FROM COARSE TO FINE-GRAINED CONCEPT BASED DISCRIMINATION FOR PHRASE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Phrase detection requires methods to identify if a phrase is relevant to an image and localize it if applicable. A key challenge in training more discriminative phrase detection models is sampling negatives. However, sampling techniques from prior work focus primarily on hard, often noisy, negatives disregarding the broader distribution of negative samples. To address this problem, we introduce CFCD-Net, a phrase detector that differentiates between phrases through two novels methods. First, we generate groups that consist of semantically similar words we call concepts (*e.g*. {dog, cat, horse, ...} vs. {car, truck, ...}), and then train our CFCD-Net to discriminate between a region of interest and its unrelated concepts. Second, for phrases containing fine-grained mutually-exclusive words (*e.g*. colors), we force the model into selecting only one applicable phrase for each region using our novel fine grained module (FGM). We evaluate our approach on the Flickr30K Entities and RefCOCO+ datasets, where we improve mAP over the state-of-the-art by 1.5-2 points. When considering only the phrases affected by our fine-grained reasoning module, we improve by 3-4 points on both datasets.

1 INTRODUCTION

Phrase grounding aims to identify image regions related to free form phrases(s). A key difference in the various types of phrase grounding tasks rely on the assumptions (if any) that are made. For example, in phrase localization and referring expression tasks (*e.g*., (Kazemzadeh et al., 2014a; Plummer et al., 2017b)) methods assume the images contain a given phrase and they need only localize it. This results in an evaluation that is analogous to category-specific proposal methods like bounding box regression (Girshick, 2015) or bag of regions (Gu et al., 2009). In contrast, for phrase detection the goal is to locate all relevant regions across a database of images given a phrase, making it analogous to object detection where categories are defined by a natural language phrase. While recent methods have been very successful in the localization-only task (*e.g*. (Kamath et al., 2021; Li et al., 2022)), phrase detection remains a significant challenge. This is due, in part, to the long-tailed distribution of phrases, *i.e*., there may be tens of thousands of annotated queries in a dataset, but many occur only a few times. Thus, learning to identify when a phrase is not relevant to an image is key to good performance. Prior work used hard-negative sampling to boost performance (*e.g*., (Hinami and Satoh, 2018)). However, this generates many false negatives during training, and, as illustrated in Figure 1(a), the larger pool of semantically similar negative phrases may still embed near the ground truth phrase, causing many false positives during inference.

To address this issue, we introduce a Coarse-to-Fine-grained Concept-based Discrimination Network (CFCD-Net), a method that uses semantically coherent groupings of words (concepts) to perform both coarse and fine grained discrimination. A straightforward approach to distinguish between a large set of phrases would be simply using all the phrases in a batch during training. However, this approach has two major drawbacks. First, phrase grounding datasets are very sparsely labeled, so many unannotated phrases can simply be false negatives (see Appendix D). Second, using all phrases in a single batch would not fit into GPU memory as phrase grounding datasets can have tens or hundreds of thousands of unique training phrases (Plummer et al., 2020).

Prior work avoids memory issues by augmenting a batch with a limited number of positive phrases (Plummer et al., 2020) or negative phrases (Hinami and Satoh, 2018) during training. However, as mentioned earlier, this may result in semantically similar phrases being embedded nearby

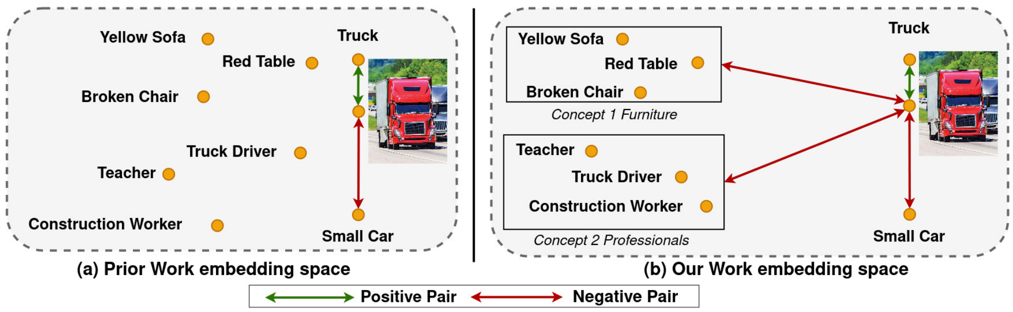


Figure 1: **Comparing embedding spaces**. (a) prior work used hard-negative mining to make their model more discriminative (*e.g.*, (Hinami and Satoh, 2018)), which misses a large distribution of negatives (b), training with our CFCD-Net involves creating a set of semantically coherent words (concepts) (*e.g.* {Sofa, table, Chair}, then, training a model to discriminate a region from its unrelated concepts. The resulting embedding space better separates the region from its unrelated phrases.

the ground truth phrases, reducing performance. Instead, CFCD-Net separates the words belonging to the training phrases into groups with the same part-of-speech. Then we automatically construct a set of concepts, *i.e.*, semantically coherent bags of words, for each group. For the concepts related to nouns (*e.g.*, (dog, cat, horse) and (car, truck, bike)), we perform a coarse discrimination.First, for each phrase we obtain a set of unrelated concepts (*i.e.*, concepts that share no semantically similar nouns with the phrase), which we refer to as negative coarse concepts (NCC). Then, given a region/phrase pair, we use the phrase's NCCs as negatives. As illustrated in Figure 1(b), this encourages separation between the region and a wide array of unrelated phrases.

This approach enables CFCD-Net to minimize false negatives during training, as most unannotated positive phrases belong to the same concept (*e.g.*, the phrase "a cow" would not use the concept "domestic animals" as a negative). In addition, since the set of concepts is small (< 70 in our experiments), we can include all of them in every batch. Our approach is similar in spirit to the distributional sampling approach of Wu et al. (2017), which demonstrated that balanced batches representing the entire dataset boosts performance over hard-negative mining. However, we create concept groups only once in a short preprocessing step (only a few seconds on a CPU). In contrast, Wu et al. (2017) periodically computes pairwise distances between all instances in the training set, which would take hours on a single GPU every time it was recomputed (details in Section 4.1).

While NCCs help provide a more robust representations of nouns in our datasets (which typically refer to objects), another challenge that helps differentiate phrase detection to task like object detection is the need to recognize attributes, which are commonly captured by adjectives in phrases. We find that adjective concepts work in the opposite way as nouns, where within an adjective concept the words refer to difficult fine-grained differences, while phrases can have words in multiple concept bags. For example, let us assume we have concepts for colors (*e.g.*, (red, green)) and some texture patterns (*e.g.*, (striped, plaid)). A shirt could be both red and striped, but a striped shirt should not also be a plaid shirt. We take advantage of these adjective concepts containing mutually exclusive words by adding a fine-grained module (FGM) to our model. Instead of discriminating between the concepts themselves as outlined by our method NCC above, FGM differentiates between the fine-grained words within the concept and then augments the main model score with its prediction. While there are other potentially useful parts-of-speech that may belong to a phrase (*e.g.*, verbs), we find that existing datasets contain too few of them to make a significant impact on performance.

Our contributions can be summarized as:

- A novel model, CFCD-Net, that improves over SOTA by 1.5-2 average mAP on phrase detection by mining semantically coherent concepts and then learning to discriminate between a region of interest and its unrelated negative coarse concepts (NCC).
- A fine grained reasoning module (FGM) that boosts performance by 1-4 average mAP over affected phrases by learning visual cues to differentiate between visually similar instances.
- A novel method for automatically mining semantically coherent concepts that are visually similar and with minimum outliers that improves the distribution of our minibatches and thus better represent the training data.

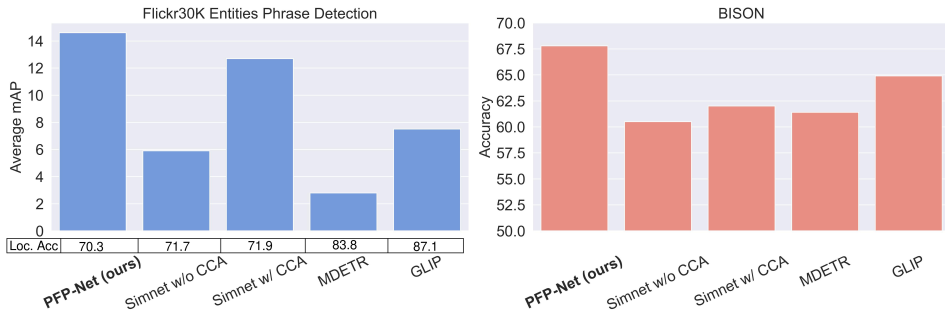


Figure 2: Comparison across methods (Plummer et al., 2020; Kamath et al., 2021; Li et al., 2022) on phrase grounding (left) and Binary Image Selection (BISON) (Hu et al., 2019) (right). Methods along the x-axis are ordered by their localization-only performance. We see that methods designed for localization are not correlated with detection performance (left). However, we see comparing the left and right plots that improvements on phrase detection are correlated with improved performance on downstream tasks like BISON. See Section 2 for discussion

2 Related work and discussion on applications

Most prior work in phrase grounding has focused on the localization-only tasks, where you are provided a ground truth image-phrase pair and have to identify the relevant image region (*e.g.*, (Bajaj et al., 2019; Hu et al., 2016; Chen et al., 2017; Kazemzadeh et al., 2014a; Plummer et al., 2017a; 2018; Yang et al., 2019)). However, Plummer et al. (2020) demonstrated that these methods overfit to the localization task, *i.e.*, they improve performance on localization, but reduce detection performance. This is partly because distinguishing between similar phrases in the localization-only task is unnecessary since most images only contain a reference to a single object of the same type (Plummer et al., 2020). Thus, a localization model that is given the phrase *a young teenager* as input could look for the object category "person" to identify the right object most of the time, whereas a detector also has to determine if it exists at all. As shown in the left side of Figure 2, this means that improved localization performance does not boost performance on detection.

We argue the disconnect between localization-only and phrase detection performance partially explains why improving localization performance only leads to small improvements on downstream tasks (Liu et al., 2017; Datta et al., 2019; Plummer et al., 2017b). Once any grounding method is integrated, even doubling localization performance leads to negligible differences in performance on downstream tasks Plummer et al. (2017b).Thus, most vision-language work uses object and attribute detectors, rather than models trained on localization tasks, to represent images, such as the commonly used bottom-up features of Anderson et al. (2018). In contrast, since phrase detection can also identify what phrases are relevant to an image, it is directly correlated with improved downstream task performance. To help illustrate this point, we compared several grounding methods on the Binary Image Selection (BISON) Hu et al. (2019) benchmark. In this task a model is given an sentence and is asked to to choose between two semantically similar images. In our experiment, we extracted the set of noun phrases from the given sentence, and then ranked images by averaging the grounding scores of the phrases. In Figure 2(right), we see methods that performed well on phrase detection also did well on BISON. Refer to Appendix A for more details.

Some prior work falls between detection and prior work in localization, where a ground truth image-phrase pair is not provided (Hinami and Satoh, 2018; Zhang et al., 2017), but they severely limit the number of negative phrases seen by each image at test time. Thus, methods from these tasks also often do not generalize to phrase detection (Plummer et al., 2020).

Some recent work (*e.g.*, (Kamath et al., 2021; Li et al., 2022)) also tested their models on long tailed object detection benchmarks like LVIS (Gupta et al., 2019a). We note, however, that phrase detection is considerably more challenging, as phrases include information about attributes (*e.g.*, "a **red** shirt") and spatial relationships (*e.g.*, "a cup **on top of** a white table"), in addition to a long tailed set of object categories. Thus, as our experiments show, methods that work well on LVIS (*e.g.*, Li et al. (2022)) do not necessarily generalize to phrase detection.

3 COARSE-TO-FINE GRAINED DISCRIMINATION USING CONCEPTS

The goal of phrase detection is to detect and localize all instances of a phrase within a dataset of images. More formally, assume we have a dataset of images X where for each $x \in X$, regions of interests R_x are annotated with boxes B_x and phrases P_x. Denote the space of possible phrases as $P = \bigcup_{x \in X} P_x$. Thus, for each $p \in P$ and $x \in X$, the task of phrase detection involves determining whether p is relevant to x, and if so localize it with a bounding box. Note that this task involves zero-shot evaluation by definition, since while there are phrases shared between P_{train} and P_{test}, $P_{train} \neq P_{test}$, as well as evaluating few-shot and common phrases. Moreover, as we pointed out earlier, detection is a generalization of localization. More concretely, in localization, for each image x, only $P_x \subset P$ are evaluated, whereas, in detection, the entirety of P is used. This difference is key for the disconnect between improvements on localization and improvements for downstream tasks as we demonstrate in Figure 2 and discuss in Section 2.

We improve the discriminative power of detection models by using Concepts. We define a concept c as a bag of semantically coherent words that share the same part-of-speech (*e.g.*, nouns). We mine for a set of concepts C where each $c \in C$ represents a unique semantic concept (*e.g.*, Vehicles: $\{$car, bike, truck$\}$). Based on these concepts, we introduce two novel methods of discrimination: a coarse method that uses noun based concepts C^n outlined in Section 3.1, and a fine grained methods that uses adjective based Concepts C^a outlined in Section 3.2.

3.1 EXPANDING BATCH COVERAGE THROUGH COARSE NEGATIVE CONCEPTS (NCC)

A critical goal of phrase detection is to ensure that the model can discriminate between a region of interest r and all unrelated phrases in P. Thus, a natural solution to this problem is to expand the batch coverage of negative region-phrase pairs. Following this approach, Hinami and Satoh (2018) mined for hard negative phrase-region pairs to increase their model discriminative power. However, as we discussed in the introduction, these methods miss a large pool of negative phrases that are not "hard" but may still end up embedded reasonably close to the ground truth region embedding. A naive solution would be to consider all possible phrases P. However, detection datasets are sparsely labeled. Thus, it is hard to know if an unannotated phrase is not actually a positive for a region (*e.g.* a region might be labeled with *skier* but not *woman*). Moreover, since phrase grounding datasets contain tens or hundreds of thousands of phrases, we can not fit all of them in one batch due to GPU memory constraints. To mitigate this problem, we propose grouping the nouns in the training dataset into semantically coherent groups, *i.e.* noun based concepts C^n. Then, we pair every phrase in the dataset with its set of unrelated concepts. Consequently, given a positive phrase-region pair (p, r), most of the unannoated positive phrases will be contained in one concept (*e.g. domestic animals*). Thus, we are more confident that all unrelated concepts should be true negatives. Furthermore, since our concepts group only nouns, we are able to limit the concepts to at most 70 per dataset. This makes it feasible to fit all the concepts in one batch. Moreover, since our concepts span the majority of the language space in the training set, our batch now has a balance of easy to hard negatives, which has been shown to improve performance in prior work Wu et al. (2017). We outline our full pipeline of concept generation and phrase-concept assignment below.

Concept Generation and Assignment: To obtain the noun based concept set C^n, we first extract nouns from our dataset and use a language embedding to represent them. Then, we use the embeddings to cluster the nouns into our semantically similar set of concepts. Given the resulting set of concepts C^n, we assign each phrase to its set of relevant concepts C^n_p. A simple solution is to pair a phrase with a concept with which it shares a noun. This approach, however, disregards phrase-concept pairs that don't share nouns but are nevertheless viable pairs. Thus, we make these assignments using semantic similarity as detailed below.

First, given a phrase p, and phrase nouns p_n, we consider concepts related to phrase p which are the result of a simple noun-phrase matching: $C^n_{p1} = \{c : c \cap p_n \neq \emptyset, c \in C^n\}$. We also consider concepts C^n_{p2} which result from a similarity-based assignment process. This method is designed for when a phrase is still relevant to a concept but they do not explicitly share any nouns (*e.g. hoody* with concept: *shirt and sweatshirt*)). More concretely, we look up the language representation for each noun $n \in p_n$. We will denote this representation as $\widetilde{n}$. We also compute a concept c representation by averaging its nouns' textual feature vectors. We will denote this as $\widetilde{c}$. Now assume

$\text{sim}(\widetilde{n},\widetilde{c}) = \frac{\widetilde{n}\,\widetilde{c}}{||\widetilde{n}||\,||\widetilde{c}||}$, we compute an "association" score as follows:

$$\text{Assoc}(n,c) = \frac{\exp(\text{sim}(\widetilde{n},\widetilde{c})/\tau)}{\sum_{j=0}^{M}\exp(\text{sim}(\widetilde{n},\widetilde{c})/\tau)} \quad (1)$$

From here the concepts associated with phrase p based on the semantic similarity are:

$$C_{p2}^{n} = \{c : \text{Assoc}(n,c) > \gamma, n \in p_n, c \in C^n\} \quad (2)$$

This way, the phrase is assigned to concepts with which it shares the greatest semantic similarity. The final set of phrase concepts is therefore, $C_p^n = C_{p1}^n \cup C_{p2}^n$. Thus, a phrase near multiple concepts will be assigned to all of them, and so we will only contrast phrases against groups that we are confident are negatives.

A traditional choice of clustering-algorithm/textual embedding is GLoVE Pennington et al. (2014)/K-Means. However, this combination would result in many noisy concepts (discussed further in Section 4.2). Alternatively, we note that textual embeddings that are visually grounded and designed to make fine-grained distinctions, like ViCO (Gupta et al., 2019b). In addition, density based clustering, like DBSCAN (Deng, 2020), only clusters items that are very close together rather than forcing every item to belong to a cluster like K-means, so we can be more certain they belong to the same concept. As we will show, these choices enabled us to produce a set of concepts with minimal noise. This, in turn, has improved our concept-phrase assignment process in Equation 2. In addition to ViCO, we experimented with transformer based embeddings such as BERT (Su et al., 2019). However, we noticed that there was no semantic similarity in the generated concepts. This is because individual nouns do not have enough context which is critical for transformers to work well. Therefore, we do not report the performance of these embeddings in our experiments section.

After assigning each phrase p to its set of related concepts C_p, we use the phrase's unrelated concepts $\overline{C_p} = C \backslash C_p$ as negative samples. More concretely, we pair $\overline{C_p}$ with the region r associated with phrase p to obtain negative concept-region pairs. These pairs are then simply concatenated to the original ground truth phrase-region positive pairs during training.

3.2 Concept based Fine-Grained Discrimination

In addition to the the mined negative concepts obtained in Section 3.1 which improve the model's robustness against a wide spectrum of nouns (*e.g.*, furniture, vehicles), we also make use of adjective based concepts C^a. However, we note that these concepts behave in a different way than noun based concepts. More concretely, adjective based concepts group together words that refer to fine grained differences (*e.g.*, color, texture patterns). Therefore, we can not discriminate between these concepts because a phrase can have multiple adjectives from multiple concepts (*e.g.*, *red striped shirt*). However, we can confidently discriminate between words within one concept since the concept members are mutually exclusive (*e.g.*, red vs. green). We obtain the set of adjective based concepts C^a using the same approach outlined in Section 3.1, except we use adjectives rather than nouns.

Having obtained the set of adjective concepts, we introduce a novel module, Fine Grained Module (FGM), to discriminate between each concept members. Formally, the FGM module encodes image regions using a set of convolutional layers. It then performs multi-label classification on the members of each adjective based concept (*e.g.*, colors) $c \in C^a$, then uses them to augment the main model region-phrase scores. Formally, given concept $c \in C^a$, $a \in c$, let R be the number of regions/fine-grained adjective pairs, s^a be the region-adjective score, and l^a be its 0/1 label indicating whether it is a positive/negative region-adjective pair, then:

$$L_{FGM} = \sum_{i}^{R} l_i^a \log s_i^a + (1 - l_i^a)\log(1 - s_i^a)). \quad (3)$$

Module Inference with FGM: At test time, the FGM module's scores are augmented with the base model's phrase-region scores. Given a phrase p, the phrase adjectives p_a and $a \in p_a$, a phrase-region pair score s^p, and adjective-region score s^a, then the final score s^f:

$$s^f = (1 - \lambda_c)s^p + \lambda_c s^a, \quad (4)$$

where λ_c is a scalar that applies for every $a \in c$. With this, the final loss for CFCD-net is:

$$L_{final} = L_{base} + L_{FGM}, \quad (5)$$

where L_{base} is the loss used to train the phrase detection backbone which our model is agnostic to.

Table 1: mAp Split by frequency of training instances. **(a)** contains results reported in prior work or produced using their code. **(b)** contains ablations of our model that compares the performance of our its three components (NPA, NCC, and FGM). See Section 4.1 for discussion

		Flickr30K Entities				RefCOCO+			
	#Train Samples	zero-shot	few-shot	common	mean	zero-shot	few-shot	common	mean
(a)	QA R-CNN (Hinami and Satoh, 2018)	3.9	4.3	8.9	5.7	0.9	1.5	9.3	3.9
	Subquery (Yang and et al., 2020)	–	–	–	–	0.7	1.3	9.2	3.7
	FAOG (Wang and Specia, 2019)	3.2	3.5	7.6	4.8	0.7	1.1	8.9	3.6
	MDETR (Kamath et al., 2021)	1.5	2.1	4.8	2.8	1.4	2.6	10.1	4.7
	R-CLIP (Cai et al., 2022)	8.8	6.2	3.9	6.3	4.2	3.5	2.8	3.5
	GLIP (Li et al., 2022)	4.3	6.9	11.3	7.5	2.0	4.5	13.0	6.5
	SimNet (Wang et al., 2018)	4.7	4.4	8.6	5.9	2.0	3.3	13.1	6.1
	CCA (Plummer et al., 2017b)	8.6	10.5	17.2	12.1	5.7	8.4	20.3	11.5
(b)	SimNet w/CCA (Plummer et al., 2020)	9.7	11.2	17.3	12.7	6.0	10.2	20.1	12.1
	+ NCC	10.2	11.9	18.1	13.5	6.1	10.2	21.7	12.7
	+ NPA	10.1	12.0	18.9	13.7	5.9	10.5	22.9	13.1
	+ NPA + NCC	10.6	12.6	19.6	14.1	6.3	10.5	23.4	13.4
	CFCD-Net (ours, NPA + NCC + FGM)	**10.9**	**12.9**	**19.8**	**14.6**	**6.6**	**10.6**	**23.7**	**13.7**
	# Unique Phrases	1783	2764	472	5019	5653	2293	48	7994
	# Instances	1860	4373	8248	14481	5758	3686	1171	10615

Table 2: Performance of the phrases impacted by our FGM module compared with the previous SOTA (SimNet w/CCA). See Section 4.1 for discussion

	Flickr30K Entities				RefCOCO+			
#Train Samples	zero-shot	few-shot	common	mean	zero-shot	few-shot	common	mean
SimNet w/CCA (Plummer et al., 2020)	13.0	14.7	12.0	13.2	7.0	10.3	11.0	9.4
+ NPA + NCC	12.9	13.9	17.1	14.7	7.1	10.6	17.1	11.6
(CFCD-Net) + NPA + NCC + FGM	**15.0**	**17.2**	**18.5**	**16.8**	**7.8**	**11.7**	**17.9**	**12.5**

3.3 Hard Negative Mining

Our proposed method Negative Coarse Concepts (NCC) ensures we cover a large portion of the language space that is uncovered by prior work hard negative mining efforts (*e.g.*, (Hinami and Satoh, 2018)). Nevertheless, discriminating between closely related phrases (*e.g.* car vs. truck) is important for good performance. Therefore, to ensure model's best performance, we use Negative Phrase Augmentation (NPA) (Hinami and Satoh, 2018), a hard negative mining method developed for open-vocabulary localization to train our model. We adapt NPA to the phrase detection task as follows: For each phrase p in the validation set, we record the non-ground truth regions that the model is likely to associate p with. Then, we register the phrases associated with these regions as hard negative candidates for p. We store these candidates in a "Confusion Table" and update it every 3 epochs. Note that for each phrase, we remove candidate phrases that are not mutually exclusive with p using WordNet (Miller, 1995) following (Hinami and Satoh, 2018). This process is critical since a model might rightfully associate the phrase *vehicle* with a region only labeled with *car*. Thus, without checking for mutual exclusivity, NPA will consider *vehicle* a hard negative with respect to *car*. Now, for each positive phrase-region pair (p, r) in the batch, we sample a candidate hard negative phrase $\overline{p}$ and concatenate the negative pair $(\overline{p}, r)$ to the batch.

4 Experiments

Datasets: We evaluate CFCD-net on two common phrase grounding datasets. First, we use Flickr30K Entities (Plummer et al., 2017b) that consists of 276K bounding boxes in 32K images for the noun phrases associated with each image's descriptive captions (5 per image) from the Flickr30K dataset (Young et al., 2014). We use the official splits (Plummer et al., 2017b) that consist of 30K/1K/1K train/test/validation images. Second, we evaluate on RefCOCO+ (Yu et al., 2016),

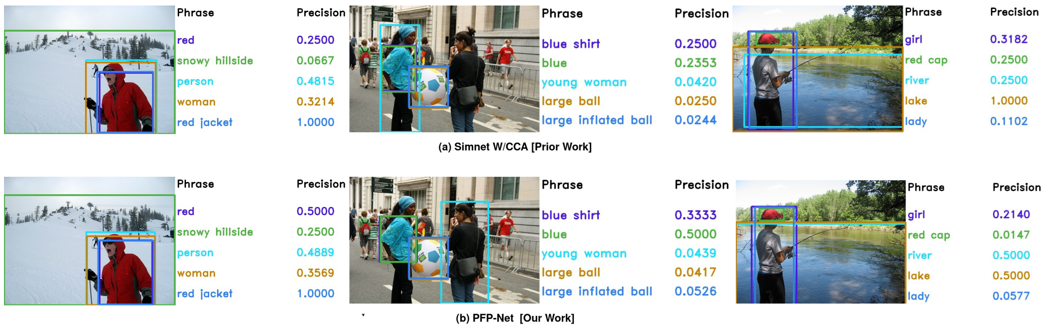


Figure 3: Comparison between CFCD-Net vs prior work on positive (ground truth) precision scores. See Section 4.2 for discussion

Table 3: Comparison between CFCD-Net and transformer based phrase localization models evaluated on Flickr30k. See Section 4.1 for discussion

Model	Training Data	Inference Time
MDETR (Kamath et al., 2021)	COCO, VG, Flickr30k(200k)	7 Days
GLIP (Li et al., 2022)	FourODs,GoldG+, COCO(24M)	1 Day
Our Model	Flickr30k(30k)	$<$ 1 hour

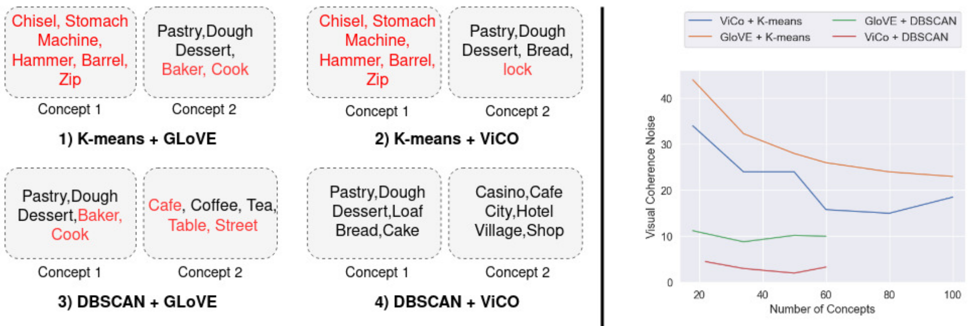
which consists of 19,992 images from the COCO dataset (Lin et al., 2014) that have been labeled with 141,564 region descriptions. We use the official split (Yu et al., 2016), which splits the train/val and testing sets 16K/1.5K/1.5K. Both datasets are licensed under creative commons.

Metrics: We follow the evaluation protocols of Plummer et al. (2020). For every image we obtain the most likely region and confidence score for every phrase in our test split. For any ground truth phrases in an image, we consider them successfully localized if the predicted bounding box has at least 0.5 intersection-over-union with its ground truth bounding box. Then, we compute average precision (AP) for each phrase and then split them into zero-shot, few-shot, and common sets, based on if they didn't occur in our training split, if they had between 1-100 occurrences, or if they occurred more than 100 times, respectively. We then report an overall mAP for each set of phrases, as well as the average of them for an overall performance score. This procedure ensures that the zero-shot and few-shot phrases are not over represented compared to the common phrases, since the zero- and few-shot sets have more unique phrases, but represent a smaller portion of overall instances.

Implementation details. For a fair comparison with state of the art (Plummer et al., 2020), We use their detection backbone (Faster-RCNN) coupled with ADAM (Kingma and Ba, 2017) optimizer and using their hyperparameter settings. For example, we encode images with a ResNet-101 He et al. (2016) and encode phrases by averaging HGLMM Fisher Vectors (Klein et al., 2015). We set all hyperparameters introduced by our work via grid search. Specifically, when creating our concept groups using DBSCAN we set ϵ (the threshold to determine clustering tolerance), and γ, τ in Section 3.1. The used values were $\gamma = 0.2, \tau = 0.01$ for both datasets. $\epsilon = 0.43$ for Flickr30K and $\epsilon = 0.53$ for RefCOCO+. We trained our model using a single NVIDIA RTX 8000 GPU.

4.1 NCC + FGM Results

Table 1 compares our model (CFCD-Net) (averaged over 3 runs) with the state-of-the-art in phrase grounding. Comparing the last line of Table 1(b) to the results from prior work in Table 1(a) and the first line of Table 1(b) we get 1.5-2 point gain in mAP over the state of the art. This gain came mostly from common phrases, where we achieved a 2-3 point gain over prior work. Especially notable is our results using recent transformer-based models MDETR (Kamath et al., 2021), GLIP (Li et al., 2022), and an adaptation of CLIP (Radford et al., 2021) to phrase grounding, R-CLIP (Cai et al., 2022). Note that these methods performed well on phrase localization and challenging object detection benchmarks like LVIS (Gupta et al., 2019a). This is further evidence that simply optimizing for the localization objective as in prior work does not yield discriminative models even as data is scaled up. Furthermore, as reported in Table 3, our model inference time is far shorter. This because both MDETR and GLIP uses cross modal transformers to fuse language/vision data adding significant



(a) Qualitative Comparison between clustering-algorithms/textual-embeddings

(b) Qualitative Comparison between clustering-algorithms/textual-embeddings

Figure 4: Qualitative and Quantitative comparison between difference choices of clustering-algorithm/textual-embedding for concept generation. Refer to Section 3.1 for discussion

Table 4: The table contains ablations of our model that compares the performance of its NCC component using different concept generation methods (*i.e*. clustering algorithm/embedding combinations) to state of the art (SimNet w/CCA). See Section 4.2 for discussion.

	Flickr30K Entities				RefCOCO+			
#Train Samples	zero-shot	few-shot	common	mean	zero-shot	few-shot	common	mean
Baseline [SimNet w/CCA]	9.7	11.2	17.3	12.7	6.0	10.2	20.1	12.1
NCC w/ K-means + GLoVE	10.0	11.8	**18.5**	13.4	6.0	9.9	20.4	12.1
NCC w/ K-means + ViCo	10.1	11.9	18.3	13.4	6.0	9.8	20.7	12.2
NCC w/ DBSCAN + GLoVE	9.9	**11.9**	18.4	13.4	**6.1**	**10.4**	20.7	12.4
NCC w/ DBSCAN + ViCo	**10.2**	**11.9**	18.1	**13.5**	**6.1**	10.2	**21.7**	**12.7**

computational complexity. In contrast, our computationally efficient fusion based on element-wise product followed by lightweight fully connected layers results in an inference time less 1.1% that of either model.

In Table 1(b) we also report the contribution of each component of our model. We observe that our NCC component makes significant improvements over both datasets. Note that although the improvements are quite similar to those made by NPA (Hinami and Satoh, 2018), they come at a fraction of the cost. Specifically, NPA must update a confusion table to select good negatives for every phrase in the training pool during training, which took 3 hours on Flickr and 4 hours on RefCOCO+ using 4 NVIDIA RTX8000 GPUs every time the confusion table was updated (every 3 epochs in our experiments). In contrast, the groups used by NCC are computed only once and takes a few seconds on a CPU using precomputed language features. Nevertheless, as shown in Table 1(b), we find we get best performance when both methods are combined (NPA + NCC).

In addition to the gains from using NCC and NPA, Table 1(b) reports that the FGM module from Section 3.2 further boosts performance gain for both datasets over all phrases. However, since FGM only affects a subset of phrases, in Table 2, we report performance on only the phrases affected by our FGM module, where we obtain a 3-3.5 point gain on Flickr30K Entities and RefCOCO+. Finally, Figure 3 provides additional insight into our model behavior. For each positive phrase, we compute its precision score using the phrase score as a threshold. Thus, higher precision means that there are fewer false positives. Compared to prior work, our model consistently maintains higher precision for each positive phrase. For example, in Figure 3(b) our model has higher precision for phrase ”snowy hillside” in the image on the left than prior work model in Figure 3(a).

4.2 NCC Hyperparameter Analysis

Clustering-Algorithm/Textual-Embedding. One of our main contributions is having the model discriminate between a region and its unrelated concepts (negative coarse concepts (NCC)) as outlined in Section 3.1. These concepts are mined using a clustering algorithm over noun embeddings

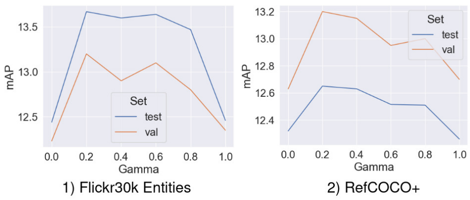


Figure 5: Effect of Concept Size $|C|$

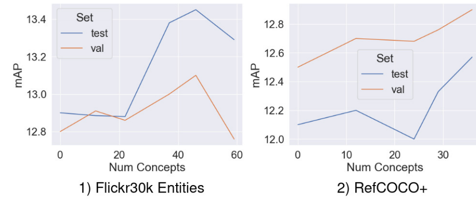


Figure 6: Effect of γ in Equation 2

extracted from the training set. A traditional choice of clustering-algorithm/textual-embedding is K-Means/GLoVE. However, through our experiments, we observed that this combination is ineffective at improving detection performance results in Table 4 demonstrates. To understand this behavior, we first qualitatively examined the resulting concepts in Figure 4 (a)(1). Note that the concept on the left includes just outliers. This is because K-means does not impose a constraint over each cluster density, *i.e.*, outliers in the embedding space are clustered with other concepts even though there is little evidence they belong together. Furthermore, even when the concept is less noisy (concept on the right), it includes words that are semantically similar but not visually similar (*e.g.* Baker vs Dough). This likely harms our concept-phrase assignment process. For example, given a phrase *tall baker*, a textual embedding that is not visually grounded would likely assign a concept containing (”bread, baker, dough”) as related to the phrase. However, a visually grounded embedding would only assign concepts that exclusively contain humans as related. To address these shortcomings, we first use ViCO (Gupta et al., 2019b), a textual embedding that incorporates visual similarity, which should ensure that concepts rule out words that are semantically similar but not visually similar. Furthermore, we use DBSCAN (Deng, 2020), whose clusters are formed from words that fall within a density threshold ϵ. Thus, the method is more effective at ruling out outliers. Note samples of the resulting concepts in Figure 4 (a)(4), which are more visually similar and include fewer outliers. We further document this change by calculating ”visual coherence noise.” We calculated the metric for each clustering-algorithm/textual-embedding by counting the number of resulting concepts where at least %50 of the words were not visually similar. Indeed, ViCO/DBSCAN performs the best, which is then reflected in better performance in Table 4.

Number of Concepts. We also investigate the effect of changing the number of concepts in Figure 5. We progressively increase DBSCAN’s density threshold for each dataset until we can not generate more concepts. We observe that more concepts improve performance on both datasets. This is probably because more concepts split the language distribution into finer portions, thus improving the accuracy of our phrase-concept assignment in Section 3.1.

Concept-Phrase Assignment. Moreover, we examine the effect of changing Section γ Equation 2 parameter in Figure 6 that controls whether a given concept is assigned to a phrase (*i.e.* they are related). We vary the parameter between 0 (a phrase is assigned to every given concept) and 1 (a phrase is only assigned to concepts that share nouns with). We note that performance is best for values of γ between (0.2, 0.6). Performance drop for $\gamma < 0.2$ as we assign wrong concepts to phrases. The performance also drops when $\gamma > 0.8$, which indicates the importance of our similarity-based matching component. In other words, simply assigning phrases to concepts that only share nouns with is not sufficient. This is because, as noted in Section 3.1, many concepts are relevant to certain phrases but do not share words.

5 Conclusion

In this work, we introduced a new phrase detection mode (CFCD-NET) that significantly improves performance by 1.5-2 points on two phrase detection data-sets. It does so by incorporating visually coherent clusters (concepts) to sample negative concept-region that effectively improve the model discriminative abilities when compared to prior work. Our model further improves performance by incorporating a novel fine grained module that learns to discriminate between adjective fine grained tokens. Notably, our approach even outperforms recent transformer-based methods like MDETR (Kamath et al., 2021) and GLIP (Li et al., 2022) on phrase detection performance while requiring less training data and significantly faster inference speeds. In addition, although our experiments used the Faster R-CNN framework to fairly compare to prior work, our contributions are modular and can be adapted to any underlying detection framework.

Code of Ethics Statement Phrase detection methods like ours can help to better understand the content of images, which can be very useful for downstream tasks like image captioning and visual question answering. This can be extremely useful for some tasks like answering questions about images posed by people who have visual impairments, as many of these questions may not be answerable given the image content Gurari et al. (2018). Methods performing well on phrase detection much better than methods developed for other phrase grounding tasks, as the primary difference is exactly to be able to identify if a phrase is present in an image in addition to localizing it. However, these benefits also come with some risks, such as enabling some applications in tasks like surveillance by allowing a user to quickly locate specific entities in a database of images. We also note that although our approach provides significant performance improvements over the state-of-the-art, absolute performance on this task is still quite poor. While this suggests there are ample opportunity for researchers to improve performance on this task, it also indicates that the results of these systems should not be trusted blindly. We also note that Flickr30K Entities and COCO may contain personally identifiable information, but they remain standard benchmarks on phrase grounding tasks, making them an important comparison.

Reproducibility Statement We provide source code with the supplementary material to reproduce the results. Moreover, we provide extensive details for training the model and setting the hyperparameters used in this study in Section 4. Finally, all our reported results as indicated in Section 4 were averaged over 3 runs to ensure statistical significance.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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 (CVPR)*.

Mohit Bajaj, Lanjun Wang, and Leonid Sigal. 2019. G3raphground: Graph-based language grounding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. 2022. X-detr: A versatile architecture for instance-wise vision-language tasks.

Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-guided regression network with context policy for phrase grounding. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 824–832.

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/CVF International Conference on Computer Vision (ICCV)*.

Dingsheng Deng. 2020. Dbscan clustering algorithm based on density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pages 949–953.

Ross Girshick. 2015. Fast R-CNN. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chunhui Gu, Joseph J. Lim, Pablo Arbelaez, and Jitendra Malik. 2009. Recognition using regions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1030–1037.

Agrim Gupta, Piotr Dollár, and Ross Girshick. 2019a. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tanmay Gupta, Alexander Schwing, and Derek Hoiem. 2019b. Vico: Word embeddings from visual co-occurrences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

Ryota Hinami and Shin'ichi Satoh. 2018. Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2615, Brussels, Belgium. Association for Computational Linguistics.

Hexiang Hu, Ishan Misra, and Laurens van der Maaten. 2019. Binary image selection (bison): Interpretable evaluation of visual grounding. *ArXiv*, abs/1901.06595.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr - modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1780–1790.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014a. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014b. Referit game: Referring to objects in photographs of natural scenes. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4176–4182.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bryan A. Plummer, Paige Kordas, M. Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. 2018. Conditional image-text embedding networks. In *The European Conference on Computer Vision (ECCV)*.

Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. 2017a. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017b. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Bryan Allen Plummer, Kevin Shih, Yichen Li, Ke Xu, Svetlana Lazebnik, Stan Sclaroff, and Kate Saenko. 2020. Revisiting image-language networks for open-ended phrase detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv:1908.08530*.

Josiah Wang and Lucia Specia. 2019. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(2):394–407.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. 2017. Sampling matters in deep embedding learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Z. Yang and et al. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *The European Conference on Computer Vision (ECCV)*.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. 2017. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A BISON BENCHMARK

In Section 2 of our paper we used the BISON benchmark Hu et al. (2019) to demonstrate how good performance on phrase detection leads to improved downstream task performance (unlike the localization-only task evaluated in most prior work). In this benchmark, the model is asked to select the caption that best fits an image from a pair of semantically similar images. Formally, given input image I and and a set of noun phrases representing a the caption S, we would use a scoring function (*i.e.*, a phrase grounding method) to measure the relationship between each phrase S_{p_i}, *i.e.*, PhraseScore(S_{p_i}, I). Then, we compute the Image-Caption score ImgScore(S, I) as follows:

	Flickr30K Entities				RefCOCO+			
#Train Samples	zero-shot	few-shot	com-mon	mean	zero-shot	few-shot	com-mon	mean
MDETR Kamath et al. (2021)	1.6	2.6	8.0	4.0	1.4	2.8	11.0	5.0
GLIP Li et al. (2022)	4.3	6.8	13.7	8.3	1.9	4.4	12.6	6.3
CCA Plummer et al. (2020)	8.9	10.7	18.9	12.9	5.7	8.3	20.9	11.7
SimNet w/CCA Plummer et al. (2020)	9.4	11.2	19.8	13.5	6.2	10.3	20.5	12.3
+NCC	10.6	12.1	20.4	14.3	6.2	10.3	21.8	12.8
+NPA	10.3	12.4	20.8	14.5	5.9	10.5	23.6	13.3
+NPA+NCC	10.7	12.7	21.4	15.0	6.4	10.5	23.8	13.5
CFCD-Net (+ NPA + NCC + FGM)	**11.0**	**13.4**	**21.7**	**15.4**	**6.5**	**10.5**	**24.2**	**13.8**

Table 5: mAP split by frequency of training instances where **augmented positive phrases (PPA)** from Plummer et al. (2020) is used for evaluation. The table compares our model three components (NPA, NCC, and FGM) to state of the art (SimNet w/CCA). See Section 4.1 for discussion

Table 6: Phrase Localization accuracy of prior work SimNet w/CCA and our model three components: NPA, NCC, FGM. Section C for discussion.

	Flickr30K Entities	Ref COCO+
SimNet w/CCA	71.9	57.5
+NPA	70.6	56.3
+NPA+CR	70.3	55.9
PFP-Net(+NPA+CR+FGM)	70.3	55.9

$$\text{ImgScore}(S, I) = \frac{1}{|S|} \sum_{i=0}^{|S|} \text{PhraseScore}(S_{p_i}, I) \tag{6}$$

We rank images for a given caption using Eq. (6), then select the image with the bigger score.

B Positive Phrase Augmentation (PPA)

Table 5 reports the performance of CFCD-Net using positive phrase augmentation (PPA) Plummer et al. (2020), which reduces annotation sparsity by pairing ground truth phrases with plausible positive phrases using WordNet Miller (1995). We note that PPA does not change the relative gains of the detection methods, but obtains higher absolute performance.

C Phrase localization performance

As discussed in Section 2, we note that localization and detection performance are not causally related. However, to be complete, we report localization performance of our model in Table 6. We note that our methods' localization numbers are on par with previous work.

D Evaluation dataset selection

The sparsity of phrase detection datasets annotations poses significant challenges on evaluation. Real world datasets annotations can not cover all the possible positive cases. For example, a region annotated with only the phrase *blue shirt* can also be correctly labeled with *clothing*. Plummer et al. (2020) attempted to mitigate this problem by introducing Positive Phrase Augmentations (PPA) where structures like WordNet Miller (1995) were used to derive additional positive samples. However, this problem is not limited to issues with synonyms. Phrases might have different structures

Table 7: **Dataset annotations' false negative rate.** See Section D for more details.

	False Negative%
Referit Kazemzadeh et al. (2014b)	74%
Visual Genome Krishna et al. (2016)	72%
Flickr30k Entities Plummer et al. (2017b)	31%
RefCOCO+ Yu et al. (2016)	40%

but can convey the same meaning (*e.g*. frisbee that is round vs a round frisbee), While the authors of Flickr30K Plummer et al. (2017b) limited the structure of their annotations to mitigate this problem, this is not the case for datasets like Visual Genome Krishna et al. (2016) or Referit Kazemzadeh et al. (2014b) therefore, their validity for phrase detection evaluation is not clear. To quantitatively document this issue, we sampled 30 random phrases from each dataset and considered the top 5 most similar phrases using the visual based language representation ViCo Gupta et al. (2019b). For each of these top 5 phrases, we manually counted the number of false negatives. We report the average results in Table 7. Note that both Referit and Visual Genome suffer from significantly higher false negative rates than Flickr30k Entities. Thus, they are not viable evaluation datasets for phrase detection. We instead use Flickr30k Entities and RefCOCO+ Yu et al. (2016) which was collected using the same underlying game as Referit but with improved data collection standards. Thus the phrases were more appearance focused and concise as evident in the lower false negative rate Table 7.