Weakly-Supervised Visual-Textual Grounding with Semantic Prior Refinement

Davide Rigoni[1,2]
davide.rigoni.2@phd.unipd.it
Luca Parolari[1]
luca.parolari@unipd.it
Luciano Serafini[2]
serafini@fbk.eu
Alessandro Sperduti[1,3]
alessandro.sperduti@unipd.it
Lamberto Ballan[1]
lamberto.ballan@unipd.it

[1] University of Padova
Padova, Italy
[2] Fondazione Bruno Kessler (FBK)
Povo, Italy
[3] University of Trento
Trento, Italy

Abstract

Using only image-sentence pairs, weakly-supervised visual-textual grounding aims to learn region-phrase correspondences of the respective entity mentions. Compared to the supervised approach, learning is more difficult since bounding boxes and textual phrases correspondences are unavailable. In light of this, we propose the Semantic Prior Refinement Model (SPRM), whose predictions are obtained by combining the output of two main modules. The first untrained module aims to return a rough alignment between textual phrases and bounding boxes. The second trained module is composed of two sub-components that refine the rough alignment to improve the accuracy of the final phrase-bounding box alignments. The model is trained to maximize the multimodal similarity between an image and a sentence, while minimizing the multimodal similarity of the same sentence and a new unrelated image, carefully selected to help the most during training. Our approach shows state-of-the-art results on two popular datasets, Flickr30k Entities and ReferIt, shining especially on ReferIt with a 9.6% absolute improvement. Moreover, thanks to the untrained component, it reaches competitive performances just using a small fraction of training examples.

1 Introduction

Visual-textual Grounding (VG), i.e. the task of locating objects referred by natural language sentences, requires a joint understanding of both visual and textual modalities. Depending on the amount of annotations used during training, VG can be tackled in a different manner. In this work, we focus on the weakly-supervised setting [3, 11, 35, 37, 40] in which the only available annotation refers to image-sentence pairs. In other words, it is only known which sentence describes each image in the dataset, but not the objects in the image referred by the textual phrases composing the sentence. In contrast, in a fully-supervised setting the

model is trained using all the region-phrase pairs [4, 11, 16, 29, 31, 39], which in practice is a difficult and very expensive annotation to collect.

To this end, we propose a simple model referred as to Semantic Prior Refinement Model (SPRM), whose predictions are obtained by combining two modules: (i) the first, which **does not require training**, for each textual phrase returns a rough alignment with a candidate object bounding box, (ii) while the second, composed by two **trained** sub-components, refines the rough alignments in the final phrase-bounding box alignments. Given a textual phrase and an image as input, the model recognizes the most relevant objects in the image using a pre-trained object detector, and predicts the bounding box referred by the phrase adopting the two aforementioned modules. Specifically, the rough alignment is based on the similarity score (i.e. concept similarity) between the head of the textual phrase and the predicted label of the bounding boxes. Here, the key idea is that the head of the phrase should be very similar (semantically speaking) to the content of the bounding box and, thus, to its class.

The model is trained to maximize the multimodal similarity between an image and a sentence describing that image, while minimizing the multimodal similarity of the same sentence and a new unrelated image, adequately selected. We investigated the model performances on the Flickr30k Entities and the ReferIt datasets, showing that our model presents consistent and competitive results in both datasets. Moreover, we evaluated our model performance in low-data environments, showing that our model can still achieve surprising results even when trained with just a tiny fraction of training examples

Our contributions can be summarized as follows: (i) we propose a new model which is based on the novel idea of first predicting a rough alignment between the phrase and a bounding box, and then refining the prediction; (ii) we conduct extensive experiments on the popular Flickr30k Entities and ReferIt datasets, showing state-of-the-art results (in the weakly-supervised setting); (iii) our model, even when trained on a small fraction of the available examples (e.g. 10%), achieves consistently competitive results.

2 Related Works

In the literature, there are many works related to our proposal. Attention was successfully used to generate spacial attention masks able to localize regions referred by phrases preserving linguistic constraints [38], using self-supervision [14], or even by leveraging a multimodal semantic space [1]. Attention was also employed to reconstruct a subset of the input like query subject, location, and context [21], or the full input along with the proposal's information [9]. Basically, the idea is to reconstruct the input from selected relevant features such that the model learns to ground entities mentioned in the text. Along with traditional grounding systems, a regression loss is implemented to refine the bounding boxes coordinates [23]. Spacial transformer [13] was successfully employed to compute correlation scores between phrase and image's spacial features map [40].

Following the encoder-decoder architecture, a slightly different approach consists of learning to ground entity-region by randomly blending arbitrary image pairs, which are reconstructed conditioned by the corresponding texts [3]. Leveraging the idea of a similarity measure between the two modalities, other works developed a contrastive learning framework where the model localizes entity-region by image-sentence supervision: the contrastive examples may be guided by replacing words in sentences [11], or either distilling knowledge in order to compute accurate similarity scores [36]. Using the bounding box's labels and attributes in the representation of image features allows to compute meaningful embedding

that can be compared with the language modality [37]. Eventually, object detectors may be combined to extract redundant and robust information about regions [35].

A different approach instead overcomes weak supervision by learning to ground through caption-to-image retrieval task: learning caption-to-image retrieval intrinsically means learning to ground, i.e., the model learns to return the correct image given a caption if and only if it has properly understood how to ground the caption with the image [5].

3 Problem Definition

Formally, given an image $\boldsymbol{I} \in \mathcal{I}$ and a sentence $\mathrm{S} \in \mathcal{S}$ the VG task aims to learn a map $\gamma : \mathcal{I} \times \mathcal{S} \rightarrow 2^{\mathcal{Q}_\mathrm{S} \times \mathcal{B}_{\boldsymbol{I}}}$, where $\mathcal{Q}_\mathrm{S}$ is the domain of the noun phrases defined on S, and $\mathcal{B}_{\boldsymbol{I}}$ is the domain of all the bounding boxes defined on $\boldsymbol{I}$. More precisely, the set $\mathcal{Q}_\mathrm{S}$ is defined as $\{\boldsymbol{q}_j\}_{j=1}^{m}$, where m is the number of noun phrases and $\boldsymbol{q}_j \in \mathbb{N}^2$ is a vector containing the initial and final character positions in the sentence S.

In this work, given an image $\boldsymbol{I}$, we deploy a pre-trained object detector to extract the set of bounding box proposals $\mathcal{P}_{\boldsymbol{I}} = \{(\boldsymbol{c}_k, \boldsymbol{h}_k, \boldsymbol{l}_k)\}_{k=1}^{p} \subset \mathcal{B}_{\boldsymbol{I}}$, where $\boldsymbol{c}_k \in \mathbb{R}^4$ represents bounding box coordinates, $\boldsymbol{h}_k \in \mathbb{R}^v$ is the v-dimensional vector representing the bounding box features, and $\boldsymbol{l}_k \in \Theta$ denotes the class with the highest probability to represent the content of the bounding box over the object detector pre-defined set of categories Θ. The bounding box proposal's classification is a common feature offered by most object detectors and will be used in Section 4.1 to define the concept similarity.

In the weakly-supervised approach, a training set of n examples is defined as $\mathcal{D} = \{(\boldsymbol{I}_i, \mathrm{S}_i)\}_{i=1}^{n}$. In other words, only the information about sentence S_i describing the image $\boldsymbol{I}_i$ is available at training time, while it is unknown which region $\boldsymbol{b} \in \mathcal{B}_{\boldsymbol{I}}$ is described by a noun phrase $\boldsymbol{q} \in \mathcal{Q}_\mathrm{S}$. Hence, we learn $\gamma(\boldsymbol{I}, S)$ such that it returns a subset $\Gamma \subseteq \mathcal{Q}_\mathrm{S} \times \mathcal{P}_{\boldsymbol{I}}$ where each couple $(\boldsymbol{q}, \boldsymbol{p}) \in \Gamma$ aligns the noun phrase $\boldsymbol{q}$ to the bounding box proposal $\boldsymbol{p}$.

4 Our Method

Figure 1 depicts our Semantic Prior Refinement Model (SPRM) architecture, which is composed mainly of two modules. One is the *Concept Branch* (CB) (see Section 4.1), responsible for predicting a first rough set of region-phrase correspondences. Those alignments are obtained through a process named "concept similarity" that captures the semantic information conveyed by prior knowledge in object detector and word embedding. In particular, it compares the word embeddings of the phrase's head and the bounding box class to get unimodal scores. No training is required. The information is matched by relying on two important assumptions: (i) the proposal's label semantically describes the bounding box content, (ii) and the word embedding space represents the semantic similarity of the words. Moreover, the CB includes a positional heuristic that helps to reduce ambiguity for candidate alignments.

The other module (see Section 4.2) is made by two sub-components, namely *Visual Branch* and *Textual Branch*, and it is trained to learn a multimodal embedding space for region-phrase correspondences given image-sentence pairs. The multimodal representations are constructed to maximize the similarity of region-phrase pairs when both come from the same example, while minimizing the similarity between the regions from the positive example and phrases from another example. The second refined set of alignments is obtained by measuring the similarity between learned multimodal visual and textual features for the

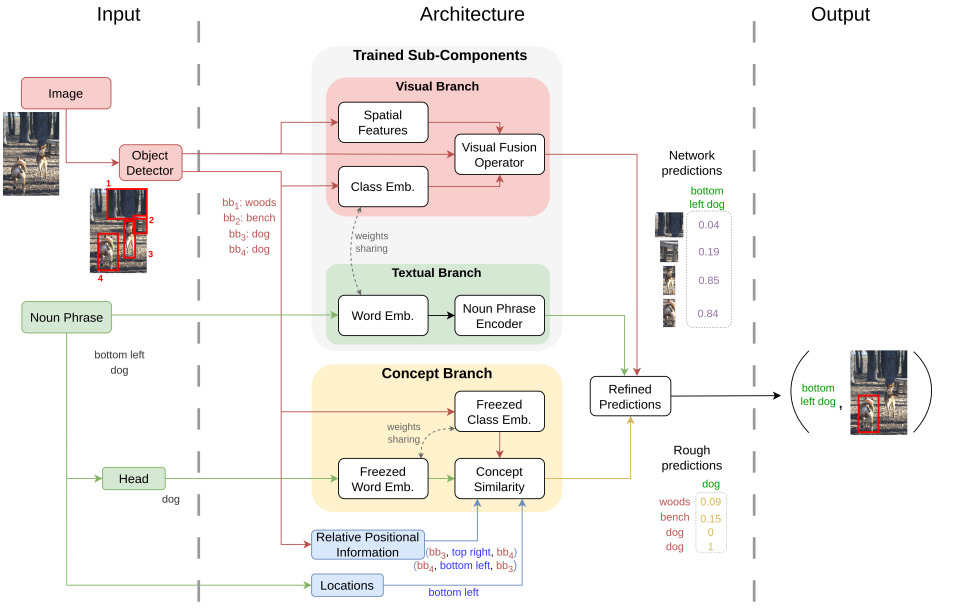

Figure 1: **Our model architecture overview.** The model computes a first rough set of alignments by leveraging prior knowledge from the object detector and word embedding (i.e. *Concept Branch*). A simple positional heuristic is injected as an extra source of prior knowledge to reduce ambiguity for candidate alignments. Then, the visual and textual branches (i.e. *Trained Sub-Components*) match learned multimodal features to predict a second, refined set of alignments. The two sets are then combined together by the *Refined Predictions* module to compute final scores for grounding.

bounding box proposal and noun phrase. The resulting scores are then combined by the prediction refinement module (see Section 4.3) to produce final scores. The candidate alignment is chosen to be the proposal with maximum similarity with the noun phrase.

4.1 Concept Branch

The *Concept Branch* (CB) is designed to face the most important problem in the weakly-supervised VG: the unavailability of region-phrase ground truths. We make use of external sources of knowledge to fill this gap. The CB leverages a pre-trained object detector to abstract the content of an image's region through the bounding box classification label, that is the concept expressing the content of the region. The bounding box classification label is a common feature in most object detectors allowing them to express the content of the bounding box as a concept in the language domain. To understand the concept expressed by a textual phrase, we use an off-the-shelf NLP parser deployed to extract the head of the phrase [14]. In fact, the head of a textual phrase determines its syntactic category. Then, by means of a pre-trained word embedding that conveys prior knowledge of words, the CB computes the similarity between the two concepts to obtain a rough score named "concept similarity". In this process, there is **no training** involved; thus the process is entirely independent of training data and can be treated as prior knowledge.

This method, although general enough to cover a vast set of cases, suffers from some limitations. First, the proposal's classification may be noisy and incorrect, driving the CB to inaccurate alignments. Second, the word embedding similarity may be biased and imprecisely captures the semantic similarity between words. Third, the CB produces equal scores when proposals have the same label. In order to deal with this issue, we adopted another source of prior knowledge based on spatial relations. For proposals with the same label, we extract relative positional information (e.g. "top"). We then match those relations with a location extracted from the phrase by a simple text search (e.g. "left" in "dog on the left").

Formally, given a set of p bounding box proposals $\mathcal{P}_{\boldsymbol{I}}$, let $E^{\mathcal{P}_{\boldsymbol{I}}} = \{\boldsymbol{e}_k^{\mathcal{P}_{\boldsymbol{I}}}\}_{k=1}^{p}$ be the corresponding set of g-dimensional vectorial embeddings, where each $\boldsymbol{e}_k^{\mathcal{P}_{\boldsymbol{I}}}$ is the embedding of the bounding box class $\boldsymbol{l}_k$, for $1 \leq k \leq p$. Given a noun phrase $\boldsymbol{q}_j$ composed by a sequence of $L(\boldsymbol{q}_j)$ words $W^{\boldsymbol{q}_j} = [w_1^{\boldsymbol{q}_j} \dots w_{L(\boldsymbol{q}_j)}^{\boldsymbol{q}_j}]$, let $E^{\boldsymbol{q}_j} = \{\boldsymbol{e}_i^{\boldsymbol{q}_j}\}_{i=1}^{L(\boldsymbol{q}_j)}$ be the set of words embedding of size g associated with each word in the noun phrase $\boldsymbol{q}_j$. Let $\boldsymbol{s}_j^t \in \mathbb{R}^6$ and $\boldsymbol{s}_k^v \in \mathbb{R}^6$ be two multi-hot vectors that encode locations in $\boldsymbol{q}_j$ and relations in the k-th proposal, respectively. Then, the concept similarity score for each proposal is:

$$\boldsymbol{S}_{jk} = f_{mask}\left(\boldsymbol{\xi}_j, \boldsymbol{e}_k^{\mathcal{P}_{\boldsymbol{I}}}, \boldsymbol{s}_j^t, \boldsymbol{s}_k^v\right) = \begin{cases} f_{\text{sim}}\left(\boldsymbol{\xi}_j, \boldsymbol{e}_k^{\mathcal{P}_{\boldsymbol{I}}}\right) & \text{if } \left(\boldsymbol{s}_k^v\right)^\top \boldsymbol{s}_j^t \geq 0; \\ -1 & \text{otherwise.} \end{cases}$$

where f_{sim} is a similarity measure (e.g. cosine similarity) and $f_{\text{sim}}(\boldsymbol{\xi}_j, \boldsymbol{e}_k^{\mathcal{P}_{\boldsymbol{I}}})$ returns the similarity score between word embeddings of phrase's head $\boldsymbol{\xi}_j$ and proposal's label $\boldsymbol{e}_k^{\mathcal{P}_{\boldsymbol{I}}}$. The function f_{mask} returns this similarity score only when the phrase and the proposal share at least one spatial reference, otherwise -1 is returned.

4.2 Visual and Textual Branches

Given the set of bounding box proposals $\mathcal{P}_{\boldsymbol{I}}$ detected in the image $\boldsymbol{I}$ by the object detector, for each of them, our model extracts the spatial features $H^s = \{\boldsymbol{h}_k^s\}_{k=1}^{p}$ where $\boldsymbol{h}_k^s \in \mathbb{R}^5$, as indicated in [29]. Moreover, contrary to the *Concept Branch*, the Visual and Textual branches adopt trainable word embeddings $\overline{E}^{\mathcal{P}_{\boldsymbol{I}}} = \{\overline{\boldsymbol{e}}_k^{\mathcal{P}_{\boldsymbol{I}}}\}_{k=1}^{p}$ and $\overline{E}^{\boldsymbol{q}_j} = \{\overline{\boldsymbol{e}}_i^{\boldsymbol{q}_j}\}_{i=1}^{L(\boldsymbol{q}_j)}$ associated to the bounding box classes and to the words of the noun phrases, respectively.

Initially, both visual and spatial features are concatenated and then projected on a smaller dimensional space, thus leading to a set of new vectorial representations $H^{||} = \{\boldsymbol{h}_k^{||}\}_{k=1}^{p}$, with $\boldsymbol{h}_k^{||} = \boldsymbol{W}^{||}\left(\boldsymbol{h}_k^s || \boldsymbol{h}_k\right) + \boldsymbol{b}^{||}$, where $||$ indicates the concatenation operator, $\boldsymbol{h}_k^{||} \in \mathbb{R}^g$, $\boldsymbol{W}^{||} \in \mathbb{R}^{g \times (5+v)}$ is a matrix of weights, and $\boldsymbol{b}^{||} \in \mathbb{R}^g$ is a bias vector. The new representation is then summed to the word embedding of the bounding box label to obtain the final visual features $\boldsymbol{h}_k^v = \boldsymbol{h}_k^{||} + \overline{\boldsymbol{e}}_k^{\mathcal{P}_{\boldsymbol{I}}}$, where $\boldsymbol{h}_k^v \in \mathbb{R}^g$.

Given the set $\overline{E}^{\boldsymbol{q}_j}$ of trainable word embeddings associated with the noun phrase $\boldsymbol{q}_j$, the textual branch applies a function f_{enc} to generate only one embedding $\boldsymbol{h}_j^t \in \mathbb{R}^\tau$ for each phrase $\boldsymbol{q}_j$. This textual features extraction is defined as $\boldsymbol{h}_j^t = f_{enc}(\overline{E}^{\boldsymbol{q}_j})$.

Note that the embeddings $\overline{E}^{\mathcal{P}_{\boldsymbol{I}}}$ and $\overline{E}^{\boldsymbol{q}_j}$ are generated with trainable modules that share the weights among each other (weights sharing). So, during training, the word embeddings learn multimodal embeddings for the visual and textual information.

4.3 Refined Predictions

The prediction module is in charge of refining the rough predictions $\boldsymbol{S}_{jk}$, i.e., the *Concept Branch* predicted scores, using the visual $\boldsymbol{h}_k^v$ and textual $\boldsymbol{h}_j^t$ features. Initially, starting from $\boldsymbol{h}_k^v$ and $\boldsymbol{h}_j^t$, the model predicts the probability $\boldsymbol{P}_{jk}$ that a bounding box proposal of index k is referred by the noun phrase $\boldsymbol{q}_j$ as $\boldsymbol{P}_{jk} = f_{\text{sim}}(\boldsymbol{h}_k^v, \boldsymbol{h}_j^t)$, where f_{sim} is a similarity measure between vectors. Please note that in our work, we adopt the cosine similarity function; therefore, $\boldsymbol{h}_k^v$ and $\boldsymbol{h}_j^t$ have the same vector dimension, i.e. $g = \tau$.

Finally, the rough predictions are refined in $\hat{\boldsymbol{P}}_{jk} = \omega * \boldsymbol{P}_{jk} + (1-\omega) * \boldsymbol{S}_{jk}$ using an hyper-parameter $\omega \in \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. Therefore, the model predictions are not constrained to values defined by concept similarity, but they co-work for the final predictions.

4.4 Loss Function

Inspired by [37], we adopt a contrastive loss. The contrastive objective $\mathcal{L}$ aims to learn the visual and textual features by maximizing the similarity score between paired image-sentence examples and minimizing the score between the negative examples.

Formally, given two training examples $(\boldsymbol{I}, \mathrm{S}), (\boldsymbol{I}', \mathrm{S}') \in \mathcal{D}$ such that $\mathrm{S} \neq \mathrm{S}'$ and $\boldsymbol{I} \neq \boldsymbol{I}'$, the loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = -\underbrace{f_{pair}(\boldsymbol{I}, \mathrm{S})}_{\text{Positive example}} + \underbrace{f_{pair}(\boldsymbol{I}', \mathrm{S})}_{\text{Negative example}}, \qquad f_{pair}(\boldsymbol{I}, \mathrm{S}) = \frac{1}{m} \sum_{j=1}^{m} \max_k \frac{\hat{\boldsymbol{P}}_{jk}}{\sum_i^p \hat{\boldsymbol{P}}_{ji}},$$

where f_{pair} is the similarity function defined over the multimodal pair image-sentence, m is the number of queries in S and $\hat{\boldsymbol{P}}_{jk}$ is the predicted similarity between noun $\boldsymbol{q}_j$ and proposal $\boldsymbol{p}_k$. Basically, the goal of f_{pair} is to aggregate the similarity scores of all the region-phrase pairs, determining the degree to which the phrases correspond with the content of the image.

In contrast to what is done in [37] where for each positive example, several negative examples built from the batch are considered, we adopt just a specific negative example $(\boldsymbol{I}', \mathrm{S})$. The negative example is built from the example $(\boldsymbol{I}', \mathrm{S}')$, selected from the batch precisely to be the one where the sentence S' is the most similar to the sentence S. This allows the model to focus on fine-grained region-phrase details that differ between the two examples. Precisely, given a training example $(\boldsymbol{I}, \mathrm{S}) \in \mathcal{B}$, the negative example is chosen as:

$$(\boldsymbol{I}', \mathrm{S}') = \underset{(\boldsymbol{I}'', \mathrm{S}'') \in \mathcal{B}'}{\operatorname{argmax}} f_{sim}(\zeta(\mathrm{S}''), \zeta(\mathrm{S})), \qquad \zeta(\mathrm{S}) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{L(\boldsymbol{q}_j)} \sum_{i=1}^{L(\boldsymbol{q}_j)} \boldsymbol{e}_i^{\boldsymbol{q}_j}$$

where $\mathcal{B}' = \mathcal{B} \backslash \{(\boldsymbol{I}, \mathrm{S})\}$. Thus, the similarity is measured in the word embedding space.

5 Experiments

5.1 Datasets and Evaluation Metrics

In this work, we have evaluated our model on the Flickr30k Entities [25] and ReferIt [17] datasets.[1] The Flickr30k Entities dataset contains 32K images and 360K queries, while the

[1] We considered the two most largely adopted datasets among the 15 papers used as comparison in our work.

ReferIt [17] dataset contains 20K images and 120K queries. Following the previous works in the area, we adopted *Accuracy* as the main evaluation metric. Namely, given a noun phrase, it considers a bounding box prediction to be correct if and only if the *Intersection over Union* value between the predicted bounding box and the ground truth bounding box is at least 0.5. Moreover, we also calculate the *Pointing Game Accuracy* for comparison purposes [1, 5, 23, 27]. *Pointing Game Accuracy* considers an example to be positive whether the center of the predicted bounding box lies wherever inside the ground-truth box.

5.2 Model Selection and Implementation

The model selected for evaluating the test set of the Flickr30k Entities and of the ReferIt datasets is chosen on the epoch that better performs in terms of *Accuracy* in the validation set. We search for the best hyper-parameters on both Flickr30k Entities and ReferIt datasets, independently on the considered fractions of training data $\{5\%, 10\%, 50\%, 100\%\}$ used for learning. We selected 10^{-5} as the learning rate step and we have adopted GloVe [24] as word embeddings, where $\tau = g = 300$. In our work, f_{enc} is implemented with a LSTM [32] neural network. The vector $\boldsymbol{h}_j^t$ is the τ-dimensional LSTM output of the last word $w_{L(\boldsymbol{q}_j)}^{\boldsymbol{q}_j}$ in the noun phrase $\boldsymbol{q}_j$. The bounding box proposals $\mathcal{P}_{\boldsymbol{I}}$ are extracted with the Bottom-Up Attention [2] object detector with a confidence score of 0.1 for Flickr30k[2] Entities and 0.2 for ReferIt[3]. The bounding box features have a dimension of $v = 2048$. We use the cosine similarity as a similarity measure f_{sim} between vectors. Our SPR model code is publicly available online[4].

5.3 Experimental Results

We compare our model to several approaches in the literature on the Flickr30k Entities and ReferIt datasets. We also assess our model performance when trained only with a small number of training examples. Indeed, the untrained *Concept Branch* module should give stability to the model even when it is trained on a small training set, as it should help to counter the overfitting trend that occurs with small datasets.

5.3.1 Full Training Set Scheme

Table 1 compares our model results to those of several approaches in the literature. Our model proposal outperforms all other approaches on standard *Accuracy* and *Pointing Game Accuracy*. In particular, in the Flickr30k Entities, our model's improvements over the State-of-the-Art are $+0.8\%$ in Accuracy and $+2.08\%$ in P. Accuracy. While on ReferIt, the improvements are $+9.65\%$ and $+3\%$, respectively for both the metrics.

To assess the soundness of our approach we tested a variant of our model that replaces visual and textual branches, responsible to learn the multimodal embedding space, with CLIP's multimodal embeddings (referred as SPR baseline + CLIP) [26]. As the results show, in Table 1, our full SPR model still outperforms the variant with CLIP. This occurs because CLIP was trained to capture the multimodal coarse-grained information from image and sentence pairs, while in VG we need more fine-grained details regarding the alignments region-query.

[2] We used the same features of [37].

[3] https://github.com/MILVLG/bottom-up-attention.pytorch

[4] https://github.com/drigoni/SPRM/

Model	Backbone (Pre-training)	Proposals (Pre-training)	Flickr30k E. (%)		ReferIt (%)	
			↑ Acc.	↑ P. Acc.	↑ Acc.	↑ P. Acc.
Top-down Saliency [27]	InceptionV3 (IN)	-	-	50.10	-	-
KAC Net [1]	VGG16 (VOC)	SS,EB	38.71	-	15.83	-
Semantic Self-Sup. [14]	VGG16 (IN)	-	-	49.10	-	39.98
Anchored Transformer [10]	VGG16 (VOC)	EB	33.10	-	13.61	-
Multi-level Multimodal [1]	PNASNet	-	-	69.19	-	48.42
Align2Ground [5]	RN152 (IN)	BUA (VG)	-	71.00	-	-
Counterf. Resilience [10]	RN101 (IN)	F-RCNN (CC)	48.66	-	-	-
MAF [31]	RN101 (IN)	BUA (VG)	61.4	-	-	-
Contrastive Learning [11]	RN101 (IN)	BUA (VG)	51.67	76.74	-	-
Grounding By Sep. [3]	VGG16, PNASNet (IN)	-	-	75.60	-	58.21
Relation-aware [23]	RN101	F-RCNN (VG)	59.27	78.60	37.68	58.96
Contrastive KL Distill. [36]	RN101 (IN)	BUA (VG)	53.10	-	38.39	-
EARN [22]	RN101	EB, F-RCNN	38.73	-	36.86	-
RefCLIP [15]	Darknet-53	YoLo3 (VG)	-	-	42.64	-
SimMaps [33]	VGG16 (IN)	-	45.56	79.95	38.74	70.25
SPR baseline + CLIP (ours)	RN101 (IN)	BUA (VG)	56.89	77.06	40.99	57.48
SPR model (ours)	RN101 (IN)	BUA (VG)	**62.20**	**80.68**	**48.04**	**62.40**

Table 1: **Results on Flickr30k Entities and ReferIt test sets**. *Acc.* is the standard accuracy metric, while *P. Acc.* is the pointing game accuracy metric. For each work we report in the column *Backbone* the adopted visual features encoder, abbreviating ResNet-101/152 [12] with RN101/152. In the *Proposals* column we listed proposal networks or object detectors employed, where SS: Selective Search [34], EB: EdgeBox [42], F-RCNN: Fast-RCNN [28], BUA: Bottom-up Attention [2]. In both columns, the pre-training dataset is indicated in parenthesis, whenever available, following these abbreviations: VG: Visual Genome [18], IN: ImageNet [6], CC: MS-COCO [20], VOC: PASCAL VOC [9].

The hyper-parameter ω regulates the weight of the *Concept Branch* on the final predictions: the higher the value, the less the *Concept Branch* affects final predictions. For this reason, in Figure 2 we present the *Accuracy* results obtained with our model trained on the entire training set at different values of ω: $\{0.1, 0.25, 0.4, 0.5, 0.75, 0.9\}$. As shown by the chart, ω greatly affects the model performance in both datasets, allowing the model to reach its peak of performance when $\omega = 0.4$ in Flickr30k Entities and $\omega = 0.75$ in ReferIt.

5.3.2 Small Training Set Scheme

In this section, we present the results obtained with our model on the datasets where only a fraction of training examples are used for training. Figure 3 reports our model *Accuracy* results. On Flickr30k Entities, the model is able to obtain State-of-the-Art results even when trained with only 50% of the training data, while on ReferIt, even when the model is trained with 5% of the training examples, it achieves State-of-the-Art performances. As expected, the *Concept Branch* module, which does not require training, makes the model training more stable and helps to counter the overfitting trend that occurs with small datasets.

5.3.3 Model Ablation

In this section, we assess the performance of our model's components: (i) the untrained *Concept Branch*, (ii) the trained visual and textual branches, (iii) and the *Relative Positional*

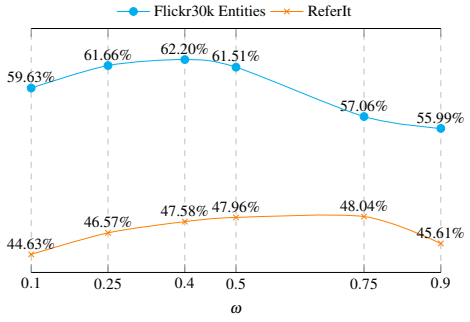

Figure 2: **Accuracy results on Flickr30k Entities and ReferIt test sets varying the ω hyper-parameter**. Results obtained by training the model on 100% of the training set.

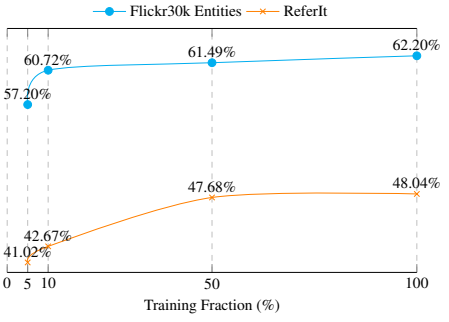

Figure 3: **Accuracy results on Flickr30k Entities and ReferIt test set by our model trained in low-data environments**. The percentage refers to the fraction of the training set considered during training.

Information component. The model achieves the best results when both the *Concept Branch* and the trained modules jointly work to produce the final predictions, as shown in Table 2.

The boost in *Accuracy* given by the *Concept Branch* is significant: +38.58% and +30.41% for Flickr30k Entities and ReferIt, respectively. As expected, the *Relative Positional Information* component constantly improves the model accuracy by +0.1% on Flickr30k Entities and by +2.6% on ReferIt. Further investigations showed that Flickr30k presents few spatial references in the queries, which explains the difference in performance gains between the two datasets.

Concept Branch	Trained Modules	Rel. Posit. Information	Flickr30k Entities (%)	ReferIt (%)
✗	✓	✗	23.52	15.03
✓	✗	✗	54.96	40.07
✓	✗	✓	55.02	42.69
✓	✓	✗	62.10	45.44
✓	✓	✓	**62.20**	**48.04**

Table 2: **Model Ablation**. Accuracy of our model's components. The *Concept Branch* contributes more to the final model performances.

5.3.4 Comparison to V&L models

The recent success of CLIP [26] in learning image-level visual representations from image-text pairs has inspired a new line of research [19, 41] to extend large V&L models on fine-grained correspondence between sentences and objects in images. In the same direction of research, in this section we compared our model to GLIP [19]. GLIP aims to learn region-level visual representations, thus enabling fine-grained alignment between image regions and textual concepts and works in a **fully-supervised fashion**. It is trained on 27M of grounding data, including Flickr30k. All the ground alignments are used, when available, during training. Thus the comparison between our model, which uses weak annotations, and GLIP is unfair. Nevertheless, we compared the two methods in the zero-shot setting, i.e. GLIP-T (B) trained only on object detection dataset Objects365 against our **untrained** Concept Branch (CB). GLIP-T (B) obtains 36.10% accuracy on Flickr30k while our CB scores 55.02%.

5.4 Limitations

Our model limitations stem mainly from the word embedding and the object detector components. In fact, the design of our proposal is well-suited for GloVe, Bottom-Up Attention, and LSTM components. However, these approaches are no more State-of-the-Art. Modern approaches, such as Large Language Models (LLMs) like BERT [7], could improve the performance of our model. Indeed, LLMs take advantage of their effective contextual capabilities to embed words in a sentence. In our architecture, LLMs can replace: (i) the LSTM in the *Textual Branch*, and (ii) the current GloVe embeddings in the *Concept Branch*. In both cases, the introduction of this new component is not straightforward, especially in the *Concept Branch*. In fact, the concept similarity scores are computed between the head of the phrase and bounding box classes. Thus, it is not clear what context the LLMs should consider during the embedding of class labels.

Furthermore, our model's dependency on the object detector performance is made explicit by the Concept Branch. Usually, in other works this dependency is hidden in the multimodal features fusion process, which relies on visual features and proposals from the object detector's output. To ensure an objective comparison of results, we use the same object detector that the current State-of-the-Art model MAF [37] utilizes. This detector is Faster-RCNN with ResNet-101, which has been trained on Visual Genome. We also make sure to adopt precisely the same features shared by MAF's authors. However, more recent object detectors could improve our model's performance.

6 Conclusion

Our work focused on tackling the task of weakly-supervised visual-textual grounding, where the lack of ground truth alignments presents a challenge for learning. Our core contribution resides in the *Concept Branch*. It captures the semantic similarity between the image's region and the phrase by matching the bounding box class and the phrase's head in a word embedding space. The alignments are obtained leveraging pre-trained object detector and word embedding, thus training is not required. The new knowledge does not depend on training data and can be treated as prior with many advantages. First, it enables compositionality as new models could be built on top of the prior to avoid starting from scratch. Second, this knowledge helps the training phase, especially in the first epochs where the model can't be guided as in the fully-supervised setting. Third, the independence of training data also enables performance stability which makes our model suitable in low-data environments. As proven by our results, this approach presents State-of-the-Art performance on Flickr30k Entities and ReferIt benchmarks. Inspired by [23], future works aim to extend our loss function to include a bounding box regression component, that has been proven to boost VG models performances. Additionally, future work will explore the use of different object detectors' categories [30]. Finally, inspired by [8], we aim to incorporate knowledge graph information in the model, enhancing the *Concept Branch* module with more structured information.

Acknowledgements. We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. This research were also supported by an UniPD BIRD-2021 Project and by the PRIN-17 PREVUE project from the Italian MUR (CUP E94I19000650001). We finally acknowledge EuroHPC Joint Undertaking for awarding us access to Vega at IZUM, Slovenia.

References

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

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

[3] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. *arXiv preprint arXiv:2104.09829*, 2021.

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

[5] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2601–2610, 2019.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. Vtkel: a resource for visual-textual-knowledge entity linking. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2021–2028, 2020.

[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*, 88:303–338, 2010.

[10] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019.

[11] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 752–768, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[14] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *arXiv preprint arXiv:1803.06506*, 2018.

[15] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2681–2690, 2023.

[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 *IEEE/CVF International Conference on Computer Vision ICCV*, 2021.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language processing (EMNLP)*, pages 787–798, 2014.

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

[19] 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. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, June 2022.

[20] 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.

[21] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620, 2019.

[22] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3003–3018, 2022.

[23] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5612–5621, 2021.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[25] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[27] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. Top-down visual saliency guided by captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7206–7215, 2017.

[28] 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*, 28:91–99, 2015.

[29] Davide Rigoni, Luciano Serafini, and Alessandro Sperduti. A better loss for visual-textual grounding. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022.

[30] Davide Rigoni, Desmond Elliott, and Stella Frank. Cleaner categories improve object detection and visual-textual grounding. In *Scandinavian Conference on Image Analysis*, pages 412–442. Springer, 2023.

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

[32] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[33] Tal Shaharabany and Lior Wolf. Similarity maps for self-training weakly-supervised phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6925–6934, 2023.

[34] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

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

[36] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021.

[37] Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language processing (EMNLP)*, 2020.

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

[39] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018.

[40] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018.

[41] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[42] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.