Doubling down: sparse grounding with an additional, almost-matching caption for detection-oriented multimodal pretraining

Giacomo Nebbia
University of Pittsburgh
Pittsburgh PA, USA
gin2@pitt.edu

Adriana Kovashka
University of Pittsburgh
Pittsburgh PA, USA
kovashka@cs.pitt.edu

Abstract

A common paradigm in deep learning applications for computer vision is self-supervised pretraining followed by supervised fine-tuning on a target task. In the self-supervision step, a model is trained in a supervised fashion, but the source of supervision needs to be implicitly defined by the data. Image-caption alignment is often used as such a source of implicit supervision in multimodal pretraining, and grounding (i.e., matching word tokens with visual tokens) is one way to exploit it. We introduce a strategy to take advantage of an underexplored structure in image-caption datasets: the relationship between captions matched with different images but mentioning the same objects. Given an image-caption pair, we find an additional caption that mentions one of the objects the first caption mentions, and we impose a sparse grounding between the image and the second caption so that only a few word tokens are grounded in the image. Our goal is to learn a better feature representation for the objects mentioned by both captions, encouraging grounding between the additional caption and the image to focus on the common objects only. We report superior grounding performance when comparing our approach with a previously-published pretraining strategy, and we show the benefit of our proposed double-caption grounding on two downstream detection tasks: supervised detection and open-vocabulary detection.

1. Introduction

A common approach for deep learning in computer vision applications is self-supervised pretraining followed by supervised fine-tuning on the task of interest [8, 11, 13, 24, 26]. In self-supervised learning, we train a model in a supervised way, but the source of supervision needs to be implicitly defined by the data. In other words, no human labeler should be involved in retrieving ground-truth values used for self-supervised training. In multimodal datasets where image-caption pairs are available, a widely utilized source of such implicit supervision is the relation between images and their captions. One way to take advantage of this relation is grounding, which refers to the idea of matching a word token (i.e., a word or parts of it) from a caption to a visual token (i.e., a part of an image) from the corresponding image. Specifically, given an image-caption pair, the model needs to learn which visual tokens should be matched with which word tokens, and to force the matched visual and word tokens' representations to be similar.

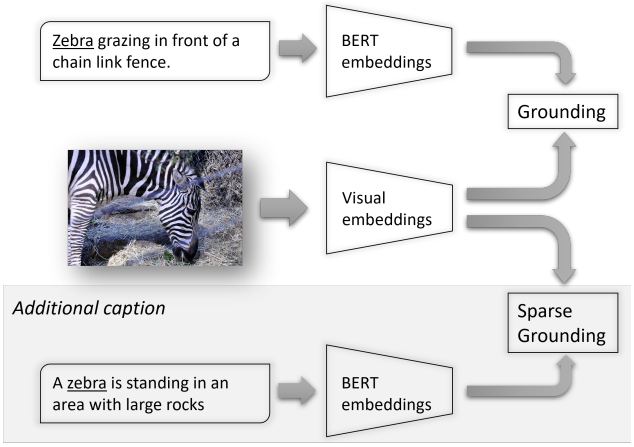


Figure 1. Proposed approach: given an image-caption pair, identify an additional caption that shares mentioned object(s) with the first caption ("zebra" in the example). Grounding is enforced between the matching image-caption, and we propose to enforce a sparse version of grounding between the image and the additional caption. Since the two captions only share mentions of one object, this sparse grounding encourages the model to only ground a few word tokens in the additional caption with the image.

Identifying new implicit relations in multimodal data to use as supervision for self-supervised pretraining and formulating new ways to take advantage of such relations are active areas of research. In addition, recent work [28, 32] has started advocating for tailoring self-supervised pretrain-

ing to the downstream application of interest: devising the pre-training objective for, say, object detection, rather than keeping it suitable for a variety of tasks [20, 30].

We contribute to both research directions by relaxing the binary matching relationship between images and captions. While previous work [6, 15, 18, 20] sees an image-caption pair as either matching or not matching, we propose to consider the relation between two captions that do not refer to the same image (and would thus not be considered matching) but mention one common object (and are, in our approach, "almost-matching"). Because mentioned objects are the link between captions, our pretraining approach is intuitively well suited for object detection.

To take advantage of the relation between captions sharing objects and use it for self-supervised pretraining, we enforce a sparse grounding between an image and the non-matching caption, where only a few word tokens are grounded in the image. The motivation behind this strategy is to encourage the model to learn to only ground the shared objects in the additional caption in the image, leading to better alignment of the visual embeddings with the word embeddings. Since the introduction of the additional caption generates a triplet (I, C, C') with one image (I) and two captions (C and C'), we will in the remainder of the text refer to our method as the *double-caption* approach.

To enforce sparse grounding in our double-caption strategy, we build on the grounding loss as defined by Zareian *et al*. [33] and use their method as our baseline. We evaluate quality of pretraining by using grounding to extract bounding boxes and report superior performance on unsupervised detection (with mAP@0.5 increasing by 26% over the baseline). Furthermore, we qualitatively show how our pretraining better grounds whole objects rather than parts, and we compare the impact of context on our pretrained model and on the baseline. Because the main goal of the proposed approach is to learn better object representations, we also evaluate the benefit of our pretraining strategy on two detection tasks: supervised detection and open-vocabulary detection, and we report improvements in performance over the baseline of 3.8% and 1.7%, respectively.

The remainder of the paper is organized as follows: Section 2 presents related work, Section 3 details our proposed double-caption approach, Section 4 describes our experimental design, Section 5 reports our results, and Section 6 draws our conclusions and outlines potential future research directions.

2. Related Work

Self-supervision has become a standard pretraining step to improve performance on a number of downstream computer vision tasks [8, 11, 13, 24, 26]. The source of self-supervision can come from the images themselves (e.g., solving jigsaw puzzles [24], colorizing images [34], inpainting [25], or predicting image rotations [17]), or can exploit the additional information provided by multimodal datasets like COCO [5, 19] or Conceptual Captions [4, 29]. Spurred by recent success in the application of the Transformer architecture [31] to Natural Language Processing (models like BERT [7] and GPT-3 [3]), many researchers have devised Transformer-based architectures that integrate text and images [6, 15, 18, 20, 30, 33]. While initial approaches relied on image features extracted from pre-existing detectors [20, 30], recent work [15, 33] has started investigating how to overcome such limitation. Our approach also follows this direction and does not require existing object detectors.

To pretrain a model in a self-supervised fashion with multimodal data, previous approaches mainly take advantage of the implicitly defined relations between words, image regions, and between an image and its caption. Each of these relations can be formulated as a loss function, obtaining the Masked Language Modeling (MLM) loss [6, 15, 18, 20, 30], the Masked Region Modeling loss [6, 20, 30], and the Image-Text Matching (ITM) loss [6, 15, 18, 20], respectively.

While still using the MLM and ITM loss functions, additional constraints can be imposed on the visual features extracted from an image. Specifically, Zareian *et al*. [33] use grounding to make features extracted from visual regions similar to the embeddings extracted for the word tokens corresponding to those regions. With our approach, we push this idea further and enforce a sparse grounding formulation between an image and an additional caption that mentions the same objects as the image-matching caption. By adding this constraint, we challenge the binary separation of matching/non-matching image-caption pairs and we introduce a novel intermediate state for "almost-matching" captions.

Our work is also related to the recent push to design self-supervision strategies that are tailored to the downstream task of interest [28, 32, 35]. For object detection, for instance, self-supervised pretraining can be encouraged to learn object characteristics by showing a model different crops of the same image region [28] or differently augmented regions of interest [32]. Our work also aims to teach a model better feature representations for objects, but using multimodal datasets rather than images alone and leveraging the understudied relation between images and almost-matching captions.

3. Method

In this section, we explain how we select the additional caption for each image-caption pair, we summarize the most relevant characteristics of the backbone model that extracts visual and word embeddings, and we show how to formulate the double-caption grounding loss.

3.1. Choosing the third input

Our proposed approach starts with the creation of the (I, C, C') triplets, examples of which are shown in Fig. 2. We see how all additional captions would be traditionally considered non-matching (e.g., there is no parking lot in the top-left and bottom-right images, but C' mentions one in both cases).

C: A stop sign that is hanging upside down.
C': A stop sign is posted next to a parking lot.

C: a close up of a table with many plates of food
C': Some food sitting on top of a table.

C: A half eaten meal sitting on a plate.
C': Plate of food with meats, potatoes, eggs, and fruit.

C: A red bus is driving on the road.
C': A bus parked between two trucks in a parking lot.

Figure 2. Examples of (I, C, C') triplets. The underlined words represent mentions of the shared objects.

Each triplet is created by requiring that (C, C') share one and only one mentioned object. For example, we notice there is no mention of the plates in C' in the top-right image, and there is no mention of the trucks in C in the bottom-right image. In this study, we use ground-truth labels (e.g., person, bicycle, car) to define objects, but the object definition can be extended to include any other label of interest or any noun. We consider the list of synonyms introduced by Lu *et al*. [21] for each ground-truth label since many captions may not mention the labels themselves. Examples of synonyms can be found in Table 1. This list of synonyms was compiled by finding the 200 most similar words for each ground-truth label in the WordVec [23] space and was manually reviewed. Alternative approaches to constructing such a list include using transformer-based embeddings (like BERT [7]) to find similar words, or the use of WordNet synsets [10].

COCO label	Synonyms
person	girl, boy, man, woman, kid
bicycle	bike, unicycle
car	automobile, van, minivan
motorcycle	scooter, motorbike, moped
airplane	plane, aircraft, jet

Table 1. Examples of synonyms for COCO labels [21]

3.2. Base model

The base model is composed of a ResNet-based [12] visual feature extractor and a BERT encoder. During pretraining, the model learns to match visual tokens and word tokens, and forces the visual token embeddings (extracted by the visual feature extractor) to be similar to the matched word token embeddings (extracted by BERT). During fine-tuning, a Faster R-CNN [27] is used for both supervised and open-vocabulary detection, with the weights of the ResNet feature extractor being transferred to the Faster R-CNN backbone and fine-tuned. Classification for open-vocabulary detection is achieved by predicting a class embedding (rather than the class itself) for each region proposal. The predicted embedding is then matched to the BERT embeddings of the considered ground-truth labels.

3.3. Three-input grounding

Let (I, C, C') identify the triplet composed by image I, its corresponding caption C, and the additional caption C'. Grounding between (I, C) is defined as

$$\langle I, C \rangle_G = \frac{1}{n_C} \sum_{j=1}^{n_C} \sum_{i=1}^{n_I} a_{i,j} \langle e_i^I, e_j^C \rangle_L \tag{1}$$

where $\langle ., . \rangle_L$ denotes the dot product of two vectors, e_i^I is the embedding for visual token i with $i = 1, ..., n_I$, e_j^C is the BERT embedding for word token j with $j = 1, ..., n_C$, and

$$a_{i,j} = \frac{\exp \langle e_i^I, e_j^C \rangle_L}{\sum_{i'=1}^{n_I} \exp \langle e_{i'}^I, e_j^C \rangle_L} \tag{2}$$

These $a_{i,j}$ coefficients weigh each dot product between visual and word token embeddings, re-scaling their impact on the grounding loss by the relative importance of each dot product with respect to those between word embedding e_j^C and the other visual embeddings $e_{i'}^I$.

To enforce sparsity of grounding between (I, C'), we modify the grounding definition as

$$\langle I, C' \rangle_G = \frac{1}{n_{C'}} \sum_{j'=1}^{n_{C'}} \beta_{j'} \sum_{i=1}^{n_I} a_{i,j'} \langle e_i^I, e_{j'}^{C'} \rangle_L \tag{3}$$

where β'_j is defined as

$$\beta_{j'} = \frac{\exp \sum_{i=1}^{n_I} \langle e_i^I, e_{j'}^{C'} \rangle}{\sum_{j''=1}^{n_{C'}} \exp \sum_{i=1}^{n_I} \langle e_i^I, e_{j''}^{C'} \rangle} \quad (4)$$

In words, each $\beta_{j'}$ represents the overall attention of word token j' in caption C' when grounded in image I. Defining $\beta_{j'}$ as the softmax of the summation of the attention coefficients for word token j' encourages the distribution of such summations over the word tokens in C' to be sparse; only few word tokens in C' will thus be grounded in image I. In contrast, Equation 1 does not enforce such a constraint, allowing more word tokens in C to be grounded in I.

The loss functions associated with grounding between (I, C) are

$$L_G(C) = -\log \frac{\exp\langle I, C \rangle_G}{\sum_{I' \in B_I} \exp\langle I', C \rangle_G} \quad (5)$$

and

$$L_G(I) = -\log \frac{\exp\langle I, C \rangle_G}{\sum_{C'' \in B_C} \exp\langle I, C'' \rangle_G} \quad (6)$$

where B_I and B_C represent the image and caption batches, and the overall loss $L(I, C)$ is the sum of the two, plus the Masked Language Model loss and the Image-Text Matching loss [15]:

$$L(I, C) = L_G(I) + L_G(C) + L_{MLM} + L_{ITM} \quad (7)$$

The loss for (I, C') is defined analogously, replacing C with C' in $L_G(I)$ and $L_G(C)$ and omitting the MLM and ITM losses. We omit the ITM loss because C' is not I's matching caption, while we omit the MLM loss because this loss will be evaluated on C' when C' is selected as the matching caption of its corresponding image I' rather than as the additional caption in the (I, C, C') triplet. When image-caption pair (I', C') is sampled during training, the ITM and MLM loss will be computed on the pair and our newly-introduced double-caption loss will be computed between image I' and an additional caption C''.

The overall loss that is minimized during training becomes:

$$L(I, C, C') = L(I, C) + \lambda L(I, C') \quad (8)$$

where λ is a hyperparameter that determines the impact of the double-caption loss on training.

4. Experimental Setup

In our experiments, we use the COCO dataset [5, 19]. Although this study focuses on learning meaningful multimodal representations during pretraining, we also show the benefit of our pretraining approach on two downstream detection tasks: supervised detection and open-vocabulary detection.

Because we use the same architecture as Zareian *et al*. [33], we use their method as our baseline as it provides a way to directly assess the impact of introducing our double-caption loss in the pretraining step. Other state of the art methods can be used as baselines, but variations in architecture or pretraining strategy would make for a less effective evaluation of the benefit of our pretraining approach.

4.1. Dataset

For pretraining, we use the COCO Captions dataset [5] (2017 splits). All images and captions used in our experiments come from this dataset. We keep the given train/val splits and we further set aside 5,000 images (same size as the official validation split) from the training split for hyperparameter tuning and model selection. For supervised detection, we use the COCO Objects dataset [19] and we set aside the same 5,000 images for model selection. For open-vocabulary detection, we follow previous work [1] and consider 48 base classes (on which models are trained) and 17 target classes (not seen during training).

4.2. Evaluation Strategies

4.2.1 Grounding-based detection

To evaluate the quality of our pretraining approach, we introduce a way to extract bounding boxes from attention coefficients (Equation 2). These coefficients are defined for a given image-caption pair where the caption can be the actual caption describing the image or any other text we are interested in grounding in the image. In our analysis, we evaluate pretraining using (a) the original caption C, (b) the additional caption C' that shares one mentioned object with C, and (c) an artificially created caption that mentions all ground-truth labels associated with the image (e.g., "A picture of a person, a cat"). In these artificial captions, multiple instances of the same ground-truth label are mentioned once since we would expect the model to ground, say, "person" with all the image regions representing people. The three choices of text to ground in the image allow us to investigate a pretraining strategy's reliance on context: from full context in (a), to less relevant context in (b) where most of the caption does not describe the image (the only exception being the mention of one object), to no additional context in (c) where few words do not refer to objects in the image.

Since each attention coefficient $a_{i,j}$ corresponds to word token j and image region i, we extract bounding boxes by only considering attention coefficients $\geq th_{attn}$ and finding the connected components of the remaining binary map. Each connected component becomes a bounding box, to which we assign a score equal to the average attention coefficient within the just-defined bounding box. In our experiments, $th_{attn} = 0.5$.

We use these bounding boxes and compare them to the ground-truth bounding boxes (i.e., those that will be used

for evaluating supervised detection). We use different subsets of such ground-truth bounding boxes depending on the choice of text to ground with the image. When (a) the original caption is used, we evaluate using all ground-truth boxes as well as using the subset of ground-truth bounding boxes that are associated with objects mentioned by the original captions. In fact, these are the only bounding boxes retrievable through grounding between the images and their captions. When (b) we ground the additional captions C', we only evaluate using the ground-truth bounding boxes linked to the one object shared between (C, C'). This is because such objects are the only ones we can assume are represented in image I since any other object in C' cannot be mentioned by C. Finally, when (c) we consider artificial captions, we use all ground-truth bounding boxes since all objects are mentioned and can thus be grounded in the image.

4.2.2 False Positive analysis

To better understand how the proposed double-caption strategy affects pretraining, we use a previously published diagnosis tool [14] to compare the types of false positive errors committed by our pretrained model with those made by the baseline. This detection diagnosis software classifies incorrect detections into four groups: localization errors, misclassification with similar objects, misclassification with other objects, and confusion with background. These errors represent correctly-classified detections that do not sufficiently overlap with a ground-truth bounding box, detections that classify an object as another considered "similar", detections confusing an object with a non-similar one, and detections that correspond to background regions, respectively. To adapt the diagnosis tool, originally developed for the Pascal VOC dataset [9], to COCO, we define as "similar" objects that belong to the same COCO supercategory.

4.2.3 Sparsity of the additional caption

To verify that our pretraining strategy teaches a model to ground the object mentioned by both captions, we compute the percentage of (I, C, C') triplets where $j^* = \text{argmax}_j(\beta_j)$ (i.e., the word token associated with the highest overall attention) corresponds to the object shared between the two captions. For the baseline, we feed the model each (I, C') pair and compute β_j according to Equation 4.

This evaluation is two-fold as it not only verifies that training with our double-caption approach has the desired effect on the attention coefficients, but also evaluates how our pretraining successfully identifies an object in a caption regardless of its context (i.e., the other words in the caption that do not describe the image appropriately).

4.2.4 Evaluation without bounding boxes

To evaluate pretraining without extracting bounding boxes from attention coefficients, we compare the mean attention coefficients inside (and outside) ground-truth bounding boxes for the baseline and our approach. We also compute the entropy of the coefficients over the visual tokens to have a quantitative measure of the attention coefficients' distribution.

4.2.5 Supervised detection and Open-Vocabulary detection

For downstream detection tasks, we finetune Faster R-CNN [27] models starting from our pretrained models' weights. For supervised detection, we report mean Average Precision (mAP), mean Average Recall (mAR), and mAP at IoU threshold of 0.5 (mAP@0.5). For baseline results, we finetune a Faster R-CNN [27] starting from the pretrained weights made available by Zareian *et al*. [33]. For open-vocabulary detection, we report mAP@0.5, following previous work. Baseline results are taken from Zareian *et al*. [33].

4.3. Implementation

We use the code base provided by Zareian *et al*. [33] to implement our double-caption strategy. Specifically, we adopt the `R_50_C4` configuration from the maskrcnn-benchmark code [22] to extract visual tokens and a frozen BERT model to extract word token embeddings. We use spatial dropout [15] during pretraining to sample visual regions. We choose a learning rate of 0.001, reduced to 0.0001 and to 0.00001 following the scheme in previous work [33]. We set the double-caption loss weight λ to 0.1. We tune hyperparameters through a grid search over the held-out internal validation set, selecting the values that return the best loss after one epoch of training. We use a batch size of 12 and we train on 3 Nvidia Titan X GPUs, which takes about 4 days.

For supervised and open-vocabulary detection, we freeze the first two blocks of the pretrained ResNet50 and finetune the rest. We use a learning rate of 0.01 (decreased by a factor of 10 as before), batch size of 9, and train for 8 epochs on 3 Nvidia Titan X GPUs, which takes about 4 days. For open-vocabulary detection, we use a learning rate of 0.005 (decreased as before), batch size of 8, and train for 10 epochs on 2 Nvidia Titan X GPUs, which takes about 5 days.

5. Results

In this section, we report performance of our double-caption pretraining approach compared to that described in Zareian *et al*. [33], showing a 26% improvement on

mAP@0.5. We also qualitatively show how our approach results in better grounding, where whole objects rather than parts are grounded with words in the captions. We finally evaluate the merit of our pretraining strategy on downstream detection tasks and report superior performance on supervised and open-vocabulary detection.

5.1. Pretraining

5.1.1 Grounding-based detection

Table 2 and Table 3 show performance of pretraining evaluated by extracting bounding boxes from attention coefficients when grounding the original captions with the images (Section 4.2.1). Table 2 evaluates detection using all ground-truth bounding boxes in the evaluation set, while Table 3 only considers those ground-truth bounding boxes associated with a label that is mentioned in the captions. We notice how our approach's mAP@0.5 increases from the baseline's 2.9% to 3.7% (a 27.6% improvement) and from 8.8% to 11.1% (a 26% improvement), respectively.

Method	mAP@0.5	mAP	mAR
Baseline	2.9	1.0	2.8
Ours	3.7	1.3	3.3

Table 2. Pretraining performance (in percentage) considering all ground-truth bounding boxes.

Method	mAP@0.5	mAP	mAR
Baseline	8.8	2.9	8.9
Ours	11.1	3.7	10.0

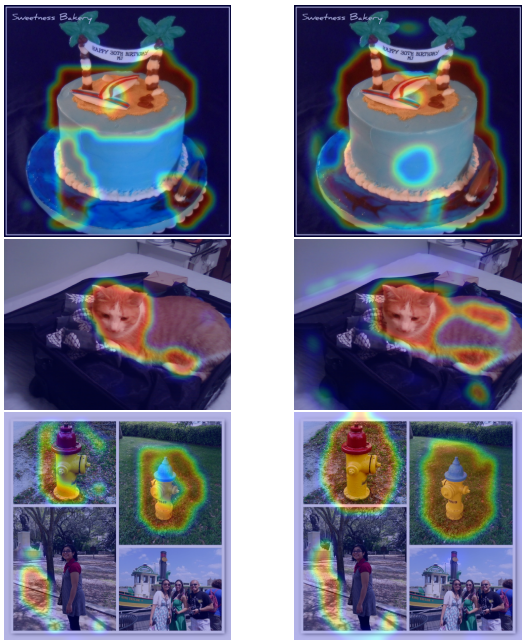
Table 3. Pretraining performance (in percentage) considering mentioned ground-truth bounding boxes.

Figure 3 shows heatmaps for the attention coefficients for the Zareian *et al*. [33] baseline and our double-caption method. We see how the baseline approach tends to ground only parts of an object (e.g., the top of the cake and the head of the cat), while our approach identifies the object in its entirety (as shown by the cake and fire hydrant examples). Intuitively, this explains the increase in performance reached by our method: more accurately localized bounding boxes.

Table 4 shows results for our evaluation of grounding between the images and the additional captions C'.

Method	mAP@0.5	mAP	mAR
Baseline	7.8	2.6	8.9
Ours	8.3	2.9	9.6

Table 4. Pretraining performance (in percentage) when evaluating using (I, C'). We only consider ground-truth bounding boxes associated with the COCO category shared by (C, C').



(a) Zareian *et al*. [33] (b) Double-caption (ours)

Figure 3. Heatmaps of the attention coefficients for (left) the Zareian *et al*. [33] baseline, and (right) our double-caption approach. All images are from the COCO 2017 validation set. The heatmaps represent the attention coefficients associated with the words "cake", "cat", and "fire hydrant".

The performance drop from Table 3 to Table 4 indicates that context provides useful information for both the baseline and our method. When context is less informative (as it is the case when replacing caption C with C'), our approach outperforms the baseline's mAP@0.5 by 6.4%.

To further remove the effect of words that do not describe any COCO object, we also repeat the grounding evaluation using artificial captions that mention each ground-truth object (e.g., "A picture of a person, a cat"). Table 5 shows the results.

Method	mAP@0.5	mAP	mAR
Baseline	3.5	1.1	4.2
Ours	3.8	1.3	4.2

Table 5. Pretraining performance (in percentage) when evaluating using artificial captions mentioning each ground-truth object (e.g., "A picture of a person, a cat").

From Table 5, we notice decreased performance for both methods, similar to what we observe in Table 4, underlining the role of context for grounding-based detection. Even in this case, though, our approach outperforms the baseline's mAP@0.5 by 8.6%.

5.1.2 False positive analysis

To further verify that our double-caption approach provides better localization than the Zareian *et al*. [33] baseline, Figure 4 shows the distribution of the false positive errors across COCO labels classified using the software by Hoiem *et al*. [14]. For each method and for each COCO class, errors are normalized over the total number of false positive errors so they add up to 1. We notice how, on average, our double-caption approach makes fewer localization (loc) mistakes than the baseline, while also making more background (bg) errors. This behavior is exemplified by the images in Figure 3, where objects appear better localized (i.e., higher attention coefficients for whole objects rather than parts), but more background pixels are also grounded with the object (e.g., the third image where more grass is grounded with "fire hydrant").

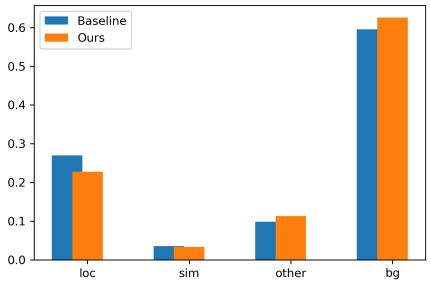


Figure 4. Distribution of false positive error type across COCO labels. Loc: localization error, sim: misclassification with object belonging to the same COCO supercategory, other: misclassification with object belonging to a different COCO supercategory, bg: misclassification as background.

5.1.3 Sparsity of the third input

Table 6 shows the fraction of (I, C, C') triplets in the validation set where the most grounded word in C' corresponds to the object shared by (C, C'). From this table, we show how our training strategy is effective at teaching the model to ground the object (C, C') share in the image, when compared with our baseline (where such behavior is not encouraged during training).

Method	% correct object
Baseline	3.4
Ours	39.2

Table 6. Percentage of (I, C, C') triplets in the validation set where the most grounded (highest $\beta_{j'}$) word in C' corresponds to the object shared by (C, C').

5.1.4 Evaluation without bounding boxes

Table 7 shows the results for our final pretraining evaluation, where no bounding boxes need to be extracted from the attention coefficients.

Method	Inside		Outside	
	Mean	Entropy	Mean	Entropy
Baseline	0.38	3.99	0.046	3.22
Ours	0.44	4.24	0.086	4.25

Table 7. Evaluating pretraining without extracting bounding boxes from attention coefficients: mean and entropy of attention coefficients inside and outside the ground-truth bounding boxes.

From Table 7, we notice higher average attention inside the ground-truth bounding boxes for our method, indicating that our pretraining focuses more on the relevant parts of the images. When evaluating outside the ground-truth bounding boxes, the average attention coefficient is low for both the baseline and our method, with our method having a slightly higher mean attention. This is in agreement with the qualitative analysis in Figure 3, where our method appears to increase coverage of whole objects at the cost of grounding some background areas as well (e.g., the grass around the fire hydrants).

Also in agreement with Figure 3, Table 7 reports higher entropy for our method both inside and outside the ground-truth bounding boxes. In other words, our method grounds more pixels within a ground-truth bounding box, covering whole objects rather than parts.

5.2. Supervised detection

Table 8 reports performance on the supervised detection task, with an improvement over the Zareian *et al*. [33] baseline of 3.2% for mAP@0.5 and of 3.8% for mAP.

Method	mAP@0.5	mAP	mAR
Baseline	43.0	23.6	37.2
Ours	44.4	24.5	38.1

Table 8. Supervised detection performance (in percentage). Our double-caption approach improves mAP by 3.8% (from 23.6% to 24.5%).

Learning a meaningful and useful multimodal representation is this study's objective. With the above analysis, we show the merit of our double-caption pretraining approach, and we leave efforts to tune our experimental design to establish a new SOTA to future work. The comparison with the approach in Zareian *et al*. [33] remains the most appropriate since it allows us to directly compare the impact of introducing our double-caption loss, while variations in model architecture in other SOTA methods would make such a direct comparison infeasible.

5.3. Open-vocabulary detection

Table 9 shows results on the open-vocabulary detection task, where we improve overall generalized mAP@0.5 by 1.7%.

Method	Base	Target	Generalized		
			Base	Target	All
Baseline	46.8	27.5	46.0	22.8	39.9
Ours	49.9	22.6	49.4	15.8	40.6

Table 9. Open-vocabulary detection mAP@0.5 (in percentage). Our double-caption approach improves mAP@0.5 by 1.7% (from 39.9% to 40.6%).

From this table, we notice how the superior overall performance is due to increased performance on the base classes on which the model is fine-tuned, which is in agreement with our findings on fully-supervised detection. Nonetheless, this improvement on the base classes comes at the cost of reduced performance on the target classes, showing how task-specific fine-tuning can limit the benefit of improving pretraining across all classes (see Tables 2,3). Lowering the learning rate ameliorates this problem, but at the same time significantly reduces performance on the base classes.

6. Conclusions and Future Work

In this study, we introduce a double-caption loss to take advantage of a previously understudied structural relation in unlabeled image-caption data: captions for different images may mention the same object(s). Given an image-caption pair (I, C), we use this relation as a source of supervision by enforcing a sparse grounding between image I and a caption C' that mentions one object that C mentions as well. We call grounding between (I, C') "sparse" because only the word tokens associated with the shared object should be grounded.

We use the work in Zareian *et al*. [33] as our baseline, and we evaluate quality of pretraining quantitatively and qualitatively, showing grounding-based detection performance of mAP@0.5 of 11.1% (vs. the baseline's mAP@0.5 of 8.8%) and highlighting how our pretraining approach appears to better capture whole objects rather than parts (Figure 3). In addition, our pretraining strategy trains models that, although sensitive to context, keep ouperforming the baseline even if context is removed (Tables 4, 5). Finally, we evaluate our pretraining strategy by fine-tuning on two detection tasks: supervised and open-vocabulary detection, reporting an improvement in mAP@0.5 of 3.2% and of 1.7%, respectively.

Given the promise of our approach, many research directions can be pursued in the future. First, additional downstream tasks can be evaluated like weakly-supervised object detection [2] or phrase grounding [16]. Second, we could increase the number of objects captions (C, C') share (although the higher the number of shared object, the less sparse grounding between (I, C') would have to become). Finally, the image originally associated with C' could be used as the third input.

Societal impact: Because the proposed method aims to better learn information from image-caption pairs, our models could include biases coming from the data. Racial and gender biases in the dataset may limit the generalizability of our models to underrepresented races/genders. Additionally, captioning itself can be biased by the human annotators' beliefs or characteristics, potentially limiting the generalizability of the learned models to captions obtained from different annotators.

Acknowledgements: This work was supported by a University of Pittsburgh Intelligent Systems Program fellowship, and gifts from Amazon and Adobe.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 4

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 8

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 4

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3

[8] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 1, 2

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[10] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[13] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 1, 2

[14] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 5, 7

[15] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2, 4, 5

[16] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 8

[17] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, Apr. 2018. 2

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[21] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 3

[22] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: March 2nd, 2022. 5

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3

[24] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1, 2

[25] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[26] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021. 1, 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 5

[28] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021. 1, 2

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[30] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[32] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[33] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 14393–14402, 2021. 2, 4, 5, 6, 7, 8

[34] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[35] Yuanyi Zhong, Jianfeng Wang, Lijuan Wang, Jian Peng, Yu-Xiong Wang, and Lei Zhang. Dap: Detection-aware pre-training with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2021. 2