Giving Commands to a Self-driving Car: A Multimodal Reasoner for Visual Grounding

Thierry Deruyttere, Guillem Collell, Marie-Francine Moens
KU Leuven
Thierry.Deruyttere@cs.kuleuven.be
Guillem.Collell@cs.kuleuven.be
Sien.Moens@cs.kuleuven.be

Figure 1: Command for a self driving car from the Talk2Car dataset: "You can park up ahead behind the silver car, next to that lamp post with the orange sign on it".

Abstract

In this paper, we propose a new spatial memory cell and a spatial reasoner for the Visual Grounding task. The goal of this task is to find a certain object in an image based on a given textual query. Our work focuses on integrating the regions of a Region Proposal Network (RPN) into a new multistep reasoning model which we call a multimodal Global Positioning System (GPS) Reasoner. The introduced model uses the object regions from an RPN as initialization of a 2D spatial memory and then implements a multistep reasoning process scoring each region according to the selected words of the query, hence why we call it a multimodal reasoner. We evaluate this new model on the recently proposed Talk2Car dataset, which is a real-world referring expression dataset containing commands for a self-driving car. The experiments show that our model, which reasons jointly over the object regions of the image and words of the query, largely improves the detection accuracy of the referred object compared to current state-of-the-art models.

Introduction

Visual Grounding (VG) is a task relevant to many real-world scenarios and is defined as follows: Given a natural language expression, localize an image region based on this expression (Yu et al. 2018; Mao et al. 2016). VG is useful for a variety of reasons. For instance, when taking a ride in a self-driving car, the passenger might want to instruct the car by saying, e.g.,"stop next to my friend with his red shirt next to the tree" (Figure 1). Another useful application of this task is service robots for the elderly. A person of age could say to a robot "get me that can of coke next to the fridge" upon which the robot has to first locate the object and then execute the command.

The approaches for VG and to the related Visual Question Answering (VQA) task, where a model has to give a textual answer about a question for a certain image, can be divided into two different paradigms. The first paradigm, which is common in VG, is a two-staged method. First, a Region Proposal Network (RPN) predicts regions for objects in the image that function as candidate regions for the sought object. Secondly, a model tries to rank these regions according to the query and the highest scoring region is selected as the answer (Hu et al. 2016; Nagaraja, Morariu, and Davis 2016; Deng et al. 2018). The second paradigm, which is also used in VQA, uses a (multistep) reasoning system that consists of multiple modules, also called cells. One of these cells is used to decompose the query and guides the search over the image in order to extract information from it. Finally, an answer is generated based on this process.

In this paper, we propose a novel method for the VG task that incorporates both the region ranking paradigm and the multistep reasoning paradigm.

To this end, we have created a new type of cell, called Global Positioning System (`GPS`) cell, that incorporates 2D spatial information from extracted regions in a spatial data structure which we call the `GPSMap`. This cell is integrated into a new multimodal model called `GPS Reasoner` which jointly reasons over the words of the query and object regions in the image. We evaluate this model on the Talk2Car dataset (Deruyttere et al. 2019) which is a referential expression dataset that contains referential commands given to self-driving cars. This dataset consists of multiple modalities (LIDAR, RADAR, Video, ...) but in this paper we only focus on the images and the referential expressions.

The main contributions of this paper are as follows:

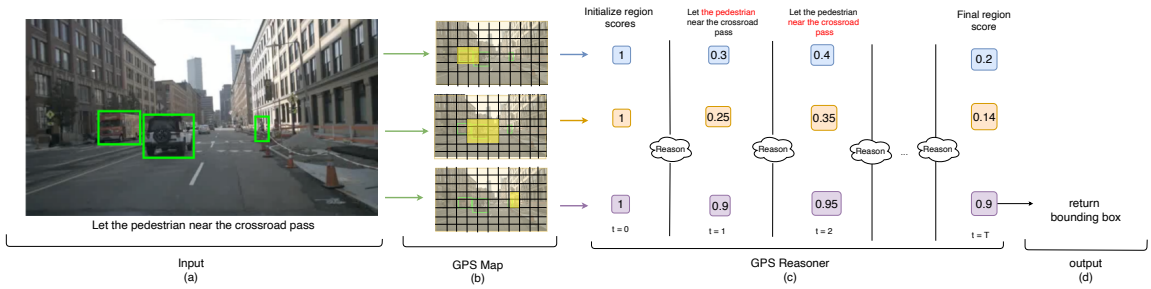

Figure 2: How the GPS Reasoner works. In step (a) an image is given together with a command and R (in this case $R = 3$) bounding boxes. In step (b), each of these bounding boxes receives an entry in the spatial data structure called the GPSMap. Step (c) represents the reasoning process of our model. At the start ($t = 0$), each region has a score of 1. While the reasoning process progresses, regions will receive new scores based how well they align with the words that are focused in that reasoning step. These focused words are indicated in red in each step. At the end of the reasoning ($t = T$), the region with the highest score is returned as the answer of the model in step (d).

1. We propose a novel integration method of decomposing a query in a multistep reasoning process while continuously ranking regions during each step leading to low scoring regions to be ignored during the reasoning process. This process leads to better coupling region proposals with decomposed queries.
2. We propose a new multimodal model called the GPS Reasoner based on this integration.
3. The GPS Reasoner uses a new Global Positioning System (GPS) cell which stores 2D spatial information in a data structure called a GPSMap.
4. We evaluate our model on the Talk2car (Deruyttere et al. 2019) dataset and show that our results improve the best state-of-the-art model by almost 9% in terms of IoU of the found referred object.

GPS Reasoner

For the VG task, when the GPS Reasoner is given an image with a query and a set of extracted object regions for this image, it should select the best region according to the query. This selection is based on the final score of the region. To assign a score to each region, the model performs a multistep reasoning process over the query, the image and the regions. In each reasoning step, the model first focuses on certain words of the query. Then, the model extracts information in parallel for each region from the image based on these words. Finally, according to how well the extracted information of each region aligns with the focused words, the model will assign scores in parallel to each of these regions. A simplified version of this process is presented in Figure 2. To execute this process, our model implements different modules that interact with each other, which we henceforth refer to as cells. The three different cells are as follows: (i) a TxtReader, based on (Deng et al. 2018; Hudson and Manning 2018; Yu et al. 2018; Hu et al. 2018), that controls the decomposition of the query text and thus dictates how the reasoning process will unfold, (ii) a novel GPS cell that functions as the 2D spatial memory of the model by using a GPSMap, (iii) an ImgReader, based on (Deng et al. 2018; Hudson and Manning 2018), that extracts information from a given image based on the control of (i) and the spatial memory from (ii). In our model, the alignment between words in the query and objects in the image remains transparent in the reasoning steps, and the model also emphasizes the joint reasoning with the two modalities. A full detailed implementation of our model will be published at a later date.

GPSMap

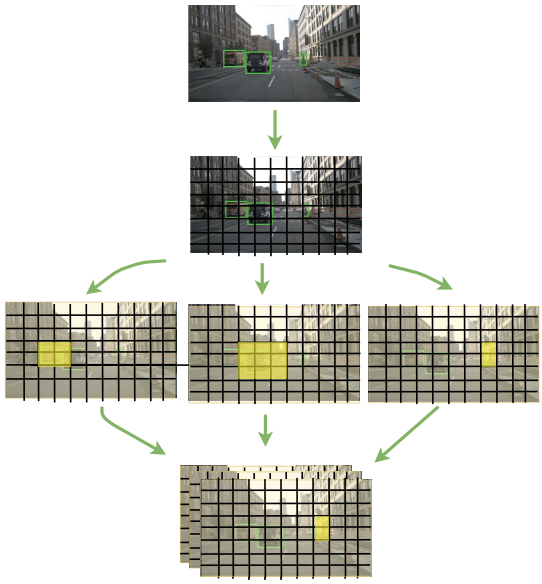

Figure 3: Example of spatial maps in the GPSMap for 3 regions (green bounding boxes). First, the image is divided into grid cells, then for each found object a separate spatial map is created where the cells belonging to the region are given a score of 1. The other cells receive a lower weight.

An important part of the GPS Reasoner, used in the

GPS cell, is the GPSMap which is a spatial data structure that is calculated a-priori and which stores the location of R found object regions from a RPN. This spatial data structure is of shape $[R \times H_f \times W_f]$ with H_f and W_f respectively the height and width of the extracted image features. To create the data structure, every bounding box r is mapped on a 2D spatial map of shape $[H_f \times W_f]$. The mapping assigns a weight of 1 to every cell of this 2D map that falls inside the bounding box r, and a lower weight elsewhere (in our case 0.5). This map also stores the score of each bounding box during the reasoning process. This process can be seen in Figure 3. The reasoning behind the spatial map is that it is used in the ImgReader to (a) indicate the location of the found object of bounding box r – this will also indicate where the reasoning process should take place – by giving those cells a weight of 1, and (b) the model should also be able to see objects that are located elsewhere in the picture, hence why we give the other cells a non-zero weight. An example of this is if we are looking for "the man next to the tree" and an entry in the GPSMap indicates the location of a man, then the ImgReader cell should look around this region for the tree. The model is thus checking which regions correspond the most with the given expression which can be seen as a bottom-up reasoning process. A top-down reasoning process would be to only take the image and reason over the full image to find the looked after region.

Dataset

In the experiments below we train and test on the Talk2Car dataset (Deruyttere et al. 2019), which contains images from the nuScenes dataset (Caesar et al. 2019) that are annotated with natural language commands, bounding boxes of scene objects, and the bounding box of the object that is referred to in a command. The Talk2Car dataset consists of commands given to a self-driving car. In total it contains 11,959 commands that belong to 9,217 images, which are either taken in Singapore or Boston during different weather (sun or rain) and time conditions (night or day). This dataset was selected because of its complex natural language commands that constrain - through modifying language expressions - the object to be found in the scene demanding reasoning over the objects in the scene and words in the command. Note also that in Singapore and Boston they drive on different sides of the road. Train, validation and test sets contain respectively 8,349 (69.8%), 1,163 (9.721%) and 2,447 (20.4%) commands. On average a command and an image each contain respectively around 11 words and 11 objects. Among the words of the commands around 21% are nouns, 21% verbs and around 6% are adjectives. In the ground truth annotations there are 23 different object categories (e.g., car, truck, man, tree). On average an image contains more than 4 objects of the same category (e.g., cars). In addition, the dataset consists of several test sets, each of which evaluate specific challenging settings while the full test is used to assess the overall performance of the model. A first sub-test set assesses the ability of a model to recognise distant referred objects. The second and third sub-test sets evaluate how well a model can cope with short and long commands respectively. The final sub-test set assesses how the model copes with ambiguity. In our case ambiguity refers to having multiple objects of the referred class in the visual scene.

Experimental setup

We evaluate the proposed GPS Reasoner against 5 different strong baselines on three different measures. Every model that uses regions will use a CenterNet (Zhou, Wang, and Krähenbühl 2019) model to extract these regions. For these models we will do tests with different amount of regions (top-k) based on their confidence score.

Baselines

1. SCRC (Hu et al. 2016).
2. Transformed MAC (Hudson and Manning 2018) for VG. **Currently the state of the art on Talk2car**.
3. STACK (Hu et al. 2018).
4. A-ATT (Deng et al. 2018).
5. GPSPrior: The last model, called GPSPrior, uses a GPSMap where only the object regions receive a weight of 1. The cells that fall outside of a region receive a 0 weight. This GPSMap is then multiplied with the extracted image features to limit the search space to only these parts. These altered image features are passed to MAC to reason with.

Measures

All the models are evaluated with three measures. The first measure is the overall accuracy of the model. This is defined as the percentage of predicted regions that have an Intersection over Union (IoU) or overlap, with the ground truth regions of over 0.5. The second measure is inference speed as the setting in the Talk2Car dataset is a time critical setting. The final measure is the number of parameters of each model.

Results

The results of our GPS Reasoner compared to our 4 baselines and the state of the art (MAC) on the Talk2Car test set are shown in Table 1. In this table we see the three different measures mentioned before. From these results we see that the GPS Reasoner clearly outperforms all the other models for any top-k confident number of regions. Note that this top-k confidence selection mechanism of bounding boxes is still very simple but it already gives a big improvement. The best GPS Reasoner model further improves the state of the art baseline (MAC) by 17% relatively in terms of IoU. In terms of number of trainable parameters, our model is roughly on par with the baselines, but is about five time slower than MAC at inference time. We argue that the difference in accuracy and inference times comes from the fact that our model reasons over all the regions integrated in the visual field while MAC reasons solely over the image. The GPS Reasoner was also trained with only ground truth bounding boxes to know the theoretical limit of the model and achieved an IoU of 68%.

Method	$IoU_{0.50}$ (%)	Inference Speed (ms)	Params (M)
MAC (Hudson and Manning 2018)	50.51	51	41.59
STACK (Hu et al. 2018)	33.71	52	35.2
SCRC (Top-32) (Hu et al. 2016)	43.80	208	52.47
GPSPrior (Top-32)	49.94	179	61.76
A-ATT (Top-16) (Deng et al. 2018)	45.12	180	160.31
GPS Reasoner (top-8)	56.85	224.7	62.25
GPS Reasoner (top-16)	**59.26**	270.5	62.25
GPS Reasoner (top-32)	58.93	359.7	62.39
GPS Reasoner (top-64)	51.74	576.2	62.39

Table 1: Performance ($IoU_{0.50}$), inference speed (evaluated on a TITAN XP) and number of parameters of the different models. All models that use object regions have been evaluated with the top-k ($k = 8, 16, 32, 64$) scoring regions. In the table we only display the best k-value for SCRC, GPSPrior and A-ATT for brevity.

Conclusion

In this paper we introduced a multimodal and multistep reasoning model, called `GPS Reasoner`, for VG tasks. The model does not only decompose the query in multiple reasoning steps, but it also decomposes the reasoning process over the extracted regions by using a `GPS Map` which allows the reasoning process over the regions to happen in an independent and parallel manner. The `GPS Reasoner` was evaluated on the Talk2Car dataset which is composed of images of city environments taken from the viewpoint of a car accompanied by commands that passengers give to the car. The proposed model that jointly reasons over the words of the command and the detected objects in the visual scene with a large margin outperforms state-of-the-art models in terms of the IoU metric that compares the ground-truth referred object with the object found. We argue that having a separate reasoning process for each region, thanks to the `GPSMap`, is certainly beneficial in environments like self-driving cars as it allows the reasoning process to remain transparent but also indicates when the model is hesitant in certain situations based on the scores of certain regions. In future work the value of this spatial map will be further explored as assistance in the navigation process of the car when executing the command. Also the multimodal multistep reasoning model gives many possibilities for future improvements which we did not yet explore in this paper. Some of these improvements are, for instance, taking into account object names recognised in the image and their probability distributions (cf. as done in linking text entities with knowledge base entities in (Le and Titov 2019)), integrating intelligent selection mechanisms that give priority to the processing of certain words or object regions based on prior knowledge, or dynamically improving the region proposals.

Acknowledgements

For this work we would like to acknowledge MACCHINA (KU Leuven, C14/18/065), the Flanders AI Impuls Programme and also CALCULUS (H2020-ERC-2017-ADG 788506). Additionally, we would like to thank Nvidia for granting us two TITAN Xp GPUs.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2019. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Deng, C.; Wu, Q.; Wu, Q.; Hu, F.; Lyu, F.; and Tan, M. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7746–7755.

Deruyttere, T.; Vandenhende, S.; Grujicic, D.; Van Gool, L.; and Moens, M. F. 2019. Talk2car: Taking control of your self-driving car. 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)*, 2088–2098.

Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4555–4564.

Hu, R.; Andreas, J.; Darrell, T.; and Saenko, K. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 53–69.

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

Le, P., and Titov, I. 2019. Boosting entity linking performance by leveraging unlabeled documents. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 1935–1945.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

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

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.