers, certain other unsupervised methods allow modification of their approach to enable evaluation on the phrase grounding task. For example, [Zhang *et al.*, 2016] and [Ramanishka *et al.*, 2017] produce word-level heatmaps and average them to get a phrase-level output. We compare with all these works in section 5.

3 Grounding Through Semantic Self-Supervision

Unsupervised learning can be interpreted as learning an energy function which assigns lower energy value for data points similar to the training set while assigning high energy value to others. In a self-supervised environment, the role of proxy task is to learn the function that pulls down the energy at the data manifold. With this in mind, we define our proxy task for visual grounding.

3.1 Proxy Task Formulation

Our model is trained for the proxy task of concept-learning. A concept is defined as the entity which is to be grounded in the image. For example, in the phrase *'white **towel** on the counter'*, the highlighted word *'towel'* is the concept. We observe that in most phrase-image pairs, the localization refers to some concept which explicitly occurs in the phrase as a single word. We hypothesize that if we induce a parameterization for localization of the phrase and use the localized regions to predict the concept present in an image, the parameterization will converge to the ground truth localization of the phrase. Given this proxy task, we're faced with two main challenges: 1) How do we identify the concept in a phrase? and 2) How do we learn concept representations in an unsupervised setting?

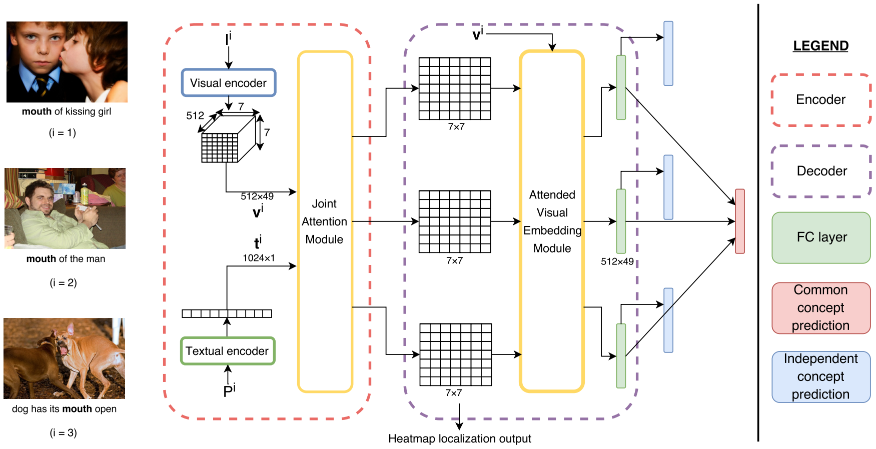

Figure 2: An overview of our model for unsupervised visual grounding of phrases. The encoder takes in a set of image-phrase pairs, indexed by i, all sharing a common concept. The encoder embeds the image and the phrase to $\mathbf{V}^i$ and $\mathbf{t}^i$ respectively. These features are used to induce a parametrization for spatial attention. Next, the decoder uses the visual attention map to predict the common concept. In addition, the decoder also predicts the common concept independently for each pair (i). For details, see Section 3.2.

For the first part, we note that identifying a concept which is to be grounded in a phrase, is a problem from the linguistics domain. We can imagine an external system which takes in as input the phrase and returns the concept. Assuming the concept is a single-word entity and exists within the phrase, a naive system can randomly pick a word from the phrase. Since most concepts to be localized are nouns, a POS tagger performs better than random sampling and in this work, we use it to find all nouns in a phrase and randomly select one of them as the concept.

For the second problem, we introduce the notion of a concept batch and learn the concept-prediction task with such batches. A concept batch, as shown in Figure 2, is one training instance for our model, which itself consists of k phrase-image pairs, all containing a common concept. The proxy task is now re-formulated as jointly decoding the common concept using all k localized feature representations in addition to independently decoding the same concept. The intuition behind training with a concept batch is that for decoding the common concept, k phrase-image pairs should encode a localized representation which is invariant to the difference in context across the k pairs. On the other hand, the proxy task of decoding independent concept (for each image in the batch) ensures two things: a) Individual and common representations are consistent b) Model cannot find a shortcut by using only few inputs from the concept batch to decode the common concept.

It is important to note that using a concept batch for learning along with a noun-based concept can be interpreted as generating weak supervision, albeit noisy in nature. Instead of an imperfect concept-identifier, if an oracle could generate a concept which always corresponded to the actual region to be grounded, then this would convert the unsupervised problem to a weakly supervised one. However in our setting, since the same image-phrase pair can be chosen with different sampled concepts during training, it is this random sampling of concepts which ensures that the model doesn't only learn a simple concept-identifier, but also generates information which can help it discriminate between the same concept in different contexts.

3.2 Encoder-Decoder Model

We adopt an encoder-decoder architecture for learning to ground as illustrated in Figure 2. The encoder uses an attention mechanism similar to [Xu *et al.*, 2015] using the joint features from visual and textual modalities. To maintain fair comparison with previous work, the image features are extracted from the last convolution layer of a VGG16 model [Simonyan and Zisserman, 2014] pre-trained on ImageNet. Similarly, the phrase features are extracted from a language model trained on next word prediction on the Google 1 Billion dataset [Chelba *et al.*, 2013] and the MS COCO captions dataset [Lin *et al.*, 2014]. As done in [Xiao *et al.*, 2017], both the model weights are frozen during training and aren't fine tuned. For the i^{th} index in the concept batch, given visual features from VGG16, $\mathbf{V}^i = f_{VGG}(I^i)$ and textual features from the language model $\mathbf{t}^i = f_{LM}(P^i)$, the attention over visual regions is given by:

$$\mathbf{f}^i_{attn} = softmax(\mathbf{f}_{joint}(\mathbf{V^i}, \mathbf{t^i})). \tag{1}$$

$$\mathbf{f}_{joint}(\mathbf{V}^i, \mathbf{t}^i) = \Phi_s(\Phi_r(\Phi_q(\Phi_p([\mathbf{V}^i, \mathbf{t}^i])))), \tag{2}$$

where $\mathbf{V}^i \in R^{m\times n}$, $\mathbf{t}^i \in R^{l\times 1}$, $\mathbf{f}_{joint}(\mathbf{V}^i, \mathbf{t}^i) \in R^{1\times n}$, $[\mathbf{V}^i, \mathbf{t}^i]$ is an index-wise concatenation operator (over the first dimension) between a matrix $\mathbf{V}^i$ and a vector $\mathbf{t}^i$ resulting in a matrix of size $((m+l)\times n)$. $\Phi(\cdot)$ corresponds to a hidden layer of a neural network and is defined as:

$$\Phi_p(\mathbf{X}) = ReLU(\mathbf{W}_p\mathbf{X} + \mathbf{b}_p), \tag{3}$$

where $ReLU(x) = max(x, 0)$, $\mathbf{W}_p \in R^{p\times d}$, $\mathbf{b}_p \in R^{p\times 1}$ and $\mathbf{X} \in R^{d\times n}$. Here n is the number of regions over which attention is defined and d is the dimensionality of each region with respective to $\mathbf{X}$.

Thus, we use a 4 layered non linear perceptron to calculate attention for each of the n regions [1]. In contrast to [Rohrbach *et al.*, 2016], we compute attention over the spatial regions of the last feature maps from VGG16 instead of computing it over bounding boxes. The four $\Phi(\cdot)$ layers gradually decrease the dimensionality of the concatenated joint features from $(m + l) \rightarrow p \rightarrow q \rightarrow r \rightarrow s$ where $s = 1$. It is important to note that the attention module is shared across all $\mathbf{V}^i$ and $\mathbf{t}^i$. Thus the encoder is common for all pairs in the concept batch. Next, we describe a decoding mechanism to predict the common and independent concept.

Given the attention weights $\mathbf{f}^i_{attn} \in R^{1\times n}$, the visual attention for common concept prediction ($\mathbf{f}_{vac}$) is computed by taking the weighted sum with the original visual features.

$$\mathbf{f}_{vac} = \sum_{i=1}^{k} \mathbf{f}^{i}_{attn}\mathbf{V}^{i} \tag{4}$$

We find that aggregating the visual attention across regions, which is commonly done in the past attention literature degrades performance for our task. Therefore we retain the spatial information and only aggregate the features across the concept batch.

Similarly, the visual attention for independent concept prediction, $\mathbf{f}^i_{vai}$ is given by the element-wise product of the attention weights and visual features.

$$\mathbf{f}^{i}_{vai} = \mathbf{f}^{i}_{attn}\mathbf{V}^{i} \tag{5}$$

Finally, both the attended features are flattened and separately connected to a fully connected layer, leading to a softmax over the concepts. In practice, we also down-sample the dimensionality of $\mathbf{f}^i_{vai}$ using $1 * 1$ convolutions before we aggregate and flatten the features.

$$\mathbf{y}_{common} = softmax(\mathbf{W}_{vac}\mathbf{f}_{vac} + \mathbf{b}_{vac}). \tag{6}$$

$$\mathbf{y}^{i}_{independent} = softmax(\mathbf{W}_{vai}\mathbf{f}^{i}_{vai} + \mathbf{b}_{vai}), \tag{7}$$

where $\mathbf{y}_{common}$ is the network prediction for the common concept and $\mathbf{y}^i_{independent}$ is the independent concept prediction for the i^{th} index in the concept batch.

3.3 Surrogate Loss

Our surrogate loss consists of two different terms, one corresponding to the common concept prediction whereas the other for the independent concept prediction. Since we decode the visually attended features to a softmax over the concept vocabulary, we use the cross-entropy loss to train our model. Given the target common concept vector $\mathbf{y}_t$ for a concept batch of size k, the proxy objective function is defined as:

$$L_{total} = L(\mathbf{y}_{common}, \mathbf{y}_t) + \frac{1}{k}\sum_{i=1}^{k} L(\mathbf{y}^{i}_{independent}, \mathbf{y}_t) \tag{8}$$

where $L(\cdot)$ is the standard cross-entropy loss.

4 Experimental Setup

In this section, we elaborate upon the implementation details, employed datasets, evaluation metric and baselines.

Dataset Statistics	Value		
	Visual Genome	ReferIt	Flickr30k
No of phrases per image	50.0	5.0	8.7
No of objects per image	35.0	-	8.9
Word count per phrase	5.0	3.4	2.3
Noun count per phrase	2.2	1.8	1.2

Table 1: Phrase-region related statistics for datasets used in evaluation. The numbers reflect the relative complexity of these datasets.

Method	Accuracy			
	Visual Genome	ReferIt (Mask)	ReferIt (Bbox)	Flickr30k
Random baseline	11.15	16.48	24.30	27.24
Center baseline	20.55	17.04	30.40	49.20
VGG baseline	18.04	15.64	29.88	35.37
[Fang *et al.*, 2015]	14.03	23.93	33.52	29.03
[Zhang *et al.*, 2016]	19.31	21.94	31.97	42.40
[Ramanishka *et al.*, 2017]	-	-	-	**50.10**
[Xiao *et al.*, 2017]	24.40	-	-	-
Ours	**30.03**	**29.72**	**39.98**	49.10

Table 2: Phrase grounding evaluation on 3 datasets using the pointing game metric. See Section 5 for explanation for ReferIt.

4.1 Implementation Details

A ImageNet pre-trained VGG16 and a Google 1 Billion trained language model are used for encoding the image and the phrase respectively. Both the visual and textual feature extractors are fixed during training. Before the attention module, both the features are normalized using a batch-normalization layer [Ioffe and Szegedy, 2015]. The concept vocabulary used for the softmax based loss is taken from the most frequently occurring nouns. Since the frequency distribution follows the Zipf's Law, around 95% of the phrases are accounted for by the top 2000 concepts, which is used as the softmax size. In the encoder, the values of p, q, r, s from Equation 2 are taken as $512, 128, 32, 1$ respectively.

4.2 Evaluation

Dataset

We test our method on the Visual Genome [Krishna *et al.*, 2017], the ReferItGame [Kazemzadeh *et al.*, 2014] and the Flickr30k Entities [Plummer *et al.*, 2015] datasets and there exist few important qualitative and quantitative differences between them. Table 1 shows some important dataset statistics which hint towards the complexity of the datasets. For example, notice that in Flickr30k, the average phrase length is just 2.3 words and average noun count is 1.2 which would mean that the region to be localized in most cases is directly present as a single word, thus changing the problem to an almost weakly supervised setting. To ensure fair comparison with the previous work of [Xiao *et al.*, 2017], we use the images from the validation set of MS-COCO which have annotations in the Visual Genome dataset as our test set. We use remaining images of the Visual Genome for training. For ReferIt and Flickr30k, we use the test sets for evaluation.

Evaluation Metric

Since our model generates localization in the form of a heatmap, we evaluate our model with the pointing game metric [Zhang *et al.*, 2016], similar to the previous work of [Xiao

[1] Since we compute attention over VGG feature maps, $n = 7 \times 7$

Loss Type	**Concept Batch Size (k)**			
	$k=3$	$k=5$	$k=7$	$k=9$
Independent concept only	27.15	27.27	28.01	28.05
Common concept only	27.52	28.94	29.18	27.90
Independent and common concept	28.25	28.91	29.89	30.03

Table 3: Analysis of different surrogate losses while varying the concept batch size.

et al., 2017; Ramanishka *et al.*, 2017]. Pointing game measures how accurate the most confident region in the predicted heatmap is with respect to the ground truth bounding box. For a given input, the predicted localization heatmap is considered a *Hit* if the pixel with the maximum value in the heatmap lies within the bounding box, else it's considered a *Miss*. The pointing game accuracy is defined as the fraction of correct localizations out of the total testing instances, i.e. $\frac{\#Hit}{\#Hit+\#Miss}$. For an image of size 224×224, the 7×7 attention map is projected back using a stride of $224/7$. Thus each of the 49 grids correspond to a 32×32 region in the original image space and the center point of the highest activated grid is chosen as the maximum value for the pointing game.

Baselines

We compare the performance of our approach with multiple baselines and previous methods. The first is a random baseline which mimics the attention-based localization of our setup, but chooses the region randomly out of the 49 image regions. The second baseline is taken from [Ramanishka *et al.*, 2017; Zhang *et al.*, 2016] where the center point of the image is taken as the max for the pointing game. Note that this baseline can produce skewed results in datasets where the phrase to be localized has a center-bias, which is what we observe with Flickr30k (as previously noted in [Ramanishka *et al.*, 2017]). We also use a visual-only baseline which selects the maximum pixel for the pointing game based on the pre-trained visual features. We use the feature maps from the last convolution layer of an ImageNet-trained VGG16 and average the channel activations to get a 7×7 map. We then choose the maximum activated grid for the pointing game. Apart from these three baselines, we also compare against weakly supervised works of [Fang *et al.*, 2015] and [Zhang *et al.*, 2016] who use an MIL based approach and an excitation backprop scheme respectively for single-word labels. As done in [Zhang *et al.*, 2016; Ramanishka *et al.*, 2017], we average the heatmaps generated for tokens present in their dictionary for obtaining the final heatmap. Finally, we also compare against the more recent unsupervised works of [Xiao *et al.*, 2017; Ramanishka *et al.*, 2017].

5 Results

We report the comparison of our method with the baselines and previous methods in this section. Table 2 summarizes the performance of our best model on the three datasets. To highlight our generalization ability, we train the proposed model on Visual Genome since it's the largest, more complex and diverse dataset out of the three and directly evaluate on the test set of all three datasets without fine tuning.

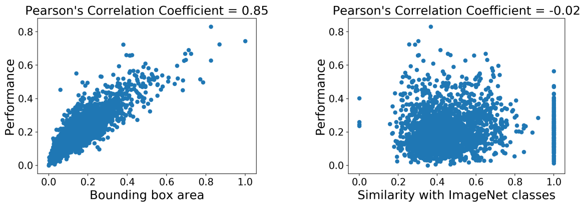

Figure 3: Variation of performance with respect to bounding box area and similarity of concept with ImageNet classes.

Visual Genome

We see that the random baseline yields the least performance as expected. Surprisingly, the VGG16 baseline fares decently well given that it does not take any phrase-related information into account. We believe this is due to the phrases often referring to some object in the image which the VGG16 features are already trained for recognizing. Our model outperforms all the baselines and improves upon the previous state-of-art work by [Xiao *et al.*, 2017] by 5.63%.

ReferItGame

[Hu *et al.*, 2016] provide segmentation mask for each phrase-region pair and use them to obtain a bounding box (bbox) which envelopes the mask completely. They then use this for their evaluation on ReferIt. Though we provide evaluation for both bbox (B) and mask (M) settings, we believe that the mask based annotations are more precise and accurate for measuring localization performance. Since both Visual Genome and ReferIt contain phrases which: a) refer to very specific regions like *'red car on corner'* and b) refer to non-salient objects like *'white crack in the sidewalk'*, both datasets have low performance with baselines like center and VGG. Our model outperforms all baselines on ReferIt too, improving upon the MIL based approach by 5.79%.

Flickr30k Entities

Flickr30k dataset has higher performance across methods as compared to the other two datasets due to the two points mentioned in the previous subsection along with the fact that Flickr30k annotates all bboxes referring to a phrase as opposed to the other datasets which only have a one-to-one phrase-bbox mapping for an image. Our model outperforms most baselines and is just 1% less than the state-of-art work of [Ramanishka *et al.*, 2017].

6 Analysis of the Approach

In this section, we examine the effects of changing the hyperparameter k (concept batch size), the significance of the two surrogate losses and the effect of the concepts with which our model is trained, followed by some qualitative outputs of our model. All the analysis in the following sections is done on the Visual Genome dataset.

6.1 Concept Batch Size and Surrogate Loss

We perform ablative studies on the two loss terms and the concept batch size k and observe certain patterns. For the discussion in this section, we use the shorthand IC (independent

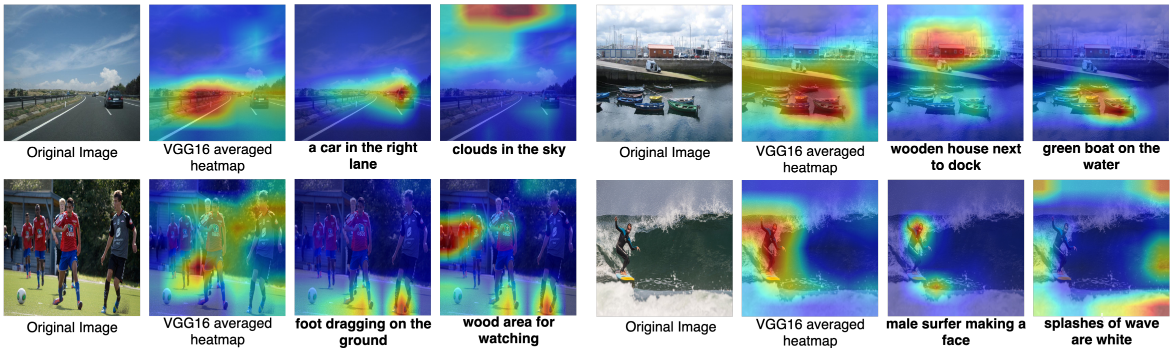

Figure 4: Qualitative results of our approach with different image and phrase pairs as input. More results and visual error analysis shown in the supplementary material on the ***arxiv*** version.

concept only), CC (common concept only) and ICC (independent and common concept) for the three loss types from Table 3. We train our model with the IC and CC loss separately, keeping everything else in the pipeline fixed. For all three settings, we vary the concept batch size k and observe some interesting trends. As shown in Table 3, for a fixed loss type, the performance increases as we increase k, the CC loss being the exception to this trend. The performance for CC loss increases up to $k = 7$, but goes down with $k = 9$. This points to a common problem with self-supervised techniques where the model finds a shortcut to reduce the loss value without improving on the performance. With only the common concept loss, the network can learn a majority voting mechanism such that not all k concept representations need to be consistent with the common concept. Thus, the network can easily optimize the proxy objective, but is not forced to learn a robust grounding for all instances in the concept batch. This is corroborated with the fact that during training, we also observe a faster convergence of CC loss for $k = 9$ than the other values. This empirically highlights the importance of the IC loss term. It also highlights the usefulness of the concept batch formulation since it improves performance in general. For a fixed k, we also observe an expected pattern. IC loss usually achieves the least performance out of the three, with CC loss coming in next. The best performance is obtained with both the losses together.

6.2 Performance Variation Across Concepts

To better understand the variation in performance across the chosen concepts, we also compute the performance across each of the 2000 concept classes. We observe a trend in the performance with concepts like *'suitcase'*, *'airplanes'* and *'breakfast'* getting close to 70% accuracy while concepts like *'screw'*, *'socket'* and *'doorknob'* getting less than 5%. We investigate two possible causes for this variability. The first is the average bounding box size associated with each of these concepts. The second is the existing knowledge of concept labels present in the ImageNet classes which our model obtains through the VGG16 based visual encoder. Figure 3 (left) shows the variation of performance with respect to the average bounding box area for each concept. We observe a strong positive correlation between the two variables, explaining the lower performance for concepts with small sizes. For computing the correlation of concept performance with the knowledge from ImageNet classes, we use a trained word2vec model [Mikolov *et al.*, 2013] and compute the maximum similarity of a particular concept across all the ImageNet classes. We plot this in Figure 3 (right) which illustrates no noticeable correlation between the two variables. This further strengthens the case for our approach since we observe that our concept performance isn't biased towards the labels present in ImageNet.

6.3 Improvement Over a Noun-Based Concept Detector

We also conduct a small experiment to verify that the model isn't simply working as a noun-based concept detector instead of modeling the complete phrase. For this, we replace the full phrase with a single noun, randomly sampled from the phrase, as the input to the textual encoder. We note a 4.7% drop in performance on the Visual Genome dataset for $k = 5$. Since training of the original model enforces only concept-level discrimination, it's interesting to see that the presence of complete phrases is useful for model performance. This shows that our model is much more than a word-level concept-detector and utilizes the full phrase for grounding.

6.4 Qualitative Analysis

We show some of our qualitative results on the Visual Genome dataset in Figure 4 in the form of localization heatmap from the attention weights for two different phrases per image. We also show the VGG16 baseline activation heatmaps. We find that our model does not simply generate a phrase-independent saliency map, but focuses even on non-salient regions if the phrase refers to it.

7 Conclusion

We propose a novel approach for visual grounding of phrases through a self-supervising proxy task formulation. Our qualitative and quantitative results point to the fact that many semantic regularities exist in the data which can be exploited to learn unsupervised representations for a variety of tasks. Thorough analysis of our model reveals interesting insights which may be useful for future research efforts in the area. Using our approach, we achieve state-of-art performance on multiple datasets.

References

[Chelba *et al.*, 2013] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint*, 2013.

[Chen *et al.*, 2017] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *ICCV*, 2017.

[Deng *et al.*, 2018] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018.

[Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint*, 2016.

[Hu *et al.*, 2016] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, 2016.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[Kazemzadeh *et al.*, 2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint*, 2014.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[Plummer *et al.*, 2017] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *CVPR*, 2017.

[Ramanishka *et al.*, 2017] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. Top-down visual saliency guided by captions. In *CVPR*, 2017.

[Ren *et al.*, 2017] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Multiple instance visual-semantic embedding. In *BMVC*, 2017.

[Rohrbach *et al.*, 2016] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014.

[Wang *et al.*, 2016] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[Xiao *et al.*, 2017] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[Zhang *et al.*, 2016] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.