What kinds of errors do reference resolution models make and what can we learn from them?

Jorge Sánchez[2], **Mauricio Mazuecos**[1,2], **Hernán Maina**[1,2] and **Luciana Benotti**[1,2]
[1]Universidad Nacional de Córdoba
[2]CONICET, Argentina
{mmazuecos,hernan.maina}@mi.unc.edu.ar
{jorge.sanchez,luciana.benotti}@unc.edu.ar

Abstract

Referring resolution is the task of identifying the referent of a natural language expression, for example "the woman behind the other woman getting a massage". In this paper we investigate which are the kinds of referring expressions on which current transformer based models fail. Motivated by this analysis we identify the weakening of the spatial natural constraints as one of its causes and propose a model that aims to restore it. We evaluate our proposed model on different datasets for the task showing improved performance on the most challenging kinds of referring expressions. Finally we present a thorough analysis of the kinds errors that are improved by the new model and those that are not and remain future challenges for the task.

1 Introduction

In the context of vision and language modelling, *reference resolution* can be understood as the task of identifying a region in an image referred by a natural language expression. This task is also known as *referring expression comprehension* (REC). It is closely related to the visual grounding problem in the sense that in both cases the goal is to identify the parts of the image that support or "ground" a given linguistic expression into the real world. The difference lies in that the referred region in REC is expected to be unique, while there may be multiple support regions in the case of visual grounding, *e.g.* the expression "a person" might refer to multiple instances of the person class, all of which can be seen as grounding candidates for the natural language phrase. Also note that, in the case of REC, the referred region may correspond to an actual physical object (or groups of objects) or to an abstract visual element that can be perceptually grouped into a meaningful entity, *e.g.* "the patch of grass at the bottom". REC aims at identifying a specific object or region unambiguously (Mao et al., 2016; Hu et al., 2016; Qiao et al., 2020).

Although the use of natural language queries to guide localization (Bansal et al., 2018; Rahman et al., 2018) has been explored in the computer vision literature in the past, it has mainly focused on the use of rather simple expressions involving class names and intrinsic visual attributes (*e.g.* colors). Humans use more complex expressions that include spatial and order relations, relative attributes, meronimy, etc. Compared to other recognition problems, REC goes beyond simple visual-linguistic matching and requires some (primitive, implicit) form of visual-linguistic reasoning, *e.g.* an expression like "the person to the left of the tree" requires that in order to locate the target object ("the person") one has to look at its relative position ("to the left of") with respect to a different element in the scene ("the tree"). REC and visual grounding are particularly relevant to other visual-linguistic problems like visual question answering (Antol et al., 2015) and visual dialog (Das et al., 2017a; De Vries et al., 2017), where being able to link different linguistic elements (words, phrases and syntactic relations) in a sentence to actual regions in an image helps establishing a "common ground" between the agents that take part in the communication process (Mazuecos et al., 2021).

Our research question is *what kinds of errors do reference resolution models make and what can we learn from them?*. This paper makes the following contributions[1].

- We propose a method for classifying referring expressions into linguistically motivated groups that allow for a disaggregated analysis.
- We identify expressions that define an spatial relation between two or more regions of the image as the main source of errors.
- We compare two strategies for accounting for the spatial dimension that improve on the state of the

[1]Code and models available at https://github.com/jadrs/rec

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1971 - 1986

art for REC on different datasets.

- We perform a systematic analysis of the errors and the strengths of the best proposed model.

The paper is structured as follows. In Section 2 we discuss related work from both psycholinguistics and machine learning, Section 3 digs into the particularities of the REC and visual grounding problems and analyzes the types of errors made by current models for the REC task. In Section 4 we propose a model that takes the observed errors into account and Section 5 evaluates the new model on different datasets. Section 6 presents a meticulous analysis of the errors that are improved by the new model and those that are not and remain future challenges.

2 Related work

The automatic processing of referring expressions (REs) has been studied for a long time (Winograd, 1972). Back then their semantic representation was a logical form and REs were classified into intrinsic (*e.g.* "*the big red car*") and relational (*e.g.* "*the car by the pedestrian*") corresponding to unary and binary logical predicates, respectively (Dale and Haddock, 1991).

In the psycholinguistics community, reference were studied as a collaborative process and focused on the construction of shared knowledge (Clark and Wilkes-Gibbs, 1986; Clark, 1996) through the use of language. Hawkins (1978) proposed a theory in which an speaker 1) introduces a referent, 2) collaborates with the hearer to locate the referred object in some shared set of objects and 3) refers to the totality of objects that satisfy the RE. Viewed as a collaborative process, reference becomes a fundamental phenomena underlying all kinds of grounded dialogs (Dale and Reiter, 1995; de Vries et al., 2017).

REs are a linguistically rich constructs. Phenomena like *overspecification*, that is when an RE has more attributes than actually needed, has been shown to help identification when the redundant attribute is easy to recognize (Paraboni et al., 2017). *Vagueness* and the use of *gradable attributes* emphasize the context dependance when using REs (Quirk et al., 1980; DeVault and Stone, 2004); something called "*big*" in one scene may be seen as "*small*" in another. REs with *syntactic ambiguities* were also shown to be used in human dialogs (Chantree et al., 2005; Khan et al., 2008). Viethen and Dale (2008) showed that even with fairly simple scenes human speakers frequently use *relational descriptions* to identify objects.

In this paper we adapt the classification of referring expressions (REs) proposed in (Krahmer and van Deemter, 2012) to the domain of 2D photos of the world. Krahmer and van Deemter typification includes the types: instrinsic (that they call unary predicates), relational (that they call binary predicates), set (that refer to a group of objects), and gradable (that we described above). Intrinsic and relational REs differ in the number of objects that are involved in the description. *Intrinsic* REs only involve the referent while *relational* REs use one or more additional objects to identify the target. Referring expressions that identify *sets* use properties that are shared by the elements in the set. For example the referents of *"the white cats"* are all cats and white. In general references to set use plural definite descriptions to identify them. The datasets used in this paper are supposed to refer to a single referent and not to a set of referents so we restrict our classification in this paper to singular REs. Finally, Krahmer and van Deemter last type corresponds to gradable REs. REs referring to objects in 2D photos of the world frequently use spatial properties that are *gradable*. For example, the attribute *to the right of* is gradable in that *"the empty sky to the right of the statue"* might refer to the sky touching the statue on the right or the sky further to the right. In this paper we restrict our analysis of gradable properties to spatial properties.

We agree with Cirik et al. (2018) that careful analysis of datasets and proposed models is crucial to make progress in REC. They performed an analysis of REC models by modifying or completely removing the REs and showed that the models could exploit biases on particular datasets to achieve competitive performance. In our work we do a disaggregated performance analysis and meticulous error analysis and not only rely on automatic performance metrics.

The introduction of the transformer architecture by Vaswani et al. (2017) enabled interesting and novel ways of fusing visual and linguistic information (Tan and Bansal, 2019; Lu et al., 2019, 2020). Besides differences on the way both modalities are merged (single- vs two-branch models, cross-modal attention layer design, etc.), these models also differ on the tasks they are (pre-)trained for, ranging from masked token prediction and text-

image matching (Li et al., 2020a; Sun et al., 2019) to more elaborated strategies such as label-region alignments (Li et al., 2020b; Guo et al., 2020) or scene-graph prediction (Yu et al., 2020). Most of these tasks can be seen as local (word-to-region) or global (text-to-image) matching tasks and, although effective for pre-training, they do not contemplate other phenomena such as composition and indirection that can be observed in grounding problems. Although large-scale pre-training has been shown to be effective for grounding (Kamath et al., 2021; Li and Sigal, 2021), such methods require the availability of large collections of image-text pairs with explicit alignments between image regions and phrases in text. Because of the scale of such data and the cost of extensive annotations, these approaches have not gone beyond image-text matching phenomena.

In this work, we take inspiration from a recently proposed family of transformer-based architectures that tackle the REC problem as a regression problem, *i.e*. given an image and a query expression, directly predict the bounding box coordinates of the object or image region referred by it (Deng et al., 2021; Du et al., 2021).

3 Problem definition and motivation

Given an image and a natural language expression, the goal of REC is to predict the location of the referred object or region, *e.g*. by predicting the coordinates of the bounding box that encloses it more tightly. The expression may include diverse linguistic constructs such as ellipsis, prepositional phrases, conjunctions, etc.

In what follows, we first introduce a simplified yet performant version of the model proposed by Deng et al. (2021). This model will serve as a strong baseline in our experiments because it allow us to evaluate the differences in performance between the baseline and the proposed extensions. Next, we propose a linguistically motivated classification scheme which allows for a *disaggregated analysis* that will guide the rest of the paper. Finally, we discuss some motivating results in light of our baseline and different types of expressions.

3.1 A strong baseline model

As in (Deng et al., 2021), the input to our model is an image I and a RE e. Its output are the bounding box coordinates of the referred object or region. The model consists on the following blocks: a image encoder, a language encoder, a transformer-based cross-modal encoder and a box prediction head. An outline of the architecture is shown in Fig. 1.

Image encoder. We feed the image to a pre-trained convolutional backbone. We use a ResNet-50 (He et al., 2016) pre-trained on ImageNet with the classification head removed. The output of this network is a feature map of size $H \times W$ and d channels, that we further project to D dimensions using 1×1 convolutions. Different from Deng et al. (2021), we do not add any additional transformer layer on top of the convolutional encoder. We flatten the resulting tensor along the spatial dimension and obtain a sequence of HW visual embeddings of dimensionality D.

Language encoder. We first map e into a sequence of T tokens and encode it using a pre-trained language model. We use a pre-trained BERT (Devlin et al., 2019) as the default encoder. The output sequence of (sub-) word embeddings is projected onto D dimensions using a fully-connected layer.

Cross-modal encoder. We concatenate both visual and language embeddings into a multimodal sequence and feed it to a cross-modal encoder consisting of a transformer architecture with L layers. Each layer in the encoder corresponds to a multi-head self-attention layer with skip connections (Vaswani et al., 2017). As in (Deng et al., 2021), we add learnable position embeddings to the input of each transformer encoder layer. The output of the cross-modal encoder is a sequence of embeddings of the same length as the input and whose elements can be seen as a re-encoding of the corresponding input embeddings. This re-encoding mechanism is based on the "similarity" between other elements of the same sequence (visual and/or linguistic) as induced by the multi-head self-attention layers within the cross-modal encoder.

Prediction head. The box prediction head is a single fully connected layer followed by a sigmoid. The output of this head corresponds to the normalized coordinates of the target bounding box. Instead of adding a specialized output token as in (Deng et al., 2021), we take as input the average of the first HW embeddings after the cross-modal encoder, *i.e*. those corresponding to the visual block. Given a training set of image-expression-box tu-

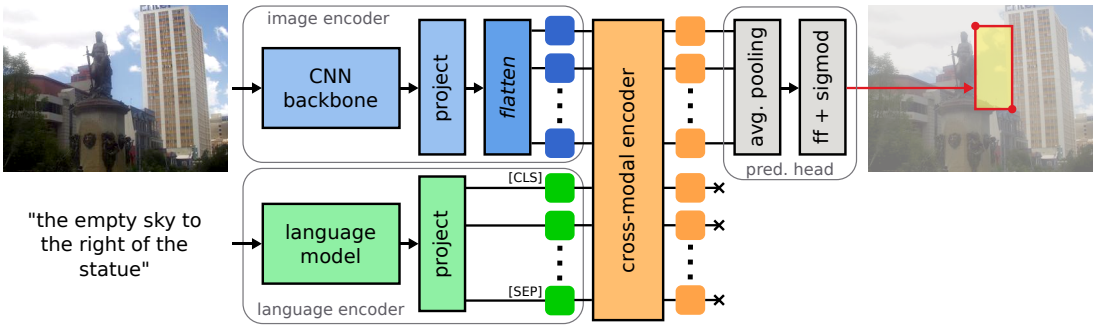


Figure 1: Our baseline model for referring expression comprehension showing its main processing blocks: image and language encoders (in blue and green respectively), cross-modal fusion (in orange) and box prediction head (in gray). The input is an image and a referring expression and the output is a bounding box.

ples $\mathcal{D} = \{(I_n, e_n, b_n), i = 1, ..., N\}$, we adopt a loss formulation based on a combination of the generalized IoU of Rezatofighi et al. (2019) and the soft L_1 loss as in Deng et al. (2021).

3.2 Types of referring expressions and errors

One of the challenges posed by the REC problem is that natural language expressions exhibit different degrees of grounding complexity. In order to better understand the limitations of current models, and motivated by previous work described in Section 2, we propose to classify them into a non-disjoint set of types as follows.

- *Spatial.* Expressions that include spatial language *e.g.* "the doorway on the far right". They include prepositions that signals a spatial relation (*e.g.* behind) or adjectives (*e.g.* left) and nouns (*e.g.* foreground) that give spatial cues.[2]
- *Ordinal.* Expressions that include ordinal adjectives that determine the position of an object inside a group, *e.g.* "2nd set of jewels".
- *Relational.* Expressions that use another object, which is related to the referent, in the RE; *e.g.* "the instrument right behind the guy with yellow hat". Relational expressions are spatial when the relation is spatial.[3]
- *Intrinsic*: Expressions that do not fall into any of the above types. They do not use the position of the object to describe it, instead they use only properties that are intrinsic to the referent no matter its position; *e.g.* "the tall metal fence".

[2]A complete list of spatial words used in this paper is in Appendix A.

[3]We say an expression is relational if it contains a preposition with one or more nouns to the left and to the right.

Type	# test	expr len	acc
All	65193	3.5 (2.6)	66.76
Intrinsic	22779	1.5 (1.0)	81.81
Spatial	42277	4.6 (2.6)	58.73
Ordinal	1173	5.9 (2.8)	28.47
Relational	13154	6.7 (3.0)	44.59

Table 1: Accuracy disaggregated by expression type for the RefItGame dataset. The expr len column shows the average expression length and its standard deviation (between parentheses). The last column of the table shows the accuracy for the baseline model on the ReferItGame test set. An RE may fall in more than one type, in that case it is counted in all the types.

Table 1 shows the frequency and lengths of the different types of REs in the ReferItGame (Kazemzadeh et al., 2014) dataset as well as the performance of the baseline model introduced in Sec. 3.1. The performance reported is the accuracy of the predicted bounding box. It is considered that the bounding box is correct if it overlaps with more than 50% of the ground truth one. The table shows that the accuracy on the spatial class is considerably lower than the accuracy for the intrinsic class although spatial expressions are more frequent. The accuracy is even lower for the relational and ordinal expressions. After performing this disaggregated analysis our intuition is that the position embeddings used in the state of the art models are not enough for capturing the spatial and relational information required to handle this type of expressions.

The Fig. 2 shows the cardinality of the intersections of the types as a matrix. Each row at the bottom corresponds to a type. Each column corresponds to a non empty set, and the bars at the top

show the size of the respective intersections. The filled dots show which type is part of an intersection. The first four columns represent those REs that fall in only one type. The last three columns show combinations of types.

The intrinsic type is disjoint from all others by definition. Almost all ordinal REs are also spatial or relational. There are very few relational REs that are not spatial (an example of such rare RE is *"the girl sleeping with the teddy bear"*). Most relational REs use an spatial preposition in the the ReferItGame dataset.

4 Restoring referential spatiality

The types of expressions that have the worst accuracy, relational and ordinal, can be associated to limitations of the model in capturing richer spatial relations that go beyond simple matching (names to image and absolute locations to image). In what follows, we propose two generic strategies to inject stronger visual priors into the model after the cross-modal fusion block. In Fig. 3 we illustrate how we reformulate the prediction head of the original model in Fig. 1. We describe the modified architecture below.

First, due to the one-to-one alignment of the embedding sequence at the input and output of the transformer, we can identify the first block of HW elements as a re-encoding of the original visual embeddings, modulated by the input expression e. We can rearrange this block as a tensor of size $H \times W \times D$ so as to restore the spatial structure lost after the flattening operation. This operation is

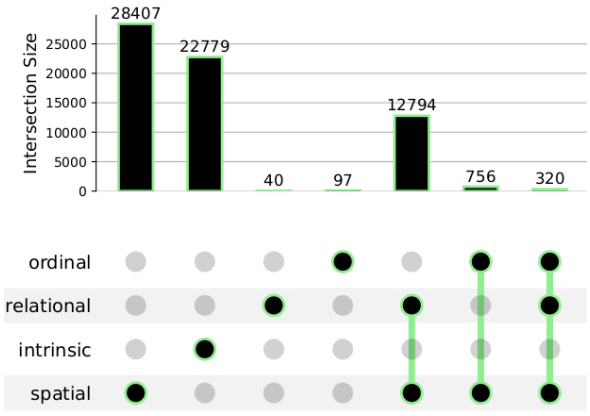


Figure 2: The RE types proposed in this paper are not disjoint. Their distribution and intersections on the ReferItGame dataset are shown in the columns. The first four columns correspond to REs that belong to only one type.

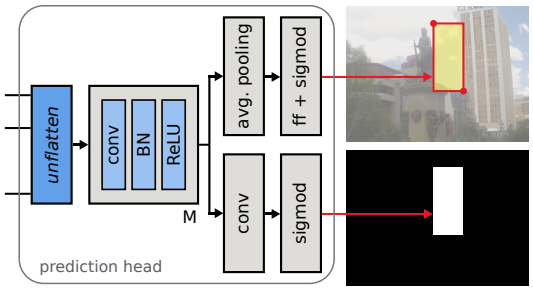


Figure 3: Prediction head for the extended model. This new prediction head replaces the one from Fig. 1 with an "unflatten" operation and a block of convolutions followed by the regression and segmentation heads. The regression head outputs a bounding box (in yellow) while the segmentation head outputs a mask (in black and white).

represented as the unflatten block in Fig. 3.

We next add a small convolutional network with M layers (convolution layer with a 3×3 kernel and stride of 1, batch normalization and a ReLU nonlinearity), on top of which we attach two different heads: a target regression head as before and a box segmentation head, as shown in Fig. 3.

The box regression head is the same as the base model but applied to the average pooled features after the convolutional block, while the additional segmentation head consists of a simple convolutional layer whose output is normalized to the interval $[0, 1]$ by a sigmoid activation. The goal of this head is to provide a spatial consistency constraint while avoiding the need of requiring additional data. Supervisory signal for this branch is obtained trivially from the existing annotations, *i.e.* by creating a binary mask from the bounding box coordinates of the target. Learning is formulated by adding an auxiliary pixel-wise binary classification term to the main loss.

The intuition behind adding M convolutional blocks after the cross-modal transformer as well as the mask prediction head is to ensure that the spatial consistency of the representation that has been broken is restored. The locality and hierarchical nature of the (stacked) convolutions as well as the complementarity of the segmentation and regression heads proves to be an effective approach to improve the accuracy of some kinds of spatial expressions as we will see in the next two sections.

5 Experiments

In the following we describe the datasets and experimental setup we use in our experiments. A

detailed description can be found in Appendix C.

5.1 Datasets

We conduct experiments on four different datasets.

ReferItGame. We use this dataset as a set for validation and error analysis. ReferItGame (Kazemzadeh et al., 2014) was collected based on a two-player game where one of the players has to write a RE based on a given object while the second player has to guess its identity by looking only at the image and the generated expression. If guessed correctly, both players receive a game point and swap their roles for the next image.

Images in this dataset are from the IAPR TC-12 corpus (Escalante et al., 2010). The dataset consists of 130121 expressions referring to 96654 different objects across 19894 images. We use a cleaned version of the dataset provided by Hu et al. (2016) and standard splits, *i.e*. 54127, 5842 and 60103 expressions for training, validation and testing, respectively.

RefCOCO and RefCOCO+. These datasets (Nagaraja et al., 2016) were collected by following the same procedure as in ReferItGame, but using images from MS COCO (Lin et al., 2014). RefCOCO contains 50000 referred objects in 19994 images. Each object is referred by an average of three expressions, for a total of 142210 REs split across 120624 expressions for training, 10834 for validation and two additional splits (testA and testB) with 5657 and 5095 expressions for testing, respectively. The testA split contains multiple people while the testB split contain multiple instances of all other objects.

RefCOCOg. RefCOCOg (Mao et al., 2016) was collected non-interactively using Amazon Mechanical Turk. This dataset contains 104560 expressions for 54822 objects in 26711 images. Compared with RefCOCO and RefCOCO+, expressions in the RefCOCOg dataset are considerably longer (an average of 8.43 vs. 3.61 and 3.53 words, respectively). We use the split proposed by (Nagaraja et al., 2016) for meaningful comparisons with other methods.

We report REC performance using average accuracy. We consider a region prediction as correct if it has an overlap (as measured by the IoU metric) of at least 0.5 with the ground truth box.

	SH	M=0	2	4	8
All	✗	71.66	72.64	73.00	73.46
	✓	71.89	73.03	73.16	73.61
Intrinsic	✗	84.76	84.63	84.93	84.10
	✓	83.79	84.93	84.14	84.27
Spatial	✗	64.28	65.91	66.53	67.36
	✓	65.18	66.27	66.74	67.73
Ordinal	✗	34.44	42.22	42.22	46.67
	✓	36.67	45.56	41.11	45.56
Relational	✗	51.29	51.70	52.70	53.57
	✓	51.74	52.99	52.99	54.11

Table 2: REC performance on the validation set of the ReferItGame datset for different expression types and model configurations. M denotes the number of convolutional blocks while "SH" denotes whether we use the segmentation head or not. M $= 0$, SH $=$ ✗ corresponds to the baseline model outlined in Sec. 3.1.

5.2 Architecture selection

In this section we evaluate the impact on performance of the changes proposed in Sec. 4 for the different types of REs discussed in Sec. 3.2. For these experiments, we set the maximum number of epochs to $E = 60$. Table 2 shows comprehension performance on the validation set of the ReferItGame dataset for different model configurations and expression types. M denotes the number of convolutional blocks and "SH" denotes whether we use the segmentation head or not. In this case, M $= 0$ and no segmentation head (SH=✗) corresponds to the baseline model outlined in Sec. 3.1. First, consider the case $M = 0$, *i.e*. no convolutional blocks after the cross-modal encoder. From the table, we see that adding a segmentation head improves performance overall, specially for the expressions that involve some degree of spatial reasoning. For the *intrinsic* type, improvement is marginal as the model is already able to solve the region-(class)target alignment problem. For more complex expressions, adding a segmentation head constraints the model to focus on the target location (and scale) and helps disambiguate references to objects/regions from its context. If we now consider $M > 1$, we observe the following. First, performance improves for all expression types, specially for *ordinal* and *relational* w.r.t. the model with $M = 0$. Second, for $M > 1$, adding a segmentation head seems to have no impact on performance. This can be attributed to a greater flexibility

of the stack of convolutions in capturing spatial and relational information.

5.3 Comparison with the state-of-the-art

Next, we compare the performance of our models against four different methods proposed recently in the literature, namely: LBYL-Net (Huang et al., 2021), VGTR (Du et al., 2021), TransVG (Deng et al., 2021) and the Referring Transformer (Ref. Tr.) (Li and Sigal, 2021). LBYL-Net is a one-stage grounding model based on modeling spatial relations between the referent and its context via a suitable convolution operator. VGTR is a transformer-based one-stage model following an encoder-decoder design and custom grounding modules. TransVG is similar to our baseline model with an additional transformer after the visual backbone and a specific output embedding that feeds the prediction head. Finally, the referring transformer model follows a similar design as DETR (Carion et al., 2020) while tackling simultaneously the RE comprehension and segmentation problems. We show performance for each model and relative improvement of the extended model with respect to the baseline in Table 3. All these models rely on a ResNet-101 as visual backbone, except LBYL which uses darknet-53.

In this paper, we presented two different models: a baseline described in Sec. 3 and a extended model that incorporates $M = 8$ convolutional layers and the segmentation head explained in Sec. 4. **We use a ResNet-50 as visual backbone and the same set of hyper-parameters and training procedure as before, explicitly avoiding dataset specific fine-tunings**. We disaggregate the performance according to the different types of expressions for both the baseline and the extended model.

First, we see that our baseline model is a strong baseline for REC. If we compare average performance (rows "All" in Table 3), we see that our baseline model performs comparably to the best performant methods in the first group. It shows the second best performance on both RefCOCO testA and testB subsets, second and third best performace in RefCOCO+ testA and testB, respectively; and achieves top performance on the RefCOCOg test set. If we consider the extended model, we observe a consistent overall improvement on all datasets. Although our goal is *not* to get the best possible performance but to highlight the importance of the different expression types when evaluating RE models, results in the table show that our design is on pair with the state of the art. This is important since both our models follow a simple design and rely on the same training protocol, which is compared to the more complex backbones and per-dataset tuning of hyperparameters of the methods in the first group.

If we compare disaggregated performance for the baseline and extended models, we observe the following. From Table 3, performance improves on all subsets and expression types, with the only exception of the *intrinsic* and *ordinal* types in the *testA* and *testB* subsets of RefCOCO and RefCOCO+ datasets, respectively. For RefCOCO, this accounts for a -0.2% decrease on performace w.r.t. to the baseline. For RefCOCO+, the difference is greater (-15% w.r.t. to the baseline). Note however that for this dataset, the number of *ordinal* expressions is 34, 4 and 32 for the *val*, *testA* and *testB* subsets. The observed decrease corresponds to a difference of only 2 examples.

In general, we observe a greater improvement for expressions that involve spatial (spatial and relational) and grouping (ordinal).

A detailed summary of these results, including performance on the validation sets and sample cardinalities for all subsets and expression types can be found in Appendix D.

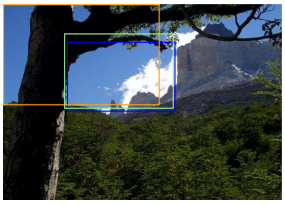
6 Error Analysis

In this section we analyze the predicted output of the extended and baseline models for the image-expression pairs of the ReferItGame validation set. The figures in this section show the extended model prediction in green and the baseline model prediction in orange. The comparative error analysis goal is to shed light over the kinds of linguistic and visual phenomena that the models are able to handle and those they are not. The analysis was divided into two parts and was carried out on a total of 351 examples. First, we explored referring expressions where the green model correctly predicted the ground truth and improved over the baseline depicted in orange. Two such examples are in Fig. 4. Second, we analysed cases where both models failed, exemplified in Fig. 5. More examples are shown in in Appendix B.

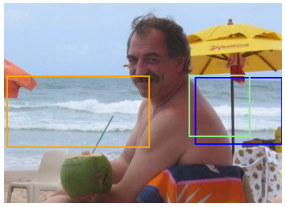
We identified different abilities a model should have in order correctly solve a broader set of expressions. They are listed below. Fig. 4 and Fig. 5 illustrate each of the skills.

Model	Type	RefCOCO		RefCOCO+		RefCOCOg
		testA	*testB*	*testA*	*testB*	*test*
LBYL	All	82.91	74.15	73.38	59.49	-
VGTR	All	82.32	73.78	70.09	56.61	67.23
TransVG	All	82.72	78.35	70.70	56.94	67.73
Ref. Tr	All	85.59	76.57	75.96	62.16	69.40
Baseline	All	84.85	74.72	75.95	59.36	69.40
(M=0, SH=✗)	Intrinsic	84.11	71.24	80.33	65.85	73.92
	Spatial	85.22	75.42	70.55	53.32	68.44
	Ordinal	75.91	48.67	50.00	40.62	46.25
	Relational	77.03	57.39	67.11	48.02	67.87
Extended	All	86.00 (+1.4%)	77.96 (+4.3%)	77.09 (+1.5%)	61.16 (+3.0%)	71.31 (+2.8%)
(M=8, SH=✓)	Intrinsic	83.94 (-0.2%)	72.32 (+1.5%)	81.18 (+1.1%)	68.25 (+3.6%)	75.25 (+1.8%)
	Spatial	86.96 (+2.0%)	79.06 (+4.8%)	72.14 (+2.3%)	54.62 (+2.4%)	70.51 (+3.0%)
	Ordinal	85.40 (+12.5%)	56.33 (+15.7%)	50.00 (+0.0%)	34.38 (-15.4%)	48.75 (+5.4%)
	Relational	79.56 (+3.3%)	60.22 (+4.9%)	68.50 (+2.1%)	49.80 (+3.7%)	70.09 (+3.3%)

Table 3: Comparison with other methods from the literature on the RefCOCO, RefCOCO+ and RefCOCOg datasets. For the baseline and the extended model, we consider performance for different expression types. Relative improvement percentage of the extended model with respect to the baseline are shown in parentheses.



(a) Sky near clouds



(b) Water near man's back

Figure 4: Examples where our proposed model (green) improves over the baseline (orange). The ground truth is shown as a blue box. The first example is classified as *fuzzy objects*. The second as *meronimy*.

(a) The empty *ski* to the right of the statue

(b) Grass directly above and behind little boy on bike

Figure 5: Examples that remain a challenge for both models (ours in green, baseline in orange, ground truth in blue). The first example is classified as *fuzzy objects*, *typo* and *directional*. The second is classified as *fuzzy objects*, *viewpoint*, and *implicit*.

- *Meronimy*: Reference parts of objects, such as peoples' clothes, parts of the body (*e.g. "the man's back"*) or parts of inanimate objects.
- *Viewpoint*: In order to resolve the reference, the hearer needs to assume the point of view of an object such as "*behind the little boy*".
- *Directional*: the expressions contains the direction in which we can find the referent relative to the location of the landmark, *e.g.* "*to the right of the statue*".
- *Fuzzy objects*: Reference to regions of objects without prototypical shapes and borders, like "*ground*", "*water*" or "*sky*".
- *Occlusion*: Reference to an object that is only partly visible because it is hidden by another object or does not fit completely in the image.
- *Typo*: the expression contains a typo that can confuse the understanding of the referring expression (e.g. "*ski*" instead of "*sky*").
- *Implicit*: the RE contains an ellipsis such as "*grass directly above*" meaning "*grass directly above the path*".

The first three types of errors require the RE to be relational. Meronimic errors require a relation between an object and a part of it. Viewpoint errors occur in REs that are not only relational, but also

spatial in general. They use a landmark, related to the referent, to change the point of view of the interpreter. Directional errors can occur when the direction of the relation in the RE is misinterpreted. Frequently this relation is spatial but not necessarily, it could correspond to an order established by other property, such as size (e.g. "*from the smallest to the biggest*). The last four types of errors may happen for all kinds of REs. Although implicit errors are more frequent in long REs such as relational REs.

6.1 Errors that are improved

During our analysis we identified that most of the cases in which the green model improves over the orange one contain *meronimy*, *fuzzy objects* or both. Fig. 4a is an example of *fuzzy object* because the sky is a kind of object that does not have a prototypical shape and borders (such as a person) and whose parts can be spread in different areas of the image. Our model is not only able to select a region near the clouds but it is also able to constrain its prediction using the natural boundary that is the tree on the left. The model is able to reconstruct the relative spatial position of the different parts of the sky.

Fig. 4b shows an example of *meronimy*. The "*man's back*" refers to a part of the man. Correctly predicting the referred region in this expression requires both identifying the back of the man (not the whole man) and the appropriate region of water (another instance of *fuzzy objects*). As the water and the sky, the man's back does not have a clear border wrt the rest of his body.

6.2 Errors that remain a challenge.

We analyzed 220 examples in which both models failed. The first observation that we found is that over 66% of the errors require more than one skill. This is only 38% for the 131 examples we annotated where the green model improves over the orange. Fig. 5 illustrates cases that require multiple skills, Fig. 5a not only has a typo (*ski* should be *sky*) but it also includes a directional relation (*to the right*). We can see how a tiny error in a character of the word *sky*, confuses the models and both predict the section of fences located below in the image, probably because they look like a set of *'ski'*s. Fig. 5b not only has fuzzy objects because of the grass but it also requires the interpretation of the viewpoint and implicit language for resolution. In order to identify the referent, the interpreter needs to take the viewpoint of the boy in the picture to correctly interpret the relation *behind* as referring to the blue box. The interpreter also needs to realize that the relation *above* implicitly means *above the bike path*.

Our second finding is that some kinds of meronimy remain a challenge and constitute 12% of the examples we annotated. We find that the most challenging meronimic relations are those that are not frequent in the training data (*e.g.* "*the eyebrows of the person*"). Similarly, some directional relations (that amount to 17% of the errors) are more challenging than others: those relations that are normally in the z-axis (*e.g.* "*behind*") lead to more errors that those in the y-axis (*e.g.* "*above*"). A complete distribution of the annotations can be found in Appendix B. Summing up, there is a lot of room for improving the grounding skills of REC models. We have identified that the main challenges present in the ReferItGame dataset are those that we defined as rare meronimic relations, viewpoint, directional, occlusion, implicit language and typos.

7 Discussion and conclusions

In this work we studied the kinds of errors that reference resolution models make. In particular, relational expressions caused a lot of errors in REC and that motivated the proposal of a model that improves over the previous SOTA for the task by restoring the referential spaciality. We used a new training objective which is segmentation prediction and added convolutional layers to a transformer.

Looking at accuracy only can obscure and hide common errors these model might have. We performed an error analysis and identified which skills the model would need in order to perform correctly. We found that our proposed model improves in dealing with fuzzy objects and meronimy, but still finds it difficult with other skills. By performing this error analysis we learnt there is yet a lot of work to do in making models consider viewpoint or being able to deal with implicit information (conveyed by the common visual context).

Our findings can help have a finer grained look at the predictions of models. This can be relevant for different areas of NLP like grounding, situated dialog systems and human-computer interaction as referring is a crucial skill in communication.

8 Ethical considerations

REC models predict bounding boxes instead of segmentation masks. Bounding boxes can include a lot of background information for some kinds of objects (imagine a broom at a 45 degrees angle). Such segmentation masks ground truths are expensive to annotate.

As in previous work, we count a prediction as correct if its IoU with the ground truth is above 50%. This binarization can obscure the quantitative analysis in border cases (49% vs 51% IoU).

Regarding datasets, we did not collect the datasets but used available ones for the task. Crowdsourcing raises ethical concerns including fair wage for crowdworkers, work load and exhaustion. In our qualitative analysis we could find examples of exhaustion in the linguistic production of crowdworkers.

Previous work, coming from the field of collaborative reference resolution, state that one of the "desired" applications was helping with surveillance systems (Li et al., 2017; Das et al., 2017b). We do not agree with this use of the technology. Despite every work in referring expressions inevitably helping towards that goal it would need retraining with specific domain data for that. Our proposed model and formulation is not aimed at surveillance nor the datasets used and should perform poorly in such setting.

We are reporting results over 30 experiments in total. Each experiment was running for 2.5 days on a single 1080ti GPU. We estimate $6.48 kgCO_2eq.$ for each experiment according to local emissions factor. Debugging, code refactoring and validation runs took around a couple of hundred additional runs.

Acknowledgment

We thank all anonymous reviewers and area chair for their insightful comments and suggestions. This work used computational resources from CCAD-UNC, which is part of SNCAD-MinCyT, Argentina.

References

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

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. 2018. Zero-shot object detection. In *ECCV*, pages 384–400.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

F. Chantree, A. Kilgarriff, A. De Roeck, and A. Willis. 2005. Disambiguating coordinations using word distribution information. In *Proceedings of Recent Advances in Natural Language Processing - 2005*.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirk. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.

Robert Dale and Nicholas Haddock. 1991. Generating referring expressions involving relations. In *Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany. Association for Computational Linguistics.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *CVPR*, pages 326–335.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017b. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, pages 5503–5512.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475. IEEE Computer Society.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. TransVG: End-to-end visual grounding with transformers. *arXiv preprint arXiv:2104.08541*.

David DeVault and Matthew Stone. 2004. Interpreting vague utterances in context. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1247–1253, Geneva, Switzerland. COLING.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2021. Visual grounding with transformers. *arXiv preprint arXiv:2105.04281*.

Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. 2010. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428.

Jia Guo, Chen Zhu, Yilun Zhao, Heda Wang, Yao Hu, Xiaofei He, and Deng Cai. 2020. Lamp: Label augmented multimodal pretraining. *arXiv preprint arXiv:2012.04446*.

J. Hawkins. 1978. *Definiteness and Indefiniteness: A Study in Reference and Grammaticality Prediction*. Routledge.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *ICCV*, pages 770–778.

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

Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. 2021. Look before you leap: Learning landmark features for one-stage visual grounding. In *CVPR*, pages 16888–16897.

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

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

Imtiaz Hussain Khan, Kees van Deemter, and Graeme Ritchie. 2008. Generation of referring expressions: Managing structural ambiguities. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 433–440, Manchester, UK. Coling 2008 Organizing Committee.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Muchen Li and Leonid Sigal. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *arXiv preprint arXiv:2106.03089*.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

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

Ken Litkowski. 2014. Pattern dictionary of english prepositions. In *ACL (1)*, pages 1274–1283.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Mauricio Mazuecos, Franco M Luque, Jorge Sánchez, Hernán Maina, Thomas Vadora, and Luciana Benotti. 2021. Region under discussion for visual dialog. In *EMNLP*, pages 4745–4759.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer.

Ivandré Paraboni, Alex Gwo Jen Lan, Matheus Mendes de Sant'Ana, and Flávio Luiz Coutinho. 2017. Effects of Cognitive Effort on the Resolution of Overspecified Descriptions. *Computational Linguistics*, 43(2):451–459.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE TMM*.

Radolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Startvik. 1980. *A grammar of contemporary english (9th Edition)*. Longman.

Shafin Rahman, Salman Khan, and Fatih Porikli. 2018. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, pages 547–563. Springer.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, Inc., USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12.

A Spatial prepositions and keywords

For the spatial prepositions, we rely on the Pattern Dictionary of English Prepositions (PDEP) (Litkowski, 2014), a publicly available lexical resource collected as a part of The Preposition Project (TPP). There are 78 prepositions in the *spatial* class. We removed prepositions with less than 10 samples, archaic and/or literary (*e.g.* 'pon, betwixt) and those used in a more technical context (*e.g.* aslant). The final list of prepositions and spatial keywords is as follows:

- Prepositions: *aboard, about, above, across, after, against, ahead of, all over, along, alongside, amid, among, around, as far as, at, atop, before, behind, below, beneath, beside, besides, between, beyond, by, by way of, down, for, from, in, in front of, in line with, in sight of, in the midst of, inside, inside of, into, near, near to, neath, of, off, on, on a level with, on top of, onto, opposite, out of, outboard of, outside, outside of, outwith, over, over against, past, round about, short of, this side of, through, throughout, to, toward, towards, under, underneath, unto, up, up against, up*

and down, up before, up to, upon, with, within, within sight of.

- Keywords: *background, back, bottom, center, corner, close, edge, end, entire, facing, far, farthest, floor, foreground, front, furthest, frontmost, ground, hidden, leftmost, left, middle, nearest, part, rightmost, right, row, side, top, upper.*

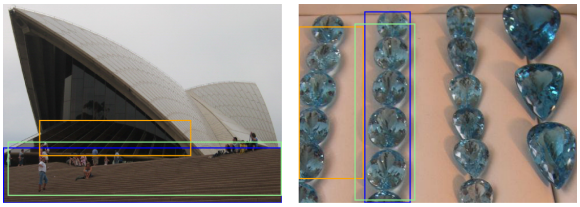
B More error analysis

Here we present more examples from the analysis to further explain the type of error these models perform and the skills they need to improve.

B.1 Errors that are improved

As mentioned in Sec. 6, the cases where the green model is better than the orange baseline, mostly occur when meronimy skills, fuzzy objects, or both are present. In Fig. 6a, we see how the green model, correctly predicts the set of steps under the opera house although this referent does not have clear borders.

In the Fig. 6b, the baseline orange model shows an error when dealing with ordinal referring expressions.



(a) Steps to the opera house. (b) 2nd set of jewels.

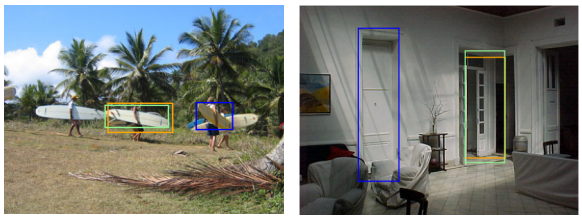
Figure 6: Errors that are improved.

The barplot in Fig. 9, shows the distribution of skill combinations found during the error analysis described in Section 6. The vertical axis details, in decreasing order, the combinations of skills that had the greatest impact in relation to the improvements observed in the green model over the baseline model. The x-axis expresses the frequency of occurrences of each of the combinations assigned to each example, in relation to the total of samples annotated.

B.2 Errors that remain a challenge

The images presented in Fig. 7, show two examples that still represent a challenge for the models.

Fig. 7b shows a frequent error found during the analysis. In this case, we observe that the models



(a) The surfboard of the blond surfboarder, the one walking towards the line. (b) Door closest to right of painting.

Figure 7: Error that remain a challenge.

detect objects further to the right than the actual referent. One possible hypothesis is that the models are not able to understand the implicit proximity that the speaker is trying to communicate. When an expression of the form "X is to the right of Y" is given, a speaker implies that he is speaking of the closest object "to the right" of the landmark.

Interpreting Fig. 7a involves identifying the imaginary line formed by the three surfers, recognizing the blond one, and inferring the direction of his walk, which is referenced by the directional preposition 'towards'.

In a similar way as shown in Fig. 9, the barplot in Fig. 10 describes the distribution of the 15 most frequent skills combinations that made both models fail.

B.3 Errors that are not real errors

Fig. 8 shows how some predictions delivered by the models, despite not coinciding with the ground truth (depicted in blue), can in fact be correct.

In particular, Fig. 8a presents an ambiguous RE. Both the ground truth and the prediction by the green model are possible interpretations. The model identifies the face of the woman that is to the left and outside the group of the three women, which is a valid interpretation given the ambiguity of the input referential expression.

For Fig. 8b, the scenario is totally different. Here

(a) Face of woman, on the left of the group of 3. (b) Donald duck above girl's head.

Figure 8: Error that are not real errors.

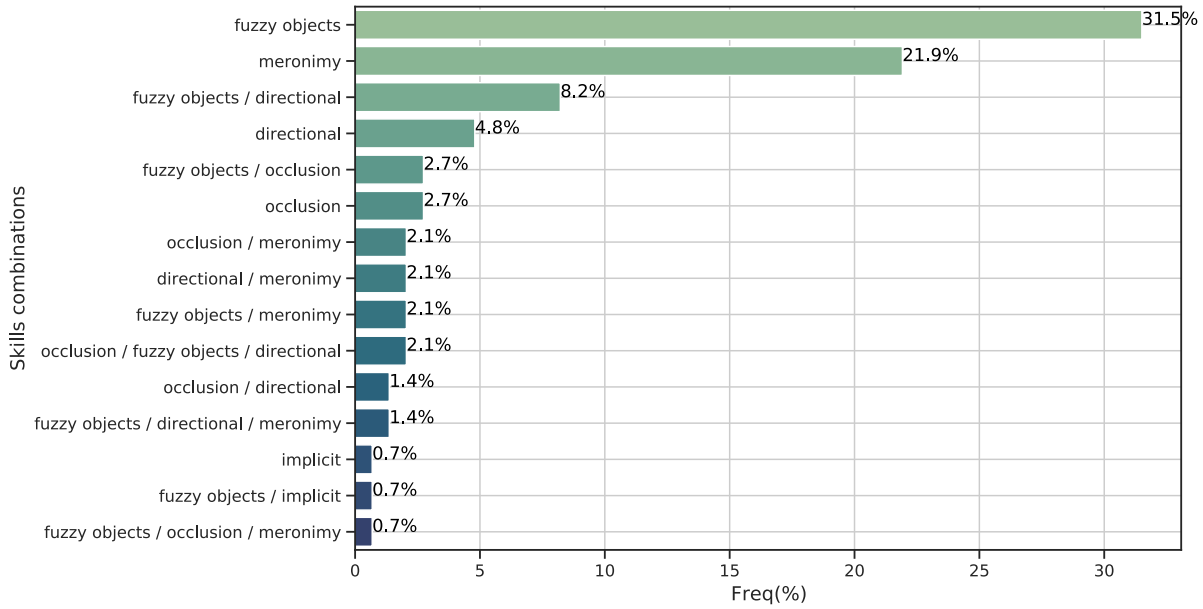


Figure 9: Distribution of the 15 most frequent skills combinations where the extended model improves over the baseline model.

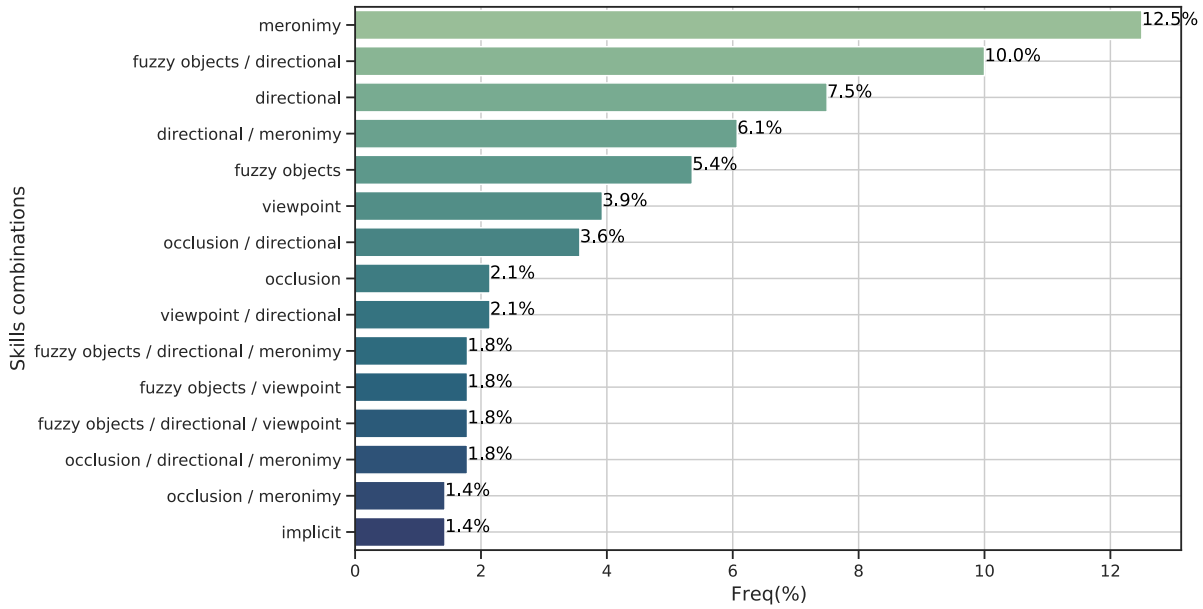


Figure 10: Distribution of the 15 most frequent skill combinations where both the extended and the baseline model fail.

the error is more related to the way in which we consider a prediction to be correct, than to a mistake made by the model. Strictly speaking, it is due to the fact that the predicted area is much smaller than that indicated as ground truth, giving an IoU $<<$ 50%; note that this decision method is mentioned as one of the limitations found in Sec. 8. The model ends up adjusting the *Donald duck* sticker more tightly than what is annotated as ground truth.

In order for our results to be comparable to previous work we did not modify the ground truth boxes.

C Detailed experimental setup

Images are first normalized by the mean and standard deviation of rgb values pre-computed on the ImageNet training set. We resize the images to 512×512 pixels while keeping the original aspect ratio by fixing the longest side to 512 and zero-

padding the shortest side accordingly.

For the visual encoder we use the pre-trained ResNet-50 (He et al., 2016) model available at the `torchvision` package from the PyTorch library (Paszke et al., 2019). This model has been trained for a 1000-way classification task on the ImageNet 2012 dataset (Russakovsky et al., 2015). We replace the output classification layer by a convolutional + normalization layer with $D = 256$ output channels. In this case, given an input image of 512×512 pixels, we obtain an output tensor of size $16 \times 16 \times 256$. We freeze the first convolutional layer of the network as well as the batch normalization layers. We apply random affine transformations (rotation, translation and scale) as the *only* augmentation strategy during training.

For the language encoder we use the *bert-base-uncased* pre-trained BERT (Devlin et al., 2019) model from the HuggingFace's Transformers library (Wolf et al., 2020). As with the visual encoder, we add a projection and normalization layer to project the embeddings output by the model to $D = 256$ dimensions. We set a maximum expression length to 32 input tokens.

Our loss function takes the form:

$$\mathcal{L}_{\text{Soft-}L_1} + \gamma \mathcal{L}_{\text{GIoU}} + \mu \mathcal{L}_{\text{segm}}$$

where the first and second terms act on the output cast by the box regression head while the third on the mask predicted by the segmentation head. We use $\gamma = 0.1$ in all our experiments. We set $\mu = 0$ for the baseline model and $\mu = 0.1$ for the extended one.

We train our models for a maximum of $E = 90$ epochs using the AdamW (Loshchilov and Hutter, 2018) optimizer with a multi-step decay schedule by a factor of 0.1 at the $\lfloor 0.6E \rfloor$ and $\lfloor 0.9E \rfloor$ epochs. Learning rate is set to 1×10^{-4} for the whole model except for the visual and language backbones (ResNet and BERT) for which we use 1×10^{-5}. Final models are chosen based on validation accuracy.

D Additional experimental results

Table 4 show an extended view of the results presented in Table 3, including recognition performance on the validation subsets for all the datasets and the sample cardinality for the different expression types considered in the paper.

Model	Type	RefCOCO			RefCOCO+			RefCOCOg		ReferItGame	
		val	testA	testB	val	testA	testB	val	test	val	test
LBYL (Huang et al., 2021)	All	79.67	82.91	74.15	68.64	73.38	59.49	-	-	-	67.47
VGTR (Du et al., 2021)	All	79.20	82.32	73.78	63.91	70.09	56.61	65.73	67.23	-	-
TransVG (Deng et al., 2021)	All	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	-	70.73
Ref. Tr (Li and Sigal, 2021)	All	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40	-	71.42
Baseline	All	81.07	84.85	74.72	69.93	75.95	59.36	70.14	69.40	69.62	66.76
(M=0, SH=✗)	Intrinsic	80.52	84.11	71.24	76.37	80.33	65.85	74.10	73.92	81.52	81.79
	Spatial	81.33	85.22	75.42	62.95	70.55	53.32	69.47	68.44	62.92	58.73
	Ordinal	64.49	75.91	48.67	35.29	50.00	40.62	56.52	46.25	34.44	28.47
	Relational	70.85	77.03	57.39	60.56	67.11	48.02	68.81	67.87	48.51	44.61
Extended	All	82.83 (+2.2%)	86.00 (+1.4%)	77.96 (+4.3%)	70.72 (+1.1%)	77.09 (+1.5%)	61.16 (+3.0%)	71.90 (+2.5%)	71.31 (+2.8%)	74.70 (+7.3%)	70.92 (+6.2%)
(M=8, SH=✓)	Intrinsic	82.01 (+1.9%)	83.94 (-0.2%)	72.32 (+1.5%)	77.34 (+1.3%)	81.18 (+1.1%)	68.25 (+3.6%)	76.53 (+3.3%)	75.25 (+1.8%)	85.15 (+4.5%)	83.94 (+2.6%)
	Spatial	83.14 (+2.2%)	86.96 (+2.0%)	79.06 (+4.8%)	63.51 (+0.9%)	72.14 (+2.3%)	54.62 (+2.4%)	71.01 (+2.2%)	70.51 (+3.0%)	68.79 (+9.3%)	64.01 (+9.0%)
	Ordinal	64.98 (+0.8%)	85.40 (+12.5%)	56.33 (+15.7%)	44.12 (+25.0%)	50.00 (+0.0%)	34.38 (-15.4%)	60.87 (+7.7%)	48.75 (+5.4%)	42.22 (+22.6%)	35.46 (+24.6%)
	Relational	72.87 (+2.9%)	79.56 (+3.3%)	60.22 (+4.9%)	60.95 (+0.6%)	68.50 (+2.1%)	49.80 (+3.7%)	70.46 (+2.4%)	70.09 (+3.3%)	56.18 (+15.8%)	51.49 (+15.4%)
Sample size	All	10834	5657	5095	10758	5726	4889	4896	9602	6333	65193
	Intrinsic	2613	1737	831	5657	3182	2378	865	1802	2290	22779
	Spatial	8180	3903	4236	5031	2523	2470	4012	7763	4028	42277
	Ordinal	414	137	300	34	4	32	46	80	90	1173
	Relational	2031	1306	744	3362	1800	1464	3456	6651	1205	13154

Table 4: Comparison with other methods from the literature on the RefCOCO, RefCOCO+, RefCOCOg and ReferItGame datasets. For the baseline and the extended model, we consider performance for different expression types. Relative improvements of the extended vs. the baseline model (in %) are shown between parentheses. Last block shows sample cardinality for each expression type and data subset.