Talk2Car: Taking Control of Your Self-Driving Car

Thierry Deruyttere[1]   **Simon Vandenhende[2]**   **Dusan Grujicic[2]**
Luc Van Gool[2]   **Marie-Francine Moens[1]**
[1]Department of Computer Science (CS)
[2]Department of Electrical Engineering (ESAT)
KU Leuven
{thierry.deruyttere, sien.moens}@cs.kuleuven.be
{simon.vandenhende, dusan.grujicic, luc.vangool}@esat.kuleuven.be

Abstract

A long-term goal of artificial intelligence is to have an agent execute commands communicated through natural language. In many cases the commands are grounded in a visual environment shared by the human who gives the command and the agent. Execution of the command then requires mapping the command into the physical visual space, after which the appropriate action can be taken. In this paper we consider the former. Or more specifically, we consider the problem in an autonomous driving setting, where a passenger requests an action that can be associated with an object found in a street scene. Our work presents the `Talk2Car` dataset, which is the first object referral dataset that contains commands written in natural language for self-driving cars. We provide a detailed comparison with related datasets such as `ReferIt`, `RefCOCO`, `RefCOCO+`, `RefCOCOg`, `Cityscape-Ref` and `CLEVR-Ref`. Additionally, we include a performance analysis using strong state-of-the-art models. The results show that the proposed object referral task is a challenging one for which the models show promising results but still require additional research in natural language processing, computer vision and the intersection of these fields. The dataset can be found on our website: http://macchina-ai.eu/

1 Introduction

Researchers have studied the problem of understanding actions communicated through natural language in both simulated (Das et al., 2017; Gordon et al., 2017; Hermann et al., 2017) and real environments (Loghmani et al., 2018; de Vries et al., 2018; Anderson et al., 2017). This paper focuses on the latter. More concretely, we consider the problem in an autonomous driving setting, where a passenger can control the actions of an **A**utonomous **V**ehicle (AV) by giving natural language commands. We hereunder argue why this problem setting is particularly interesting.

First, a recent study by Richardson and Davies (2018) has shown that the majority of the public is reluctant to step inside an AV. A possible explanation for this might be the lack of control which can be unsettling to some. Providing a way to communicate with the vehicle could help alleviate this uneasiness. Second, an AV can become hesitant in some situations (Robitzski, 2019). By giving a task or command, the passenger could guide the agent in its decision process. Third, some situations request feedback. For example, a passenger might indicate that they want to park in the shade during a sunny day. Finally, the problem of urban scene understanding is one of practical relevance that has been well studied (Cordts et al., 2016; Geiger et al., 2013). We believe all of this makes it an interesting setting to assess the performance of grounding natural language commands into the visual space.

To perform the requested action, an agent is required to take two steps. First, the agent needs to interpret the command and ground it into the physical visual space. Secondly, the agent has to devise a plan to execute the given command. This paper focuses on this former step, or more concretely: given an image I and a command C, the goal is to find the region R in the image I that the command is referring to. In this paper, to reduce the complexity of the object referral task we restrict the task to the case where there is only one targeted object that is referred to in the natural language command.

To stimulate research on grounding commands into the visual space we present the first object referral dataset, named `Talk2Car`, that comes with commands formulated in textual natural language for self-driving cars. A few example commands

(a) You can park up ahead behind **the silver car**, next to that lamppost with the orange sign on it
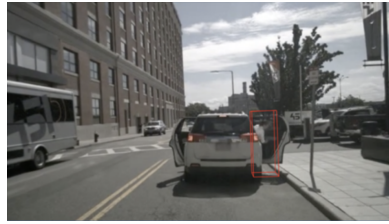
(b) **My friend** is getting out of the car. That means we arrived at our destination! Stop and let me out too!

(c) Yeah that would be **my son** on the stairs next to the bus. Pick him up please

(d) After **that man in the blue top** has passed, turn left

(e) There's **my mum**, on the right! The one walking closest to us. Park near **her**, she might want a lift

(f) Turn around and park in front of **that vehicle in the shade**

Figure 1: The `Talk2Car` dataset adds textual annotations on top of the `nuScenes` dataset for urban scene understanding. The textual annotations are free form commands, which guide the path of an autonomous vehicle in the scene. Each command describes a change of direction, relevant to a referred object found in the scene (here indicated by the red 3D-bounding box). Best seen in color.

together with their contextual images can be found in Fig. 1. Moreover, by using this new dataset we evaluate the performance of several strong state-of-the-art models that recognize the referred object of a command in the visual scene. Here we encounter several challenges. Referred objects are sometimes ambiguous (e.g., there are several cyclists in the scene), but can be disambiguated by understanding modifier expressions in language (e.g., the biker with the red jacket). These modifier expressions could also indicate spatial information. Furthermore, detecting the targeted object is challenging both in the language utterance and the urban scene, for instance, when dealing with complex and long sentences which might contain coreferent phrases, and with distant objects in the visual scene, respectively. Finally, in AV settings the speed of predicting the location of the referred object is of primordial importance.

The contributions of our work are the following:

- We propose the first object referral dataset for grounding commands for self-driving cars in free natural language into the visual context of a city environment.

- We evaluate several state-of-the-art models that recognize the referred object of a natural language command in the urban scene.

- We especially evaluate the models 1) for their capabilities to disambiguate objects based on modifying and spatial relationships expressed in language; 2) for their capabilities to cope with difficult language and visual context; and 3) with respect to prediction speed, which is important in real-life AV settings.

2 Related Work

Object Referral The `Talk2Car` dataset considers the object referral task, which requires to retrieve the correct object (region) from an image based on a language expression. A common method is to first extract regions of interest from the image, using a region proposal network (RPN). Yu et al. (2016); Mao et al. (2016) decode these proposals as a caption using a recurrent neural network (RNN). The predicted region corresponds to the caption that is ranked most similar to the referring expression. Other works based on RPN (Hu et al., 2017) or Faster-RCNN (Yu et al., 2018) have integrated attention mechanisms to decompose the language expressions into multiple sub-parts but use tailored modules for specific sub-tasks making

them less fit for our object referral task. Karpathy et al. (2014) interpret the inner product between region proposals and sentence fragments as a similarity score, allowing to match them in a bidirectional manner. Hu et al. (2016) uses an encoding of the global context in addition to the local context from the extracted regions. Hu et al. (2018) explore the use of modular networks for this task. They are comprised of multiple smaller predefined building blocks that can be combined together based on the language expression. The last three state-of-the-art models are evaluated on `Talk2Car` (section 5).

Grounding in Human-Robot Interaction When giving commands to robots, the grounding of the command in the visual environment is an essential task. Deits et al. (2013) use Generalized Grounding Graphs (G^3) (Tellex et al., 2011; Kollar et al., 2013) which is a probabilistic graphical model based on the compositional and hierarchical structure of a natural language command. This approach allows to ground certain parts of an image with linguistic constituents. Shridhar and Hsu (2018) consider the task where a robot arm has to pick up a certain object based on a given command. This is accomplished by creating captions for extracted regions from a RPN and clustering them together with the original command. If the command is ambiguous and more than one caption indicates the referring expression, the system will ask a clarifying question in order to be able to pick the right object. Due to its computational complexity during prediction, we did not select the last model in our evaluations.

Visual Question Answering The goal of VQA is to ask any type of question about an image for which the system should return the correct answer. This requires the system to have a good understanding of the image and the question. Early work (Kafle and Kanan, 2016; Zhou et al., 2015; Fukui et al., 2016) tried to solve the task by fusing image features extracted by a convolutional neural network (CNN), together with an encoding of the question. (Johnson et al., 2017; Suarez et al., 2018) experimented with modular networks for this task. Hudson and Manning (2018) proposed the use of a network made of recurrent Memory, Attention and Composition (MAC) cells. Similar to modular networks, the MAC model also

uses multiple reasoning steps, making it a suitable model in our evaluation (section 5).

Object Referral Datasets Over the years, various object referral datasets based on both real world and computer generated images have been proposed. Kazemzadeh et al. (2014) introduced the first real-world large-scale object referral dataset named `ReferIt`. Yu et al. (2016) constructed `RefCOCO` and `RefCOCO+` and as the names suggest, these two datasets are based on the `MSCOCO` dataset (Lin et al., 2014). A third dataset also based on `MSCOCO`, named `RefCOCOg` (Mao et al., 2016), contains longer language expressions than the previous two datasets. This dataset has an average expression length of 8.43 words per expression compared to 3.61 for `RefCOCO` and 3.53 for `RefCOCO+`. The dataset closest to ours is the dataset by Vasudevan et al. (2018), as it augments `Cityscapes` (Cordts et al., 2016) with textual annotations. We will henceforth refer to this dataset as `Cityscapes-Ref`. The main difference with this work, is that the `Talk2Car` dataset contains commands, rather than descriptions. A computer generated dataset named `CLEVR-Ref` was proposed by Hu et al. (2018) which has been created by augmenting the `CLEVR` dataset (Johnson et al., 2016) such that it would include referred objects. We refer to section 4.1 for a thorough comparison between these datasets and ours. Chen et al. (2019) recently released the Touchdown dataset where the goal is to navigate through a city based on a textual itinerary and locate a target (called Touchdown) at the final destination. This can also be seen as an object referral task with the addition of following an itinerary. While being a very interesting dataset and problem, it differs from the task being evaluated in this paper

3 Dataset

3.1 Dataset Collection and Annotation

The `Talk2Car` dataset is built upon the `nuScenes` dataset (Caesar et al., 2019) which is a large-scale dataset for autonomous driving. The `nuScenes` dataset contains 1000 videos of 20 seconds each taken in different cities (Boston and Singapore), weather conditions (rain and sun) and different times of day (night and day). These videos account for a total of approximately 1.4 million images. Each scene comes with data from six cameras placed at different angles on the car,

LIDAR, GPS, IMU, RADAR and 3D bounding box annotations. The 3D bounding boxes discriminate between 23 different object classes.

We relied on workers of Amazon Mechanical Turk (AMT) to extend the videos from the `nuScenes` dataset with written commands. To create commands, each worker watches an entire 20 second long video from the front facing camera. Afterwards, the worker navigates to any point in the video that is found interesting. Once the worker has decided on the frame, a pre-annotated object from the `nuScenes` dataset for that frame needs to be selected. The annotation task is completed when a command referring to the selected object is entered. The workers were free to enter any command, as long as the car can follow a path based on the command.*

We hired five workers per video who could enter as many commands per video frame as they wanted. To ensure high quality annotations, we manually verified the correctness of all the commands and corresponding bounding boxes. The verification happened in a two-round system, where each annotation had to be qualified as adequate by two different reviewers. To incentivize workers to come up with diverse and meaningful commands, we awarded a bonus every time their work received approval.

3.2 Statistics of the Dataset

The `Talk2Car` dataset contains 11 959 commands for the 850 videos of the `nuScenes` training set as 3D bounding box annotations for the test set of the latter dataset are not disclosed. 55.94% and 44.06% of these commands belong to videos taken respectively in Boston and Singapore. On average a command consist of 11.01 words, 2.32 nouns, 2.29 verbs and 0.62 adjectives. Each video has on average 14.07 commands. In Fig. 2(d) we can see the distribution of distance to the referred objects. Fig. 2(b) displays the distribution of commands over the videos. On average there are 4.27 objects with the same category as the referred object per image and on average there are 10.70 objects per image. Fig. 2(a) shows a heatmap of the location of all referred objects in the images of `Talk2Car`. In Fig. 2(c) we see the distribution of commands that refer to an object of a certain category.

---

*The path to be followed by the car when executing the command, will be added in a later version of the dataset.

3.3 Dataset Splits

We have split the dataset in such a way that the train, validation and test set would contain 70%, 10% and 20% of the samples, respectively. To ensure a proper coverage of the data distribution in each set, we have taken a number of constraints into account. First, samples belonging to the same video are part of the same set. Second, as the videos are shot in either Singapore or Boston, i.e., in left or right hand traffic, the distributions of every split have to reflect this. Third, we aim to have a similar distribution of scene conditions across different sets, such as the type of weather and the time of day. Finally, as the number of occurrences of object categories is heavily imbalanced (see fig. 2(c)), we have ensured that every object category contained in the test set is also present in the training set. With these constraints in mind we randomly sample the three sets for 10 000 times and optimize for a data distribution of 70%, 10% and 20%. The resulting train, validation and test sets contain 8 349 (69.8%), 1 163 (9.7%) and 2 447 (20.4%) commands respectively. We have also identified multiple subsets of the test set, which allow evaluation of specific situations. When the referred object category occurs multiple times in an image, attributes in modifying expressions in language including spatial expressions might disambiguate the referred object. This has led to test sets with different numbers of occurrences of the targeted category of the referred object. Longer commands might contain irrelevant information or might be more complex to understand, leading to test sets with commands of different length. Finally, referred objects at large distances from the AV might be difficult to recognize. Hence, we have built test sets that contain referred objects at different distances from the AV.

4 Datasets Comparisons

4.1 Quantitative Evaluation of **Talk2Car**

Table 1 compares the `Talk2Car` dataset with prior object referral datasets. It can be seen that `Talk2Car` contains fewer natural language expressions than the others. However, although the dataset is smaller, the expressions are of high quality thanks to the double review system we discussed earlier in section 3.1. The main reason for having fewer annotations is that the original `nuScenes` dataset only discriminates between 23 different categories corresponding to the an-
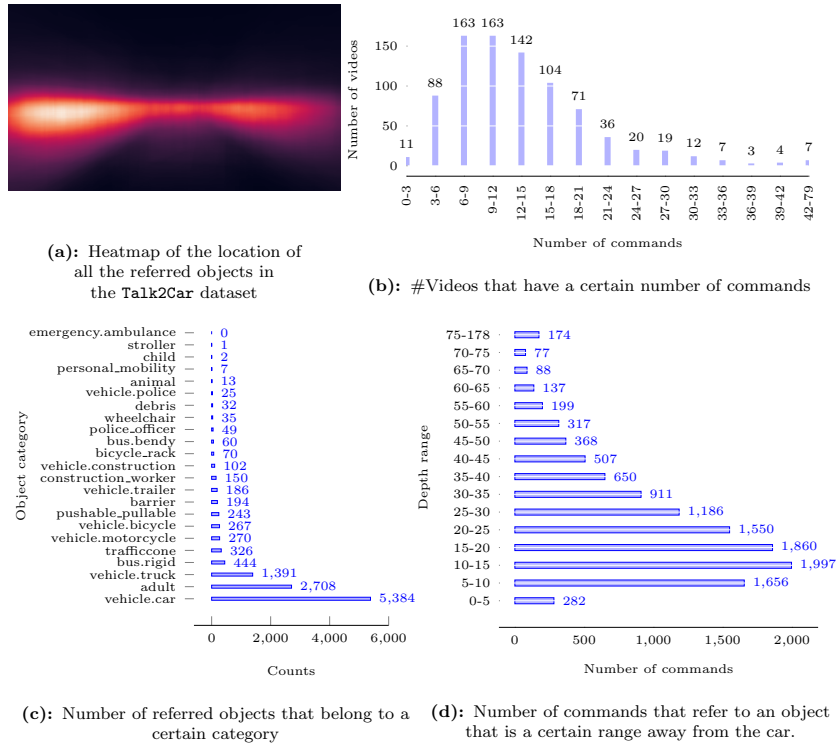
(a): Heatmap of the location of all the referred objects in the `Talk2Car` dataset

(b): #Videos that have a certain number of commands

(c): Number of referred objects that belong to a certain category

(d): Number of commands that refer to an object that is a certain range away from the car.

Figure 2: Statistics of the `Talk2Car` dataset.

notated bounding boxes. Moreover, the original `nuScenes` dataset considers the specific setting of urban scene understanding. This limits the visual domain considerably in comparison to `MS-COCO`. On the other hand, `Talk2Car` contains images in realistic settings accompanied by free language in contrast to curated datasets such as MS-COCO. Compared to most of the above datasets, the video frames annotated with natural language commands are part of larger videos that contain in total 1 183 790 images which could be exploited in the object referral task.

When we consider the average length of the natural language expressions in `Talk2Car`, we find that it ranks third, after `CLEVR-Ref` and `Cityscapes-Ref`. We did not put limitations on what the language commands could contain, which benefits the complexity and linguistic diversity of the expressions (section 4.2).

When looking at the type of modalities, the `Talk2Car` dataset considers RADAR, LIDAR and video. These modalities are missing in prior work except for video in `Cityscapes-Ref`. Including various modalities allows researchers to study a very broad range of topics with just a single dataset.

4.2 Qualitative Evaluation of `Talk2Car`

To make our discussion more concrete, we compare the textual annotations from Fig. 1 with some examples from prior work that are listed below. `RefCOCO` contains expressions such as 'Woman on right in white shirt' or 'Woman on right'. `RefCOCO+` on the other hand contains expressions such as 'Guy in yellow dribbling ball' or 'Yellow shirt in focus'. Lastly, `ReferIt` contains 'Right rocks', 'Rocks along the right side'. The language used in the above prior work is more simple, explicit and is well structured in comparison to the commands of `Talk2Car`. Additionally, the latter tend to include irrelevant side-information, e.g., 'She might want a lift', instead of being merely descriptive. The unconstrained free language of `Talk2Car` introduces different challenges, which involve co-reference resolution, named entity recognition, understanding relationships between objects, linking attributes to objects and understanding which object is the object of interest in the command.

The commands also contain implicit referrals as can be seen in the command in Fig. 1(f): 'Turn around and park in front of that vehicle in the shade'. Similar to `CLEVR-Ref`, object referral in `Talk2Car` requires some form of spatial rea-

Dataset	Images	Objects	Expressions	Avg expr length	Video	Lidar	Radar
ReferIt (Kazemzadeh et al., 2014)	19,894	96,654	130,525	3.46	✗	✗	✗
RefCOCO (Yu et al., 2016)	26,711	50,000	142,209	3.61	✗	✗	✗
RefCOCO+ (Yu et al., 2016)	19,992	49,856	141,564	3.53	✗	✗	✗
RefCOCOg (Mao et al., 2016)	26,711	54,822	85,474	8.43	✗	✗	✗
CLEVR-Ref (Hu et al., 2018)	99,992	492,727	998,743	14.50	✗	✗	✗
Cityscapes-Ref (Vasudevan et al., 2018)	4,818	29,901	30,000	15.59	✓	✗	✗
Talk2Car (Ours)	9,217	10,519	11,959	11.01	✓	✓	✓

Table 1: Statistics of and comparison with existing datasets for object referral.

soning. However, in contrast to the former, there are cases where the spatial description in the command is misleading and truthfully reflects mistakes that people make. An example is the command in Fig. 1(e), where we refer to the object as being on the right side of the image, while the person of interest is actually located on the left.

Another important difference is the type of images in each dataset. For instance, the urban images in RefCOCO are taken from the viewpoint of a pedestrian. On the other hand, the images in Talk2Car are car centric.

5 Application of the State-of-the-Art Models and their Evaluation

We assess the performance of 7 models to detect the referred object in a command on the Talk2Car dataset. We discriminate between state-of-the-art methods based on region proposals and non-region proposal methods, apart from simple baselines. [†]

5.1 Region Proposal Based Methods

Object Sentence Mapping (OSM) This region proposal based method uses a single-shot detection model, i.e., SSD-512 (Liu et al., 2016), to extract 64 interest regions from the image. We pretrain the region proposal network (RPN) for the object detection task on the train images from the Talk2Car dataset. A ResNet-18 model is used to extract a local representation for the proposed regions. The natural language command is encoded using a neural network with Gated Recurrent Units (GRUs). Inspired by (Karpathy et al., 2014), we use the inner product between the latent representation of the region and command as a score for each proposal. The region that gets assigned the highest score is returned as bounding box for the object referred to by the command.

Spatial Context Recurrent ConvNet (SCRC) A shortcoming of the above baseline model is that the correct region has to be selected based on local information alone. Spatial Context Recurrent ConvNets (Hu et al., 2016) match both local and global information with an encoding of the command. We reuse the SSD-512 model from above to generate region proposals. A global image representation is extracted by a ResNet-18 model. Additionally, we add an 8-dimensional representation of the spatial configuration to the local representation of each bounding box, $x_{spatial} = [x_{min}, y_{min}, x_{max}, y_{max}, h, w, x_{center}, y_{center}]$ with h and w respectively being the height and the width of this bounding box. For more details, we refer to the original work (Hu et al., 2016).

5.2 Non-Region Proposal Based Methods

MAC model This model (Hudson and Manning, 2018) originally created for the VQA task uses a recurrent MAC cell to match the natural language command represented with a Bi-LSTM model with a global representation of the image. A ResNet-101 is used to extract the visual features from the image. The MAC cell decomposes the textual input into a series of reasoning steps, where the MAC cell attends to certain parts of the textual input to guide the model to look at certain parts of the image. Between each of these reasoning steps, information is passed to the next cell such that the model is capable of representing arbitrarily complex reasoning graphs in a soft manner in a sequential way. The recurrent control state of the MAC cell identifies a series of read and write operations. The read unit extracts relevant information from both a given image and the internal memory. The write unit iteratively integrates the information into the cells' memory state, producing a new intermediate result.

Stack-NMN The Stack Neural Module Network or Stack-NMN (Hu et al., 2018) uses multiple modules that can solve a task by automatically

---

[†]All parameter values obtained on the validation set are cited in the supplementary material. For both the MAC and STACK-NMN we give the best results after empirically setting the number of reasoning steps.

Method	$IoU_{0.50}$ (%)	Inference Speed (ms)	Params (M)
RS	2.05	-	-
BOBB	3.15	-	-
RNM	9.70	-	-
MAC (Hudson and Manning, 2018)	**50.51**	51	41.59
STACK-NMN (Hu et al., 2018)	33.71	52	35.2
OSM (Karpathy et al., 2014)	35.31	71	43.0
SCRC (Hu et al., 2016)	38.70	90	59.5

Table 2: Performance ($IoU_{0.50}$), inference speed (evaluated on a TITAN XP) and number of parameters of the different models.

inducing a sub-task decomposition, where each sub-task is addressed by a separate neural module. These modules can be chained together to decompose the natural language command into a reasoning process. Like the MAC model, this reasoning step is based on the use of an attention mechanism to attend to certain parts of the natural language command, which on their turn guide the selection of neural modules. The modules are first conditioned with the attended textual features after which they perform sub-task recognitions on the visual features. The output of these modules are attended parts in the image which are then given to the next reasoning step to continue the reasoning process. Again, a ResNet-101 model is used to extract the image features and a Bi-LSTM to encode the natural language command. To predict the referred object this model first splits the given image into a 2D grid. Then it tries to predict in which cell located in the grid the center of the referred object lies. Once this has been predicted, the model predicts the offsets of the bounding box relative to the predicted center.

5.3 Simple Baselines

Random Selection (RS) We reuse the single-shot detection model from section 5.1 to generate 64 region proposals per image of the test set. This model randomly samples one region from the proposals and uses it as prediction for the referred object. This is done 100 times and results are averaged.

Biggest Overlapping Bounding Box (BOBB) From the heatmap in Fig. 2 (a) we can see that there is some bias of the referred objects on the left side. This model tries to exploit this information by searching a 2D bounding box that optimizes the overlap with all the bounding boxes in the training set. The algorithm is explained in Section A of the supplementary material.

Random Noun Matching (RNM) In the test set a dependency parser (Honnibal and Johnson, 2015) is used to extract the set of nouns from a given command. We keep the nouns which are substrings of the category names. Then, we randomly sample an object from the region proposals of the corresponding image. If the set of category names is empty, we randomly sample a region from all region proposals. We re-use the RPN explained in OSM for the region proposals. This method is evaluated 100 times before averaging the results.

5.4 Results and Discussion

Overall Results We evaluated all seven models on the object referral task, using both the test split from subsection 3.3 as well as multiple increasingly challenging subsets from this test set. To properly evaluate existing models against our baselines we convert the 3D bounding boxes to 2D bounding boxes. We consider the predicted region correct when the Intersection over Union (IoU), that is, the intersection of the predicted region and the ground truth region over the union of these two, is larger than 0.5. Additionally, we report average inference speed at prediction time per sample and number of parameters of each model. We report the results obtained on the test set in Table 2. The results over the challenging test subsets can be seen in Fig. 3.

In all results we see the following: First, it is clear that the simple baselines (RS, BOBB, RNM) do not perform well, which evidences the difficulty of the object referral task in the realistic settings captured in `Talk2Car`. Second, MAC performs the best on nearly all tasks and it performs significantly better than STACK-NMN which is the model that resembles MAC the most.

If we compare the two RPN systems we see that SCRC often outperforms OSM, showing that using spatial information is beneficial. Third, being
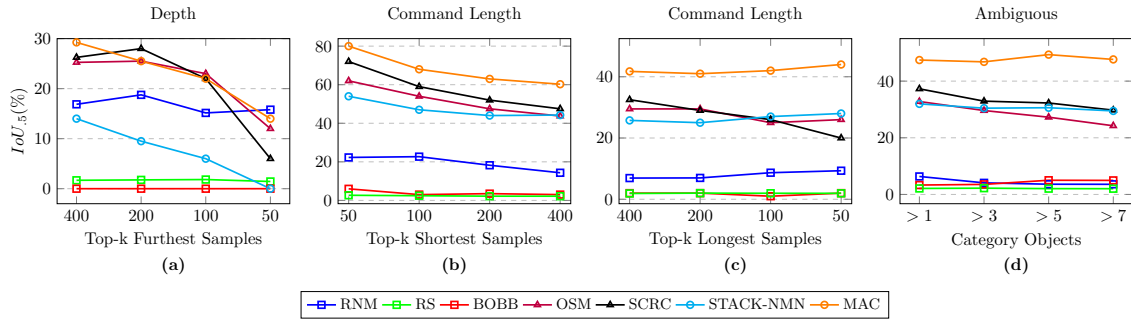
2094

Figure 3: Test performance $IoU_{0.5}$ of different methods on the challenging sub-test sets. We discriminate between A test set for the top-k furthest objects in Fig. (**a**), the top-k shortest and longest commands in Fig. (**b**) and Fig. (**c**) respectively, and in function of the number of objects of the same category in the scene in Fig. (**d**).

able to discriminate between the different object classes in the scene is important. Or put it differently, correct alignment between objects in the image and the category names mentioned in the command is a basic requirement. RNM shows us that concentrating on nouns already gives a big improvement over a purely random strategy. In a separate experiment using ground truth bounding boxes the RNM system obtained an IoU of 54% showing the importance of the alignment of a found object to the correct category name. Fourth, the command length has a negative impact on most models as can be seen in 3(c). We argue that when commands get longer there might be more irrelevant information included which the models have difficulty to cope with. Fifth, from our experiments we found that the non-RPN systems are roughly two times faster than the RPN-systems. This is due to the fact that these RPN-systems have to align every proposed region with the command. On the other hand, the non-RPN systems only have to encode the full image once and then reason over this embedding. Lastly, when looking at the ambiguity test in Fig. 3(d) we see that all models struggle when the number of ambiguous objects of one category increase except for STACK and MAC, whose performance remains fairly stable. We believe they benefit from the multiple reasoning steps before giving an answer where modifier constructions in language disambiguate the referred object. In a separate experiment we have focused on object referral in extra long commands with ambiguous objects of the same category, where we observe the same trends.

Influence of Region Proposal Quality We consider the case when we pre-train a RPN on all keyframes from the training videos, rather than only on the images with commands. It is found that the test performance of the OSM model increased from 35.31 to 40.78%. Similarly, the test performance of the SCRC model increased from 38.70 to 41.15 % showing the importance of starting from good region proposals.

Blanking out the commands (Cirik et al., 2018) found that some referential datasets have some kind of bias in the dataset when blanking out the question. We evaluated this with both SCRC and OSM by changing the question vector to a zero filled vector and we respectively got the following results. For SCRC we get 40.37% $IoU_{0}.5$ (38.70% with command), OSM: 21.65% (35.31% with command). From these results we can conclude two things. First, global information which was added to the local representation of each region in the SCRC model, contains some kind of bias that the models can learn. Second, if no global information is used, as is the case in OSM, the model $IoU_{0}.5$ actually decreases dramatically indicating that there is not a high bias in the image itself.

Influence of Using Pre-trained Word Embeddings Using pre-trained word GloVe embeddings (Pennington et al., 2014) had no effect on or even lowered the IoU obtained on the test set. We argue that words like 'car' and 'truck' are very close to each other in the embedding space but for the model to perform well it should be able to discriminate between them. We also tested ELMO (Peters et al., 2018) and BERT (Devlin et al.,

2018) embeddings but found that they gave only minor improvements for some models.

6 Conclusions and Future Work

We have presented a new dataset, `Talk2Car`, that contains commands in natural language referring to objects in a visual urban scene which both the passenger and the self-driving car can see. We have compared this dataset to existing datasets for the joint processing of language and visual data and have performed experiments with different strong state-of-the-art models for object referral, which yielded promising results. The available 3D information was neglected to be able to compare existing models but we believe that it could help in object referral as it contains more spatial information which, as seen in the experiments, is an important factor. This 3D information will help to translate language into 3D. Moreover, it will allow to perform actions in 3D based on the given command. Also, the `Talk2Car` dataset only allows people to refer to one object at a time. It also doesn't include path annotations for the car to follow, nor does it have dialogues if a command is ambiguous. In future versions, `Talk2Car` will be expanded to include the above annotations and dialogues. However, this first version already offers a challenging dataset to improve current methods for the joint processing of language and visual data and for the development of suitable machine learning architectures. Especially for cases where the ambiguity in object referral can be resolved by correctly interpreting the constraints found in the language commands, `Talk2Car` offers a natural and realistic environment to study these.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2017. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2019. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? *arXiv preprint arXiv:1805.11818*.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017. Embodied Question Answering.

Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. 2013. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 2(2):58–79.

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

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. *CoRR*, abs/1606.01847.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2017. IQA: visual question answering in interactive environments. *CoRR*, abs/1712.03316.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. 2017. Grounded Language Learning in a Simulated 3D World. pages 1–22.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. *CoRR*, abs/1807.08556.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564.

Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. *CoRR*, abs/1803.03067.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Inferring and Executing Programs for Visual Reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 3008–3017.

Kushal Kafle and Christopher Kanan. 2016. Answer-type prediction for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4976–4984.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

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.

Thomas Kollar, Stefanie Tellex, Matthew R Walter, Albert Huang, Abraham Bachrach, Sachi Hemachandra, Emma Brunskill, Ashis Banerjee, Deb Roy, Seth Teller, et al. 2013. Generalized grounding graphs: A probabilistic framework for understanding grounded language. *Journal of Artificial Intelligence Research*, pages 1–35.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Mohammad Reza Loghmani, Barbara Caputo, and Markus Vincze. 2018. Recognizing objects in-the-wild: Where do we stand? In *IEEE International Conference on Robotics and Automation (ICRA)*.

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.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Ed Richardson and Philip Davies. 2018. The changing public's perception of self-driving cars.

Dan Robitzski. 2019. Exclusive: A waymo one riders experiences highlight autonomous rideshares shortcomings. https://futurism.com/waymo-one-early-rider-autonomous-vehicle, note=Accessed: 2019-01-27.

Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*.

Joseph Suarez, Justin Johnson, and Fei-Fei Li. 2018. DDRprog: A CLEVR Differentiable Dynamic Reasoning Programmer.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Arun Balajee Vasudevan, Dengxin Dai, Luc Van Gool, and ETH Zurich. 2018. Object referring in videos with language and human gaze.

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 *Advances in neural information processing systems*, pages 5998–6008.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *CoRR*, abs/1807.03367.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*.

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

Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *CoRR*, abs/1512.02167.