TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments

Howard Chen[*]
ASAPP Inc.
hchen@asapp.com

Alane Suhr
Cornell University
suhr@cs.cornell.edu

Dipendra Misra
Cornell University
dkm@cs.cornell.edu

Yoav Artzi
Cornell University
yoav@cs.cornell.edu

Abstract

We introduce the TOUCHDOWN dataset for instruction following and spatial reasoning in a visually grounded environment. TOUCHDOWN contains instructions paired with gold executions for the tasks of navigation and spatial description resolution. TOUCHDOWN contains real-life, complex visual environments. We show through qualitative analysis that TOUCHDOWN requires complex spatial reasoning and contains a broad set of linguistic phenomena. Finally, we present a text-conditioned image feature reconstruction approach for spatial description resolution. Our results demonstrate the efficacy of our method, while highlighting the remaining challenges.

1 Introduction

Consider the visual challenges of following natural language instructions in a busy urban environment. Figure 1 illustrates this problem. The agent must identify objects and their properties to resolve mentions to *traffic light* and *American flags*, identify patterns in how objects are arranged to find the *flow of traffic*, and reason about how the relative position of objects changes as it moves to *go past* objects. Reasoning about vision and language has been studied extensively with various tasks, including visual question answering [e.g., Antol et al., 2015], visual navigation [e.g., Anderson et al., 2018, Misra et al., 2018], and interactive question answering [e.g., Das et al., 2017]. However, existing work has largely focused on relatively simple visual input, including object-focused photographs [Lin et al., 2014, Reed et al., 2016] or simulated environments [Bisk et al., 2016, Das et al., 2017, Kolve et al., 2017, Misra et al., 2018, Yan et al., 2018]. While this has enabled significant progress in visual understanding, the use of real-world visual input not only increases the challenge of the vision task, it also drastically changes the kind of language it elicits and requires fundamentally different reasoning.

In this paper, we study the problem of reasoning about vision and natural language using an interactive visual navigation environment based on Google Street View.[2] We design the task of following instructions to reach a goal position, and then resolving a spatial description by identifying the location in the observed image of Touchdown,[3] a hidden teddy bear. Using this environment and task, we release TOUCHDOWN,a dataset for navigation and spatial reasoning with real-life observations.

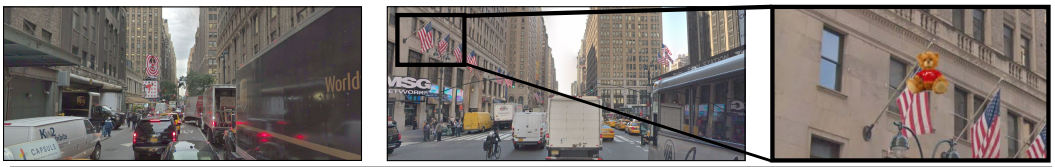
To collect diverse and challenging instructional language data, we design a series of crowdsourcing tasks, focusing on a leader-and-follower design. First, a leader worker takes a sampled route in Street View and writes natural language instructions describing the route. The worker then places Touchdown in a location in the final position and describes this location. In a second separate task, a

[*]Work done at Cornell University.

[2]https://developers.google.com/maps/documentation/streetview/intro

[3]Touchdown is the unofficial mascot of Cornell University.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.



Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Figure 1: An illustration of the task. The agent makes an initial environment observation and re-orients itself (leftmost image). The agent follows the natural language instruction to navigate to its goal position (center image). Upon reaching the goal, the agent resolves the spatial description (underlined) to a location in the observed image to locate Touchdown the bear, which is not visible to the agent. If the location is guessed correctly, Touchdown is revealed (rightmost image).

follower worker uses the written instructions to follow the route and find the target without knowledge of the correct route or target location. We use the second task to validate the instructions written by the leader. TOUCHDOWN contains 9,326 examples. Our linguistically-driven analysis of TOUCHDOWN shows that the data requires complex spatial reasoning.

We conduct an empirical analysis of the task of finding Touchdown in the final image. We present a text-conditioned image feature reconstruction approach for spatial description resolution using LINGUNET [Misra et al., 2018], a recent model for spatial resolution, and compare its performance to several baselines. Our qualitative and empirical analyses demonstrate that TOUCHDOWN poses a challenging task. TOUCHDOWN will be made available at `https://github.com/clic-lab/touchdown`.

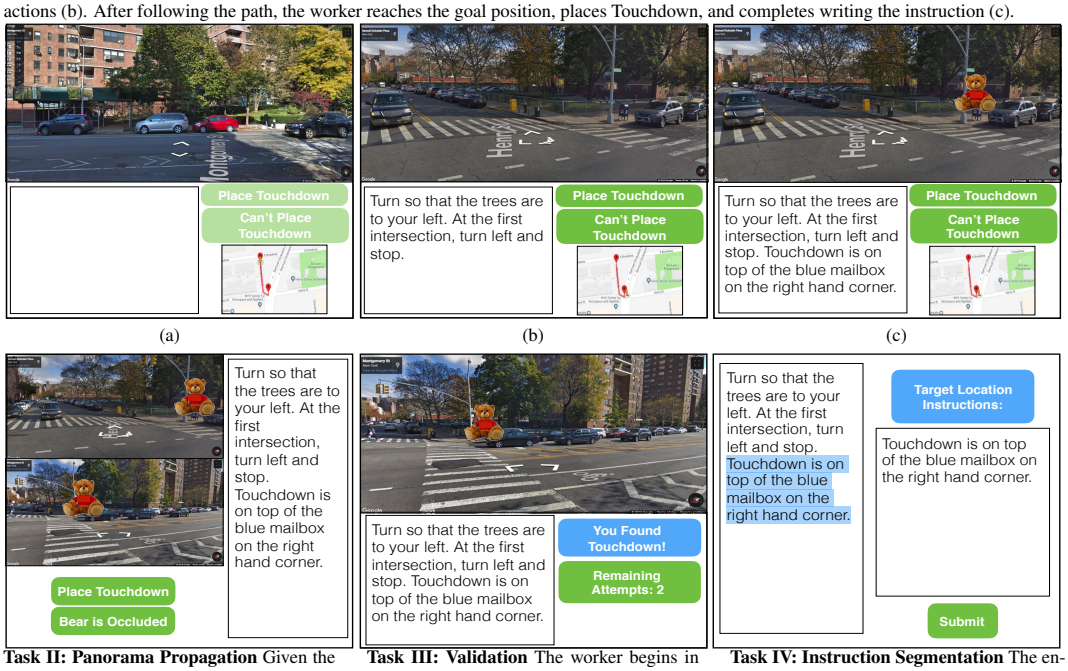
2 Related Work

Jointly reasoning about vision and language has been studied extensively, most commonly focusing on static visual input for reasoning about image captions [Chen et al., 2015, Lin et al., 2014, Suhr et al., 2017, Reed et al., 2016] and grounded question answering [Antol et al., 2015, Zitnick and Parikh, 2013]. Recently, the problem has been studied in interactive simulated environments where the visual input changes as the agent acts, such as interactive question answering [Das et al., 2017, Gordon et al., 2018] and instruction following [Misra et al., 2018, 2017]. In contrast, we focus on an interactive environment with real-world observations.

The most related resources to ours are R2R Anderson et al. [2018] and Talk the Walk de Vries et al. [2018]. R2R uses panorama graphs of house environments for the task of navigation instruction following. R2R includes 90 unique environments, each containing an average of 119 panoramas, significantly smaller than our 29,641 panoramas. We also observe that the language in our data is significantly more complex than in R2R (Section 5). Our environment setup is also related to Talk the Walk, which uses panoramas in small urban environments for a navigation dialogue task. In contrast to our setup, the instructor does not observe the panoramas directly, but instead sees a simplified diagram of the environment with a small set of pre-selected landmarks. As a result, the instructor has less spatial information compared to TOUCHDOWN. Instead the focus is on agent-to-agent conversational coordination.

3 Tasks and Evaluation

We design two tasks: navigation and spatial description resolution (SDR). The agent's goal in the navigation task is to follow a natural language instruction and reach a goal position. In the SDR task, given an image and a natural language description, the task is to identify the point in the image that is referred to by the description. Both tasks require recognizing objects and the spatial relations between them. The navigation task focuses on egocentric spatial reasoning, where instructions refer to the agent's relationship with its environment, including the objects it observes. The SDR task displays more allocentric reasoning, where the language requires understanding the relations between the observed objects to identify the target location. The navigation task requires generating a sequence of actions from a small set of possible actions. The SDR task requires choosing a specific pixel in the observed image. Both tasks present different learning challenges. The navigation task could benefit from reward-based learning, while the SDR task defines a supervised learning problem. The two



Task II: Panorama Propagation Given the image from the leader's final position (top), including Touchdown's placement, and the instructions (right), the worker annotates the location of Touchdown in the neighboring image (bottom).

Task III: Validation The worker begins in the same heading as the leader, and follows the instructions (bottom left) by navigating. When the worker believes they have reached the goal, they guess the target location by clicking in the Street View image.

Task IV: Instruction Segmentation The entire instruction is shown on the lefthand side. The worker highlights segments corresponding to the navigation and target location subtasks. On the right, the highlighted sentence is shown to the worker.

Figure 2: Illustration of the data collection process.

tasks can be addressed separately, or combined by completing the SDR task at the goal position at the end of the navigation.

Each of the tasks include several evaluation metrics. For navigation, we use three evaluation metrics: task completion, shortest-path distance, and trajectory edit distance. In SDR, we evaluate accuracy and distance error, which is the mean distance of the predicted pixel from the gold label.

4 Data Collection

We frame the data collection process as a treasure-hunt task where a leader hides a treasure and writes directions to find it, and a follower follows the directions to find the treasure. The process is split into four crowdsourcing tasks illustrated in Figure 2. The two main tasks are writing and following. In the writing task, a leader follows a prescribed route and hides Touchdown the bear at the end, while writing instructions that describe the path and how to find Touchdown. The following tasks requires following the written instructions from the same starting position to navigate and find Touchdown. Additional tasks are used to segment the instructions into the navigation and target location tasks, and to propagate Touchdown's location to panoramas that neighbor the final panorama. We use a customized Street View interface for the writing and following tasks. However, the final data uses a static set of panoramas that do not require the Street View interface. Figure 2 illustrates the data collection process.

5 Data Statistics and Analysis

TOUCHDOWN contains a total of 9,326 examples, split into 70%/15%/15% training/development/test sets. The environment includes 29,641 panoramas from New York City. In the training and development sets, instructions are on average 108.0 tokens long, where navigation segments contain on average 89.6 tokens and SDR segments contain on average 29.8 tokens. Routes include 35.2 panoramas on average. Table 1 shows linguistic analysis comparing TOUCHDOWN and R2R. Our analysis shows that resolving semantic phenomena like coordination, spatial relations, and comparisons is key to successfully completing the navigation and SDR tasks. Our data displays a significantly more diverse set of semantic phenomena compared to R2R.

Phenomenon	R2R		TOUCHDOWN		Example from TOUCHDOWN
	c	$\mu \pm \sigma$	c	$\mu \pm \sigma$	
Reference to unique entity	25	3.7	25	10.7	*... You'll pass **three trashcans** on your left ...*
Coreference	8	0.5	22	2.4	*... a brownish colored brick building with a black fence around **it**...*
Comparison	1	0.0	6	0.3	*... The bear is in the middle of the **closest** tire.*
Sequencing	4	0.2	22	1.9	*... Turn left at the **next** intersection ...*
Count	4	0.2	11	0.5	*... there are **two** tiny green signs you can see in the distance ...*
Allocentric spatial relation	5	0.2	25	2.9	*... There is a fire hydrant, the bear is **on top***
Egocentric spatial relation	20	1.2	25	4.0	*... up ahead there is some flag poles **on your right hand side**...*
Imperative	25	4.0	25	5.3	*... **Enter** the next intersection and **stop** ...*
Direction	22	2.8	24	3.7	*... Turn **left**. Continue **forward** ...*
Temporal condition	7	0.4	21	1.9	*... Follow the road **until you see** a school on your right...*
State verification	2	0.1	21	1.8	*... **You should see** a small bridge ahead ...*

Table 1: Linguistic analysis of 25 randomly sampled development examples in TOUCHDOWN and R2R. We annotate each example for the presence and count of each phenomenon. c is the number of instructions out of the 25 containing at least one example of the phenomenon; μ is the mean number of times each phenomenon appears in each of the 25 instructions.

Method	A@40px	A@80px	A@120px	Dist
RANDOM	0.21	0.78	1.89	1179
CENTER	0.31	1.61	3.93	759
AVERAGE	2.43	5.21	7.96	744
TEXT2CONV	24.82	30.40	34.13	747
LINGUNET	**26.11**	**34.59**	**37.81**	**708**

Table 2: Test results on the SDR task. We report accuracy with different thresholds (40, 80, and 120) and mean distance error.

6 Initial Experiments and Results

We evaluate three non-learning baselines on the SDR task and two learning approaches. The non-learning baselines are: (a) RANDOM: predict a pixel at random; (b) CENTER: predict the center pixel; (c) AVERAGE: predict the average pixel, computed over the training set. The learning approaches are TEXT2CONV, and LINGUNET. In both methods, we encode the the description into a fixed vector representation with a recurrence neural network. We compute image feature using RESNET18 He et al. [2016]. TEXT2CONV uses the text representation to compute convolution filters that are used to convolve over the image features. We apply a multi-layer perceptron on the outputs of the convolution, computing the distribution over pixels with a $\mathrm{softmax}$. This allows the model to learn the interaction between the text and image modalities [Chen et al., 2016], but at the cost of a large number of parameters. The second approach treats the problem of pixel-scoring as a text-conditioned image feature reconstruction problem with the LINGUNET architecture. LINGUNET uses the text representation to generate a set of 1×1 convolution filters and convolves the image feature map with them through several layers. We then perform a series of deconvolutions to generate a feature map with the same shape as the input image but with a single channel. We apply a $\mathrm{softmax}$ operation to this feature map to generate the target probabilities. Using 1×1 filters, we avoid the need to project the text representation to large convolution filters as required in the TEXT2CONV architecture.

Table 2 shows SDR test results. We observe that using LINGUNET provides a significant boost in performance over the three baselines while requiring fewer parameters.

7 Data Distribution and Licensing

We release an environment graph as panorama IDs and edges, scripts to download the RGB panoramas using the official Google API, the collected data, and our code. These parts of the data are released with a CC-BY 4.0 license. Retention of downloaded panoramas should follow Google's policies. We also provide RESNET18 image features of the panoramas by request. To follow Google's terms-of-service, we collect the emails of all people holding the image features to request data deletion upon request from Google and for other announcements. The complete license is available with the data.

8 Conclusion

We present TOUCHDOWN, a challenging dataset for visual navigation and spatial description resolution situated in a the complex New York City urban environment. Our linguistically-driven qualitative analysis and empirical results show TOUCHDOWN presents complex language and vision challenges.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

Yonatan Bisk, Daniel Marcu, and William Wong. Towards a Dataset for Human Computer Communication via Grounded Language Acquisition. In *Proceedings of the AAAI Workshop on Symbiotic Cognitive Systems*, 2016.

Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ramakant Nevatia. ABC-CNN: An Attention Based Convolutional Neural Network for visual question answering. *Transactions of the Association of Computational Linguistics*, 2016.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data Collection and Evaluation Server. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the Walk: Navigating New York City through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: Visual Question Answering in Interactive Environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

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. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, 2014.

Dipendra Misra, John Langford, and Yoav Artzi. Mapping Instructions and Visual Observations to Actions with Reinforcement Learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, 2017.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.

Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning Deep Representations of Fine-Grained Visual Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A Corpus of Natural Language for Visual Reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 217–223, 2017.

Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. CHALET: Cornell House Agent Learning Environment. *arXiv preprint arXiv:1801.07357*, 2018.

C. Lawrence Zitnick and Devi Parikh. Bringing Semantics into Focus Using Visual Abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013.