A Better Loss for Visual-Textual Grounding

Davide Rigoni
University of Padua
Bruno Kessler Foundation
Padua, Italy
drigoni@fbk.eu

Luciano Serafini
Bruno Kessler Foundation
Povo, Italy
serafini@fbk.eu

Alessandro Sperduti
University of Padua
Padua, Italy
sperduti@unipd.it

ABSTRACT

Given a textual phrase and an image, the visual grounding problem is the task of locating the content of the image referenced by the sentence. It is a challenging task that has several real-world applications in human-computer interaction, image-text reference resolution, and video-text reference resolution. In the last years, several works have addressed this problem by proposing more and more large and complex models that try to capture visual-textual dependencies better than before. These models are typically constituted by two main components that focus on how to learn useful multi-modal features for grounding and how to improve the predicted bounding box of the visual mention, respectively. Finding the right learning balance between these two sub-tasks is not easy, and the current models are not necessarily optimal with respect to this issue. In this work, we propose a loss function based on bounding boxes classes probabilities that: (i) improves the bounding boxes selection; (ii) improves the bounding boxes coordinates prediction. Our model, although using a simple multi-modal feature fusion component, is able to achieve a higher accuracy than state-of-the-art models on two widely adopted datasets, reaching a better learning balance between the two sub-tasks mentioned above.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**; **Object detection**.

KEYWORDS

Computer Vision, Visual Textual Grounding, Semantic Loss

ACM Reference Format:
Davide Rigoni, Luciano Serafini, and Alessandro Sperduti. 2022. A Better Loss for Visual-Textual Grounding. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event, .* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3477314.3507047

SAC '22, April 25–29, 2022, Virtual Event,

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507047

1 INTRODUCTION

In the last years, the scientific community has devoted much effort in developing deep learning models for computer vision and natural language processing, thanks to the increasing computational resources and the availability of new data. While deep learning models for computer vision aim to interpret and understand the visual world made by images and videos, deep learning models for natural language processing aim to interpret and understand the human natural language. In the last decade, these two research areas have made outstanding advancements that have lead to the formulation of more complex problems in which both vision an textual information are required, such as visual-question answering [2, 38, 54], image retrieval [13, 20, 21, 28], and visual grounding [4, 15, 36, 50]. Among these, of particular interest is the visual grounding problem, defined as the task of locating the content of the image referenced by a given sentence, a building block for many other real-world applications and more complex tasks. It is a challenging task, which requires the semantic understanding of the image content and its textual description, requiring the ability to predict the parts of the image content referred by a specific descriptive sentence. It can be formulated as an object detection task followed by a classification task in which, given an input image and sentence, the goal is to return only the detected object(s) in the image that represent(s) the best semantic match with the sentence. In the initial phase of research on this problem, many works have followed this formulation, developing the so called two-stage approach models [36, 49], while more recent works have chosen to address the problem by a one-stage approach model, in which the object detection and the classification problem are solved jointly [37, 46].

In the two-stage approach, the visual grounding model receives in input a set of proposal bounding boxes previously extracted by an object proposals extractor, such as Edge Boxes [55] and Selective Search [39], or by an object detector, such as Faster R-CNN [35], Single Shot multibox Detector (SSD) [24], or YOLO [33, 34]. These proposals, jointly with the given input textual sentence describing the content of the image, constitute the visual grounding model input. Usually, the model embeds the sentence in an embedding representation that tries to capture its semantic content. Then, the model predicts, for each proposal bounding box, a score that represents how much the content of the bounding box is likely to be referred by the sentence. Often, the two-stage approach models predict new coordinates for the best predicted proposal, in order to adjust the coordinates to better fit the visual content according to the sentence semantic information.

In the one-stage approach, the visual grounding model receives in input only an image and a textual sentence. Then the model learns how to extract and fuse all the visual and textual information

in order to predict the best bounding box in output, according to the input sentence. Even if this seems to be the best approach in order to reach the best results, due to the small number of assumptions made by the model, it raises some major issues: (*i)* not all the visual grounding datasets are suitable for training an object detector, due to lack of images and/or because they are not densely annotated; (*ii)* the model requires an high number of parameters, and because of that (*iii)* the training requires significant computing resources.

In the literature, there are many works adopting increasingly improved object proposals, and increasingly complex architectures than before in order to capture the visual and textual information. These models are typically constituted by two main components that focus on how to learn useful multi-modal features for grounding, and how to improve the predicted bounding box of the visual mention, respectively. Finding the right learning balance between these two sub-tasks is not easy, and the current models are not necessarily optimal with respect to this issue. In this work, we propose a model that, although using a simple multi-modal feature fusion component, is able to reach a higher accuracy than state-of-the-art models thanks to the adoption of a more effective loss function that reaches a better learning balance between the two sub-tasks mentioned above.

Our main contributions can be summarized as: (i) we present a new loss for visual proposals which considers also the object proposals semantic information, differently from the works in the literature which just consider their shapes and spatial positions in the image; (ii) we are the first to adopt the *Complete Intersection over Union* [52] loss for the visual grounding task; (iii) we introduce a new regression loss on the proposal bounding boxes coordinates which is applied to a subset of all the proposals, selected by considering the object proposals semantic information. This loss differs from the one used by the approaches in the literature, which only considers the proposal with the largest overlap with the ground truth. (iv) we experimentally show that the proposed losses improve the performance of state-of-the-art models.

2 RELATED WORKS

In this section, we report the important related works developed within three areas, namely, Referring Expression Grounding, Visual-Textual-Knowledge Entity Linking (VTKEL), and Image Retrieval.

Referring Expression Grounding. It is also known as phrase localization or visual grounding. It aims to localize the corresponding objects described by a human natural language phrase in an image. It is common to extract visual and language features independently and fuse them before the prediction. Some works apply a multi layer perceptron (MLP) [4, 5], cosine similarity [12] and element-wise multiplication. Other works apply more complex strategies such as Canonical Correlation Analysis (CCA) [31, 32], Multimodal Compact Bilinear (MCB) [14], graph structures [3] and attention methods [29]. Instead of focusing on the fusion component, [49] proposes a visual grounding model with diverse and discriminative proposals that can achieve good performance without using a complex multi-modal fusion operator. The approach presented in [1] predicts the image content location referred by the input phrase using an heatmap, applying a multi-level multi-modal attention mechanism, instead of relying on the standard bounding box. Some works focus on the weakly supervised referring expression grounding setting, in which it is available only the information about the image contents and there is no mapping among the input textual sentences and these visual locations. Given a set of bounding boxes proposals and a textual sentence in input, [36] introduces a model which learns to ground by reconstructing the given textual sentence adopting a soft attention mechanism. This approach is extended in [4] by the introduction of a novel Knowledge Aided Consistency Network (KAC Net) which is optimized by reconstructing the input query and proposal's information. A different approach is developed in [45], where an end-to-end model learns to localize arbitrary linguistic phrases in the form of spatial attention masks, using two types of carefully designed loss functions. A Variational Context model, based on the variational Bayesian method, is adopted by [50] to exploit the reciprocal relation between the referent and context. The Multi-scale Anchored Transformer Network (MATN) [51] hinges on the concept of anchors, i.e. it uses region proposals as localization anchors, learning a multi-scale correspondence network to continuously search for sentences referring to the anchors. The work presented in [15] shows that textual sentence grounding can be learned by optimizing word-region attention to maximize a lower bound on mutual information between images and caption words. [25] uses a cross-modal attention-guided erasing approach, where it discards the most dominant information from either textual or visual domains to generate difficult training samples online in order to drive the model to discover complementary textual-visual correspondences. [7] provides an accumulated attention (A-ATT) mechanism to ground the natural language query into the image using a query attention, an image attention and an objects attention.

Visual-Textual-Knowledge Entity Linking. The VTKEL task [9–11] introduces a more complex task than the referring expression task, in which an artificial agent needs to jointly recognize the entities shown in the image and mentioned in the text, and to link them to its prior background knowledge. The solution to the VTKEL problem could lead to major scientific advancement towards a better understanding of semantic information contained in the image and textual sentence, respectively. In fact, the knowledge graph allows to introduce semantic reasoning on the information contained in both the image and the textual sentence, which could lead to innovative solutions for the weakly supervised referring expression problem and for the partially annotated dataset problem.

Image Retrieval. The standard text-base image retrieval systems, given a textual sentence in input, from a set of images select the one that best matches the textual input. In particular, the best images are returned according to some metric learned through a recurrent neural network [28], correlation analysis [21] and other methods [13, 20].

3 BACKGROUND

In order to explain our work, we use the following notation: lower case symbols for scalars and indexes, e.g. n; italics upper case symbols for sets, e.g. A; upper case symbols for textual sentences, e.g. S; bold lower case symbols for vectors, e.g. $\boldsymbol{a}$; bold upper case symbols for matrices and tensors, e.g. $\boldsymbol{A}$; the position within a tensor or

vector is indicated with numeric subscripts, e.g. $\boldsymbol{A}_{ij}$ with $i, j \in \mathbb{N}^+$; calligraphic symbols for domains, e.g. $\boldsymbol{Q}$.

In our work we adopt the *Complete IoU (CIoU)* [52] loss to perform the bounding boxes coordinates regression, that is based on the *Intersection over Union (IoU)* metric. Given a pair of bounding box coordinates $(\boldsymbol{b}_i, \boldsymbol{b}_j)$, the *Intersection over Union*, also known as Jaccard index, is an evaluation metric used mainly in object detection tasks, which aims to evaluate how much the two bounding boxes refer to the same content in the image. It is defined as:

$$IoU(\boldsymbol{b}_i, \boldsymbol{b}_j) = \frac{|\boldsymbol{b}_i \cap \boldsymbol{b}_j|}{|\boldsymbol{b}_i \cup \boldsymbol{b}_j|}, \tag{1}$$

where $|\boldsymbol{b}_i \cap \boldsymbol{b}_j|$ is the area of the box obtained by the intersection of boxes $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$, while $|\boldsymbol{b}_i \cup \boldsymbol{b}_j|$ is the area of the box obtained by the union of boxes $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$. It is invariant to the bounding boxes sizes, and it returns values that are strictly contained in the interval $[0, 1] \subset \mathbb{R}$, where 1 means that the two bounding boxes refer to the same image area, while a score of 0 means that the two bounding boxes do not overlap at all. The fact that two bounding boxes that do not overlap have *IoU* score equal to 0, is the major issue of this metric: the zero value does not represent how much the two bounding boxes are far from each other. For this reason, in its standard definition, the *IoU* function is mainly used as an evaluation metric rather than as a component of a loss function for learning.

In order to solve the issue of *IoU* when considering it as a loss function, [52] proposed the *Complete IoU* loss that is defined as:

$$\mathcal{L}_{CIoU}(\boldsymbol{b}_i, \boldsymbol{b}_j) = S\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right) + D\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right) + V\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right) \tag{2}$$

$$S\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right) = 1 - IoU(\boldsymbol{b}_i, \boldsymbol{b}_j); \tag{3}$$

$$D\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right) = \frac{\rho\left(\boldsymbol{p_i}, \boldsymbol{p}_j\right)^2}{c^2}; \tag{4}$$

$$V\left(\boldsymbol{b}_i, \boldsymbol{b}_j\right) = \alpha \frac{4}{\pi^2}\left(\arctan\frac{wt_j}{ht_j} - \arctan\frac{wt_i}{ht_i}\right) \tag{5}$$

where $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ are two bounding boxes, $\boldsymbol{p_i}$ and $\boldsymbol{p}_j$ are their central points, $IoU(\boldsymbol{b}_i, \boldsymbol{b}_j)$ is the standard *IoU*, ρ is the euclidean distance between the given points, c is the diagonal length of the *convex hull* of the two bounding boxes, α is a trade-off parameter, wt_i and ht_i are the width and the height of the bounding box $\boldsymbol{b}_i$, respectively. Differently from the standard *IoU*, the *Complete IoU* is formulated in such a way to return meaningful values, leveraging the bounding boxes geometric shapes, even when two bounding boxes are not overlapped.

4 PROBLEM DEFINITION

Visual grounding is the general task of locating the components of a structured description in an image. In order to solve this task, first, it is necessary to recognize all the objects in the image and the components in the text, while after, the model needs to find the correct alignment among the nouns and the objects. Each detected object in the image is usually represented by a rectangle called bounding box, while each noun phrase detected in the text is usually called query. The bounding box is determined by its position in the image and by its dimension, while the query is determined by the position of the first character and the position of the last character in the input text.

Formally, given in input an image $\boldsymbol{I}$ and a sentence S describing some of the objects represented in $\boldsymbol{I}$, the task consists in learning a map γ from the set Q of noun phrases contained in S to a set of bounding boxes B defined on $\boldsymbol{I}$, i.e. $\gamma : \mathcal{I} \times \mathcal{S} \rightarrow 2^{\mathcal{Q} \times \mathcal{B}}$, where $\mathcal{I}$ is the domain of images, $\mathcal{S}$ is the domain of sentences, $\mathcal{Q}$ is the noun phrases domain, $\mathcal{B}$ is the domain of bounding boxes which can be defined on $\mathcal{I}$, and $2^{\mathcal{Q} \times \mathcal{B}}$ is the power set of the Cartesian product between $\mathcal{Q}$ and $\mathcal{B}$. So, given an image $\boldsymbol{I}$ containing e objects identified via the set of bounding boxes $B_{\boldsymbol{I}} = \{\boldsymbol{b}_i\}_{i=1}^{e}$, where $\boldsymbol{b}_i \in \mathbb{R}^4$ is the vector of coordinates identifying a bounding box in $\boldsymbol{I}$, and a sentence S containing m noun phrases gathered in the set $Q_{\mathrm{S}} = \{\boldsymbol{q}_j\}_{j=1}^{m}$, where $\boldsymbol{q}_j \in \mathbb{N}^2$ is a vector containing the initial and final character positions in the sentence S, $\gamma(\boldsymbol{I}, \mathrm{S})$ returns a subset $\Gamma \subseteq Q_S \times B_{\boldsymbol{I}}$ where each couple $(\boldsymbol{q}, \boldsymbol{b}) \in \Gamma$ associates the noun phrase $\boldsymbol{q}$ to the bounding box $\boldsymbol{b}$. Please, notice that the same noun phrase can be associated to several different bounding boxes, as well as the same bounding box can be associated to many different noun phrases. Following the current literature, in this paper we assume that each noun phrase is associated to one and only one bounding box. A bounding box, however, can identify more objects, e.g. several persons in the case the noun phrase is "people". A training set of n examples is defined as $D = \{(\boldsymbol{I}_i, \mathrm{S}_i, \Gamma_i^{gt})\}_{i=1}^{n}$, where Γ_i^{gt} is the set of ground truth associations for example i.

5 OUR PROPOSAL

In this section, we first describe the structure of our model, and then we describe the training procedure, which exploits the original part of our proposal, e.g. a loss function composed of novel sub-losses.

5.0.1 Model. Our model, outlined in Figure 1, follows a typical basic architecture for visual-textual grounding tasks. It is based on a two-stage approach in which, initially, a pre-trained object detector is used to extract, from a given image $\boldsymbol{I}$, a set of k bounding box proposals $\mathcal{P}_{\boldsymbol{I}}$, jointly with visual features H^v. The features represent the internal object detector activation values before the classification layers and regression layer for bounding boxes. Moreover, our model extracts the spatial features H^s from the proposals. We also assume that the object detector returns, for each bounding box proposal $\boldsymbol{p}_i \in \mathcal{P}_{\boldsymbol{I}}$, a probability distribution $Pr_{Cls}(\boldsymbol{p}_i)$ over a set Cls of predefined classes, i.e. the probability for each class $\xi \in Cls$ that the content of the bounding box proposal $\boldsymbol{p}_i$ belongs to ξ. This information is typically returned by most of the object detectors, and it will be used to define our novel loss terms.

Regarding the textual features extraction, given a noun phrase $\boldsymbol{q}_j$, initially all its words $W^{\boldsymbol{q}_j}$ are embedded in a set of vectors $E^{\boldsymbol{q}_j}$. Then, our model applies a LSTM [16] neural network to generate from the sequence of word embeddings only one new embedding $\boldsymbol{h}_j^{\star}$ for each phrase $\boldsymbol{q}_j$. Once vector $\boldsymbol{h}_j^{\star}$ has been generated from the noun phrase $\boldsymbol{q}_j$, the model performs a multi-modal feature fusion operation in order to combine the information contained in $\boldsymbol{h}_j^{\star}$ with each of the proposal bounding boxes $\boldsymbol{h}_z^v \in H^v$. For this operation, we have decided to use a simple function that merges the multi-modal features together rather than relying on a more complex operator, such as bilinear-pooling or deep neural network architectures. We leave the use of a more complex fusion operator,

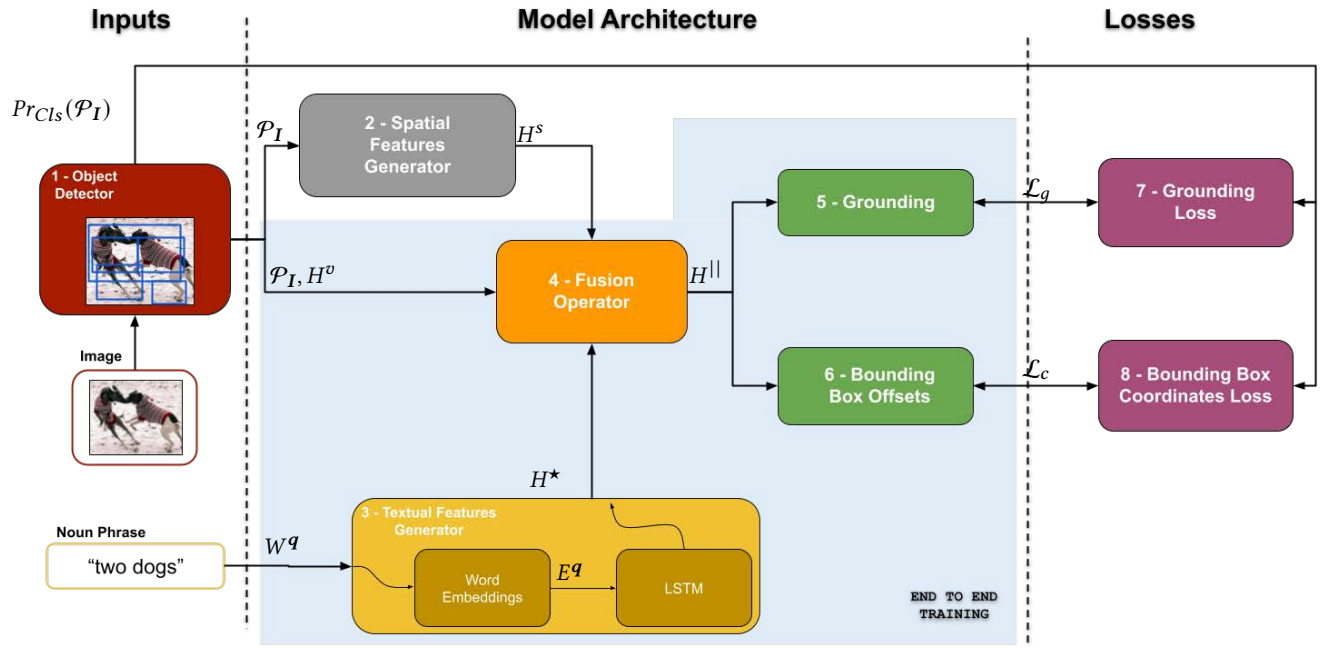

Figure 1: Our two-stage model architecture overview. (1) Initially, the image is processed by a pre-trained *Faster R-CNN* object detector in order to extract all the proposals bounding boxes from which (2) the spatial features are generated. Then, the model (3) generates the textual features from the input noun phrase using the *Textual Features Generator* module, by first retrieving each word embedding and then using an LSTM network. Finally, the model (4) fuses together all the visual, spatial, and textual features by the *Fusion Operator*, obtaining new features that are then used in the (5) *Grounding* and (6) *Bounding Box Offsets* modules, respectively. Our defined losses $\mathcal{L}_g$ (7) and $\mathcal{L}_c$ (8) are used in order to train the network end-to-end on the components included in the light blue background.

that will lead to further improvements, for future work. The multimodal fusion component we adopted returns the set of new vectorial representations $H^{||}$.

Finally, the model predicts the probability $\boldsymbol{P}_{jz}$ that a given noun phrase $\boldsymbol{q}_j$ is referred to a proposal bounding box $\boldsymbol{p}_z$. Indeed, the representations of the proposals bounding box features conditioned with the textual features can also be used to refine the proposal bounding box coordinates, that are generated by the object detector independently by the textual features. Specifically, our model does not predicts new bounding box coordinates, but offsets for the coordinates.

Technical details regarding the model are reported in the Supplementary Material.

5.0.2 Training. In this section, we present the main novel contribution of the paper, i.e. a loss function composed of novel terms. The basic idea is to exploit the semantic information associated with bounding box proposals, i.e. the probability distribution over classes of the content of a bounding box returned by the object detector, in both the loss term concerning the grounding and the loss term concerning the refinement of the bounding box coordinates. In fact, differently from most of the previous works that use the *cross-entropy (CE)* loss or the standard *Kullback–Leibler(KL) divergence* loss for grounding, our model implements a KL divergence loss in which the ground truth probability is built also considering $Pr_{Cls}(\boldsymbol{p}_i)$ with $\boldsymbol{p}_i \in \mathcal{P}_I$. Moreover, regarding the bounding boxes coordinates refinement, differently from previous works that use the $Smooth_{L1}$ *loss*, our model adopts the *CIoU loss* [52]. To the best of our knowledge, this is the first work adopting the *CIoU* loss in order to refine the final bounding boxes coordinates. Another difference with respect to all the refinement losses proposed in the literature is that we do not restrict the coordinates refinement only to the best proposal coordinates, but we extend the refinement to the subset of proposals that significantly overlap (according to an hyper-parameter) the ground truth, modulating the refinement by the agreement between the class probability of the best proposal and the class probability of the considered proposal. For the sake of presentation, we formally define the new loss terms in the following referring to a single example. The total loss is then obtained by summing up the contributions of all examples in the training set.

Given a training example $(\boldsymbol{I}, \mathrm{S}, \Gamma^{gt})$, and the bounding box proposals set $\mathcal{P}_I$, we define the loss function $\mathcal{L}$ (for a single example) as:

$$\mathcal{L} = \mathcal{L}_g(\boldsymbol{P}, \mathcal{P}_I, \Gamma^{gt}) + \lambda \mathcal{L}_c(\mathcal{P}_I, \Gamma^{gt}),$$

where $\mathcal{L}_g$ is the loss used to "shape" the grounding distribution of proposals for each specific query in input, i.e. the probability that a given proposal is associated to a given query, $\mathcal{L}_c$ is the loss related to the refinement of the bounding boxes coordinates, and λ is a trade-off parameter.

Specifically, given m the number of noun phrases and k the number of bounding box proposals, we define the entries ($j \in [1,\dots,m]$, $z \in [1,\dots,k]$) of matrix $\boldsymbol{U}$ as $\boldsymbol{U}_{jz} = IoU(\boldsymbol{b}_j^{gt}, \boldsymbol{p}_z)$ where $(\boldsymbol{q}_j^{gt}, \boldsymbol{b}_j^{gt}) \in \Gamma^{gt}$, the best proposal bounding box as $\boldsymbol{p}_{j*}$ where $j* = argmax_{z\in[1,\dots,k]}\, \boldsymbol{U}_{jz}$, and the entries ($j \in [1,\dots,m]$, $z \in [1,\dots,k]$) of matrix $\boldsymbol{C}$ containing the cosine similarity scores among the predicted class probabilities of the bounding box proposals as $\boldsymbol{C}_{jz} = Sim\left(Pr_{Cls}(\boldsymbol{p}_{j*}), Pr_{Cls}(\boldsymbol{p}_z)\right)$, where Sim is the cosine similarity function. Given these definitions, we can define the entries of the target probability $\boldsymbol{P}^{target}$ as:

$$\boldsymbol{P}_{jz}^{target} = \frac{\boldsymbol{U}_{jz}^*}{\sum_{i=1}^{k} \boldsymbol{U}_{ji}^*}, \text{ where}$$

$$\boldsymbol{U}_{jz}^* = \begin{cases} \boldsymbol{U}_{jz}\boldsymbol{C}_{jz}, & if\ \ \boldsymbol{U}_{jz} \geq \eta \\ 0, & otherwise \end{cases},$$

and η is a predefined threshold, i.e. an hyper-parameter.

On the basis of the above definitions, we define the grounding loss as:

$$\begin{aligned} \mathcal{L}_g(\boldsymbol{P}, \mathcal{P}_{\boldsymbol{I}}, \Gamma^{gt}) &= \frac{1}{m}\sum_{j=1}^{m} KL_{div}(\boldsymbol{P}_j || \boldsymbol{P}_j^{target}), \\ &= \frac{1}{m}\sum_{j=1}^{m}\sum_{z=1}^{k} \boldsymbol{P}_{jz} \log\left(\frac{\boldsymbol{P}_{jz}}{\boldsymbol{P}_{jz}^{target}}\right), \end{aligned}$$

where KL_{div} is the KL divergence function, $\boldsymbol{P}_j$ ($\boldsymbol{P}_j^{target}$) is the j-th row of $\boldsymbol{P}$ ($\boldsymbol{P}^{target}$), and $\boldsymbol{P}_{jz}$ is the model predicted probability that the noun phrase $\boldsymbol{q}_j \in Q$ refers to the image content localized by $\boldsymbol{p}_z \in \mathcal{P}_{\boldsymbol{I}}$.

Indeed, the grounding loss captures both the bounding box spatial information and the semantic information determined by the bounding box classes. Whenever a bounding box is located near the ground truth bounding box and its class probability distribution is similar to the one of the best proposal $\boldsymbol{p}_{j*}$, then the loss favours the prediction of the bounding box, otherwise the loss penalizes the bounding boxes according to their different probability distribution and spatial location. Previous works exploiting the KL divergence aims to maximize the probability of a proposal bounding box just considering their spatial location.

We now define the novel refinement loss. In order to do that, given a query $\boldsymbol{q}_j$, we need to define the following subset $\mathcal{S}_j \subseteq \mathcal{P}_{\boldsymbol{I}}$ of proposals:

$$\mathcal{S}_j = \{\boldsymbol{p}_z \mid \boldsymbol{p}_z \in \mathcal{P}_{\boldsymbol{I}} \wedge \boldsymbol{U}_{jz}^* \geq 0\},$$

which allows us to define our loss $\mathcal{L}_c$ as:

$$\mathcal{L}_c(\mathcal{P}_{\boldsymbol{I}}, \Gamma^{gt}) = \frac{1}{m}\sum_{j=1}^{m}\sum_{\boldsymbol{p}_z \in \mathcal{S}_j} \hat{\boldsymbol{U}}_{jz}\mathcal{L}_{CIoU}(\boldsymbol{p}_z, \boldsymbol{b}_j^{gt}),$$

where $(\boldsymbol{q}_j^{gt}, \boldsymbol{b}_j^{gt}) \in \Gamma^{gt}$, and

$$\hat{\boldsymbol{U}}_{jz} = \frac{\boldsymbol{U}_{jz}^*}{max_{z\in[1,k]}\boldsymbol{U}_{jz}^* + \epsilon},$$

in which ϵ is a small value added to avoid division by 0, and $max_{z\in[1,k]}$ is the maximum function applied along the indexes $z \in [1,k]$. Intuitively, for each bounding box proposal which overlaps with the ground truth (according to the parameter η), this loss refines the coordinates proportionally to the "semantic" of the bounding box. Note that adopting the normalized scores $\hat{\boldsymbol{U}}_{jz}$, the model does not penalize the loss on the best proposal bounding box $j*$.

We would like to highlight that our work is the first proposing the exploitation of the probabilities distributions over the object detector classes to address the supervised visual grounding task. However, in weakly-supervised visual-textual grounding (*not our task*) some works (e.g. [40]) leverage the information of the bounding box class with the *highest* probability.

6 EXPERIMENTAL ASSESSMENT

We have compared our model results on two widely adopted datasets (i.e., Flickr30k Entities and ReferIt) considering several competing approaches in the literature, including state-of-the-art models. In addition to that, in order to prove the usefulness of our losses independently by our model architecture, we have also adopted our losses on the DDPN model. The choice of this model was due to: (i) publicly available code[1]; (ii) published results on both Flickr30k Entities and ReferIt datasets, with state-of-the-art results on ReferIt; (iii) and exploitation of the same object detector used in our work.

6.1 Datasets and Evaluation Metric

Flickr30k Entities and ReferIt constitute the two most common datasets used in the literature, although other datasets have been used (e.g., [6, 18, 27, 53]). The Flickr30k Entities dataset [32, 48] contains 32K images, 275K bounding boxes, 159K sentences, and 360K noun phrases. The ReferIt [19] dataset contains 20K images, 99K bounding boxes, and 130K noun phrases. This dataset differs from Flickr30k Entities since it does not contain sentences, which means that the noun phrases are mutually independent. We refer the reader to the Datasets Details section of the Supplementary Material for more details.

Aligned with the works in the literature, we adopted the standard *Accuracy* metric. 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.

6.2 Model Selection and Implementation Details

To evaluate our model on the test set of Flickr30k Entities and ReferIt datasets, we have chosen the epoch in which the model achieved the best *Accuracy* metric on the validation set. We have performed a grid search for the best hyper-parameters mainly for the Flickr30k Entities dataset, ad exception of the losses hyper-parameters visible in Section 6.3.2. For the ReferIt dataset, we have used the other hyper-parameters values selected on the Flickr30k Entities dataset. We have used the Adam optimizer with exponential learning rate scheduler set to 0.9, and the following values for the learning rate: $\{0.05, 0.03, 0.01, 0.005, 0.001\}$, $c : \{2048, 2053, 2060\}$, and $\eta : \{0.1, 0.3, 0.4, 0.45, 0.5, 0.55\}$. Other hyper-parameters are

[1]We have adapted the official code: https://github.com/XiangChenchao/DDPN.

Table 1: Results obtained on Flickr30k test set. *Accuracy* indicates in percentage the standard accuracy metric. All values are copied from the original articles. "*" indicates that the reported model accuracy is referring to the version of the model in their ablation study, since the complete model uses query dependency information that we do not exploit.

Model	Accuracy (%)
SCRC [17]	27.80
SMPL [43]	42.08
NonlinearSP [42]	43.89
GroundeR [36]	47.81
MCB [14]	48.69
RtP [32]	50.89
Similarity Network [41]	51.05
IGOP [47]	53.97
SPC+PPC [31]	55.49
SS+QRN [5]	55.99
SeqGROUND [8]	61.60
CITE [30]	61.89
QRC net [5]	65.14
YOLO [46]	68.69
DDPN [49]	73.30
CMGN [26]*	73.46
SL-CCRF [23]	74.69
Ours	**75.55**
DDPN [49] using our losses	74.33

fixed to single values. For the textual features: $w = 300$, $t = 500$, and the LSTM network uses only one hidden layer of dimension t. For the image features, we have extracted a fixed number $k = 100$ of proposals for each image, $v = 2048$ from the ResNet-101's layer *pool5_flat*, and $s = 5$. In both datasets, we have found that the best model *Accuracy* is achieved at epoch 9 of training with learning rate set to 0.001 and $c = 2053$. For Flickr30k Entities we have set $\eta = 0.3$ and $\lambda = 1$, while for ReferIt we have set $\eta = 0.5$ and $\lambda = 1.4$. The code is publicly available on GitHub [2]. We refer the reader to the Implementation Details section of the Supplementary Material for more details.

6.3 Results

Table 1 reports the results obtained on the Flickr30k Entities dataset by our approach and many other approaches presented in the literature, including the most recent state-of-the-art models reported at the bottom part of the table. Concerning the model CMGN developed in [26], for the sake of a fair comparison, we have reported the performance obtained using the same setting of our model. In fact, the complete version of the CMGN model achieves an *Accuracy* of 76.74%, but exploiting query dependency information that we could exploit as well. The integration of this information in our model is left for future work. It can be noted that our approach significantly improves over competing approaches. Moreover, the DDPN model where our losses are used (last row of the table) shows a significant improvement in performance (1.03%) with respect to the original version.

[2]https://github.com/drigoni/Loss_VT_Grounding

Table 2: Results obtained on ReferIt test set. *Accuracy* indicates in percentage the standard accuracy metric. All values are reported from the original articles.

Model	Accuracy (%)
SCRC [17]	17.93
GroundeR [36]	26.93
MCB [14]	28.91
CITE [30]	34.13
IGOP [47]	34.70
[44]	36.18
QRC net [5]	44.10
[22]	44.20
[46]	59.30
DDPN [49]	63.00
Ours	66.02
DDPN [49] using our losses	**66.66**

Table 2 reports the results obtained on the ReferIt dataset by our approach and the subset of the competing approaches reported in Table 1 that can be applied to this dataset, plus additional approaches that have been assessed on this dataset[3]. Our model improves the *Accuracy* value by 3.02% when compared to the state-of-the-art model (i.e., DDPN) for this dataset, representing a more significant gain than the one obtained on Flickr30k Entities. On the other hand, adopting our losses in DDPN leads to the best performance, with an improvement over the original version of 3.66%. In the ReferIt dataset, each sentence corresponds to a single query independently from the others. In contrast, in Flickr30k Entities, a sentence could contain more queries that are semantically related among them. For this reason, models that apply complex multimodal feature fusion components that aim to capture information among the queries extracted by the sentence in input sometimes do not consider the ReferIt dataset. Thus, the set of the models used as comparison in the ReferIt dataset is not the same as in Flickr30k Entities and these reasons could explain the higher gain in *Accuracy* obtained in ReferIt than Flickr30k Entities.

We have also calculated the *Point Game Accuracy* which is recently used for a few models addressing the weakly-supervised task. It considers a prediction to be correct if and only if the center of the predicted bounding box is contained in the ground truth bounding box. In particular, our model obtains 87.96% and 78.0% on Flickr30k Entities and ReferIt, respectively. These values are far better than the ones reported in the literature, and they suggest that a significant subset of predictions that are considered to be wrong according to the *Accuracy* metric, still refer to bounding boxes that have a significant overlap with the ground truth.

6.3.1 Qualitative Results. Figures 2,3, and 4 show qualitative examples predicted by our model on the test set of both Flickr30k Entities and ReferIt datasets. In our model predictions, we have noticed that when the query refers to a small object in the image, most of the time our model predicts a very close bounding box, but not enough to have the IoU score over the 0.5 value. This is the case for the query "a tennis ball" in the figure 2. More examples are

[3]Some of them do not define an acronym, so we just use the reference to the paper.

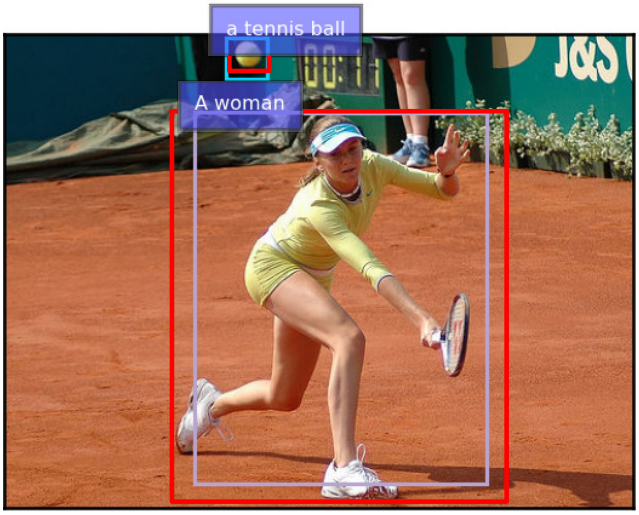

Figure 2: This picture reports a qualitative example of our model on the Flickr30k test image id: 23016347**. The ground truth bounding boxes associated with each query are reported in red.The prediction for the query "a tennis ball" is evaluated as wrong, even if the bounding box is very close to the ground truth.**

Figure 3: This picture reports a qualitative example of our model on the ReferIt test image id: 14651 **. The ground truth bounding box is reported in red. The complete sentence in input is reported at the bottom of the figure. The predicted bounding box presents an intersection over union value with the ground truth of 0.08.**

reported on the Qualitative Results section of the Supplementary Material.

6.3.2 Ablation Study. Our loss is composed by two main components and by two hyper-parameters. Here, we report the contribution of each part of the loss using different hyper-parameters values. We have performed a set of experiments where the grounding component is alternatively the cross-entropy, the KL divergence or the proposed semantic KL divergence, and the regression component is alternatively the Smooth L1 or the proposed semantic CIoU. Moreover, different values for the hyper-parameters are considered. The obtained results (Table 3) show that the major contribution to the improvement is given by the *Complete IoU loss with semantic information*, which improves the model *Accuracy* by $\sim$ 2.6% and $\sim$ 3.9% on Flickr30K Entities and ReferIt datasets, respectively. Significant improvements are also obtained by using the semantic KL divergence in place of cross-entropy or the CIoU-Sem instead of the standard CIoU. Moreover, results show that our approach is not much sensitive with respect to the hyper-parameters values, and, more importantly, the *Accuracy* on the validation set indeed represents well the *Accuracy* on the test set on both datasets.

7 CONCLUSION AND FEATURE WORK

This paper introduced a novel loss for Visual-Textual Grounding, jointly with a simple two-stage approach model. The novel loss combines a grounding loss and a bounding box coordinates refinement loss, both based on semantic information, i.e. a probability distribution over a set of pre-defined classes, returned by the object detector. The experimental assessment showed that the proposed approach was able to reach a higher accuracy than state-of-the-art models, even without using a more complex multi-modal feature fusion component. Specifically, we have compared our results versus several models in the literature over two commonly used datasets, Flickr30K Entities and ReferIt. With respect to the best state-of-the-art approaches, on the Flickr30K Entities dataset, we obtained an improvement of 0.86%, while on the ReferIt dataset, our model improved the state-of-the-art performance by 3.02%. By applying the proposed loss to the DDPN model we were able to significantly improve the performance of the model on both datasets, demonstrating its usefulness independently from the proposed model.

Since this model uses a simple multi-modal feature fusion component, there is space for trivial improvements, including a more sophisticated multi-modal feature fusion component, such as bilinear-pooling and deeper architectures, as well as the exploitation of dependencies among the queries contained by the input sentence. Future work will also address more sophisticated object detectors, and the idea to include different forms of information, such as a scene graph and prior knowledge.

8 SUPPLEMENTARY MATERIAL

We have reported all the supplementary material available at the following address: https://www.math.unipd.it/~drigoni/files/SAC_2022_A_Better_Loss_for_Visual_Textual_Grounding_Supplementary.pdf.

REFERENCES

[1] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2019. Multi-Level Multimodal Common Semantic Space for

Table 3: Accuracy obtained on Flicker30k Entities and ReferIt datasets as the losses functions and hyper-parameters values change. *CE* indicates the cross-entropy loss, *SmoothL1* indicates the Smooth L1 loss, *KL-Sem* indicates our KL loss with the semantic information and *CIoU-Sem* indicate our Complete IoU loss with the semantic information. The baseline model does not use the η parameter.

Losses		Hyper-par.		Flickr30k (%)		ReferIt (%)	
Gr.	Reg.	λ	η	Val.	Test	Val.	Test
CE	SmoothL1	0.8	/	**71.25**	**71.82**	64.24	**61.81**
		1	/	71.08	71.61	64.19	61.29
		1.2	/	71.18	71.21	**64.65**	61.64
KL	SmoothL1	0.8	0.4	71.51	72.06	63.58	61.38
		0.8	0.5	72.16	**72.55**	64.57	**62.69**
		1	0.4	71.76	72.34	63.93	61.65
		1	0.5	**72.58**	72.18	**64.82**	62.49
KL-Sem	SmoothL1	0.8	0.4	72.22	72.72	64.38	61.78
		0.8	0.5	72.42	72.41	64.99	62.12
		1	0.4	**72.54**	**72.88**	65.04	62.47
		1	0.5	72.34	72.83	**65.45**	**62.72**
CE	CIoU-Sem	0.8	0.4	73.99	74.56	**67.66**	**65.47**
		0.8	0.5	73.60	74.24	67.41	65.07
		1	0.4	**74.07**	**74.82**	67.60	65.42
		1	0.5	73.90	74.24	67.24	65.15
KL-Sem	CIoU-Sem	0.6	0.5	75.17	75.38	68.23	66.31
		0.8	0.5	75.27	**75.67**	68.70	66.12
		1	0.5	75.41	75.53	68.72	66.52
		1.2	0.5	75.23	75.34	68.88	66.37
		1.4	0.5	75.13	75.36	**68.97**	66.02
		1	0.3	**75.60**	75.55	68.64	66.49
		1	0.4	75.40	75.64	68.56	**66.54**
		1	0.6	74.48	74.68	68.02	65.31

Image-Phrase Grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12476–12486.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*. 2425–2433.

[3] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. 2019. G3raphground: Graph-based language grounding. In *ICCV*. 4281–4290.

[4] Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *CVPR*. 4042–4050.

[5] Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-Guided Regression Network with Context Policy for Phrase Grounding. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 824–832.

[6] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020. Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 10083–10092.

[7] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual Grounding via Accumulated Attention. In *CVPR*. IEEE Computer Society, 7746–7755.

[8] Pelin Dogan, Leonid Sigal, and Markus H. Gross. 2019. Neural Sequential Phrase Grounding (SeqGROUND). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*. Computer Vision Foundation / IEEE, 4175–4184.

[9] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. 2020. Jointly Linking Visual and Textual Entity Mentions with Background Knowledge. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 264–276.

[10] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. 2020. On Visual-Textual-Knowledge Entity Linking. In *ICSC*. IEEE, 190–193.

[11] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. 2020. VTKEL: a resource for visual-textual-knowledge entity linking. In *ACM*. 2021–2028.

[12] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. 2018. Deep semantic-visual embedding with localization. In *RFIAP 2018-Congrès Reconnaissance des Formes, Image, Apprentissage et Perception*.

[13] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NeurIPS*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 2121–2129.

[14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *EMNLP*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 457–468.

[15] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive Learning for Weakly Supervised Phrase Grounding. In *ECCV (Lecture Notes in Computer Science)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.), Vol. 12348. Springer, 752–768.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *CVPR*. 4555–4564.

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*. 787–798.

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferIt Game: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.

[20] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[21] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399* (2014).

[22] Jianan Li, Yunchao Wei, Xiaodan Liang, Fang Zhao, Jianshu Li, Tingfa Xu, and Jiashi Feng. 2017. Deep Attribute-preserving Metric Learning for Natural Language Object Retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan (Eds.). ACM, 181–189.

Figure 4: This picture reports a qualitative example of our model on the Flickr30k test image id: 6059154572**. The ground truth bounding boxes associated with each query are reported in red. The complete sentence in input is reported at the bottom of the figure. All bounding boxes are predicted correctly.**

[23] Jiacheng Liu and Julia Hockenmaier. 2019. Phrase Grounding by Soft-Label Chain Conditional Random Field. In *EMNLP-IJCNLP*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 5111–5121.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *ECCV*. Springer, 21–37.

[25] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*. 1950–1959.

[26] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning Cross-Modal Context Graph for Visual Grounding. In *AAAI*. AAAI Press, 11645–11652.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 11–20.

[28] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2015. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In *ICLR*, Yoshua Bengio and Yann LeCun (Eds.).

[29] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*. 6087–6096.

[30] Bryan A. Plummer, Paige Kordas, M. Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. 2018. Conditional Image-Text Embedding Networks. In *Computer Vision - ECCV 2018 - 15th European Conference (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11216. Springer, 258–274.

[31] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In *ICCV*. IEEE Computer Society, 1946–1955.

[32] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*. IEEE Computer Society, 2641–2649.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*. 779–788.

[34] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99.

[36] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of Textual Phrases in Images by Reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9905. Springer, 817–834.

[37] Arka Sadhu, Kan Chen, and Ram Nevatia. 2019. Zero-shot grounding of objects from natural language queries. In *ICCV*. 4694–4703.

[38] Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *CVPR*. 4613–4621.

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

[40] Josiah Wang and Lucia Specia. 2019. Phrase Localization Without Paired Training Examples. In *ICCV*. IEEE, 4662–4671.

[41] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2019. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 394–407.

[42] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. IEEE Computer Society, 5005–5013.

[43] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. 2016. Structured Matching for Phrase Localization. In *Computer Vision - ECCV 2016 - 14th European Conference (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9912. Springer, 696–711.

[44] Fan Wu, Zhongwen Xu, and Yi Yang. 2017. An end-to-end approach to natural language object retrieval via context-aware deep reinforcement learning. *arXiv preprint arXiv:1703.07579* (2017).

[45] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*. 5945–5954.

[46] 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 *ICCV*. 4683–4693.

[47] Raymond A. Yeh, Jinjun Xiong, Wen-Mei W. Hwu, Minh N. Do, and Alexander G. Schwing. 2017. Interpretable and Globally Optimal Prediction for Textual Grounding using Image Concepts. In *NeurIPS*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 1912–1922.

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics* 2 (2014), 67–78.

[49] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking Diversified and Discriminative Proposal Generation for Visual Grounding. In *IJCAI*, Jérôme Lang (Ed.). ijcai.org, 1114–1120.

[50] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *CVPR*. 4158–4166.

[51] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. 2018. Weakly supervised phrase localization with multi-scale anchored transformer network. In *CVPR*. 5696–5705.

[52] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. 2020. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *arXiv preprint arXiv:2005.03572* (2020).

[53] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. 2020. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *CoRR* abs/2005.03572. arXiv:2005.03572

[54] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167* (2015).

[55] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*. Springer, 391–405.