Retouchdown: Releasing Touchdown on StreetLearn as a Public Resource for Language Grounding Tasks in Street View

Harsh Mehta
Google Research
harshm@google.com

Yoav Artzi
Cornell University
yoav@cs.cornell.edu

Jason Baldridge
Google Research
jridge@google.com

Eugene Ie
Google Research
eugeneie@google.com

Piotr Mirowski
Deepmind
piotrmirowski@google.com

Abstract

The Touchdown dataset (Chen et al., 2019) provides instructions by human annotators for navigation through New York City streets and for resolving spatial descriptions at a given location. To enable the wider research community to work effectively with the Touchdown tasks, we are publicly releasing the 29k raw Street View panoramas needed for Touchdown. We follow the process used for the StreetLearn data release (Mirowski et al., 2019) to check panoramas for personally identifiable information and blur them as necessary. These have been added to the StreetLearn dataset and can be obtained via the same process as used previously for StreetLearn. We also provide a reference implementation for both Touchdown tasks: vision and language navigation (VLN) and spatial description resolution (SDR). We compare our model results to those given in Chen et al. (2019) and show that the panoramas we have added to StreetLearn support both Touchdown tasks and can be used effectively for further research and comparison.

1 Introduction

Following natural language navigation instructions in visual environments requires addressing multiple challenges in dynamic, continuously changing environments, including language understanding, object recognition, grounding and spatial reasoning. Until recently, the most commonly studied domains were map-based (Thompson et al., 1993) or game-like (Macmahon et al., 2006; Misra et al., 2017, 2018; Hermann et al., 2017; Hill et al., 2017). These environments enabled substantial progress, but the complexity and diversity of the visual input they provide is limited. This greatly simplifies both the language and vision challenges. To address this, recent tasks based on simulated environments include photo-realistic visual input, such as Room-to-Room (R2R; Anderson et al., 2018), Talk-the-Walk (de Vries et al., 2018) and Touchdown (Chen et al., 2019), all of which rely on panorama photos.

A major challenge of creating simulations that use real-world photographs is they at times capture bystanders and their property. This raises privacy concerns and requires additional care to check for and ensure personally identifiable information (PII) is removed from research resources that are made publicly available. Existing resources adopt different strategies to address this. The Matterport3D dataset (Chang et al., 2017), which underlies the R2R task, is focused on real-estate data that is curated to exclude PII. This approach is limited to environments of a specific type: houses that are for sale. Academic resources that focus on urban street scenes opted to manually collect panoramas from scratch and scrub them for PII (de Vries et al., 2018; Weiss et al., 2019). This is laborious and costly—especially the first stage of collecting the panoramas. As a result, such resources cover relatively small areas.

Google Street View has world-wide scale coverage of street scenes. Each panorama in Street View has gone through a process to protect the privacy of bystanders and their property. Individuals can also request specific panoramas to be removed. As such, it is a resource with the potential to transform the research community's ability to study problems such as street scene understanding and navigation. Touchdown relies on 29,641 panoramas from Street View; however, because raw images cannot be distributed according to the Street View terms-of-service,[1] these are not provided with the Touchdown data. Instead, only image feature vectors are available for direct download with the data, and access to the raw panoramas is subject to availability through APIs governed by Street View's terms of service.

[1] https://www.google.com/help/terms_maps/

Proceedings of the Third International Workshop on Spatial Language Understanding (SpLU 2020), pages 56–62
November 19, 2020.
https://doi.org/10.18653/v1/P17

Research can be done within a company and shared via publication without releasing data; for example, Cirik et al. (2018) discussed models for instruction-conditioned navigation in Street View. However, the full impact of the data and research about it can be better realized by making at least some portion of such resources available to the broader research community. In this context, StreetLearn (Mirowski et al., 2018, 2019) stands out as a publicly available resource of Street View data that has been approved for dissemination and use for academic research.[2] StreetLearn contains 114k panoramas from New York City and Pittsburgh that have been manually checked for PII, ensuring, for example, that faces and license plates are blurred. The dataset can be easily accessed. Researchers interested in working with the data simply fill a form stating their goals and commit to update the data periodically with newer versions as they are released. This process balances the ability of researchers to use the data with preserving the privacy and rights of individuals impacted by the data. For example, periodic updates allow Google to respond to user takedown requests.

To increase the accessibility of Touchdown and provide an example of how important data can be responsibly released, we integrate the Touchdown task and its corresponding Street View data into a new version of StreetLearn. This paper reconciles Touchdown's mode of dissemination with StreetLearn's, which was designed to adhere to the rights of Google and individuals while also simplifying access for researchers and improving reproducibility. We also provide open source implementations[3] of both the vision-and-language navigation and spatial description resolution tasks, which we show to have a consistent performance with the results in the original Touchdown paper. We hope that this release of data and code will enable the entire research community to make further progress on these problems and to consider new questions and tasks enabled by this limited but significant slice of Street View data.

2 Process

Touchdown includes tasks for natural language navigation and spatial reasoning in realistic urban environments. Touchdown uses Street View panoramas of New York City to define a large-scale navigation environment. It includes 9,326 human-written instructions and 27,575 spatial description resolution tasks. Touchdown's instructions were written by people and emphasize attributes of the visual environment as navigational cues. This makes Touchdown a valuable resource for research on following natural language instructions in visual environments. This contrasts with the template-based navigation instructions used by Hermann et al. (2020), which were generated by Google Maps API and used with StreetLearn panoramas.

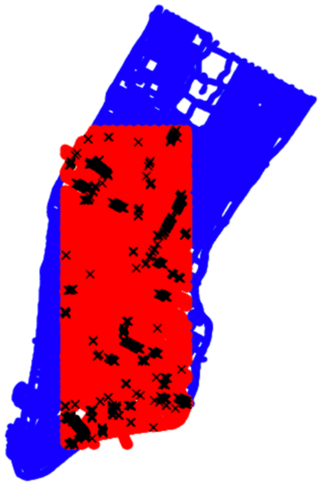

Figure 1: The overlap between the StreetLearn (blue) and Touchdown (red) panoramas in Manhattan. There are 710 panoramas (out of 29k) that share the same ID in both datasets (in black).

Unfortunately, the development and release of Touchdown introduced several challenges that complicate working with the data. Even though Touchdown itself does not contain Street View data, it references specific Street View panoramas and depends on access to them via the Street View API. This requires any researcher that wishes to work on the data to download large amounts of data using the API, which is inconvenient, error-prone and not aligned with the current Google Maps terms-of-service. Also, the panoramas available through the API periodically change, potentially making parts of the data unavailable. This means there is no hope for consistent versioning (which hurts reproducibility) regarding panorama availability because the data collected by each researcher is dependent on the particular time they access it. Finally, individual researchers or research groups cannot themselves comply with takedown requests.

[2] http://streetlearn.cc

[3] https://github.com/google-research/valan/tree/master/touchdown

Figure 2: One of the panoramas taken from the dataset which shows transient objects being referenced in the navigation text. *"Stop here, and turn left. You will now be walking down a narrow lane with parked **cars** on both sides. There should be a payphone on your right and a fire hydrant (behind silver poles) on your left. Walk down this lane, and on your left you will soon see a shop with gray columns between the windows and a blue sign with yellow trim."*

Instead, having the panoramas available as part of StreetLearn allows for necessary updates and consistent sharing of the panoramas.

To address these challenges, we collect, check and release the Touchdown panoramas as part of an update to the 114k existing StreetLearn panoramas, which cover regions of New York City and Pittsburgh. As shown in Figure 1, StreetLearn encompasses the entire region of New York City contained in Touchdown; however, the StreetLearn panoramas themselves are not sufficient for supporting the Touchdown tasks themselves. This is for several reasons.

- The **granularity** of the panorama spacing is different. Figure 1 shows that most of the panoramas are different. Touchdown has roughly 25% of the panos but covers half of Manhattan compared to StreetLearn.
- The language instructions refer to **transient objects** such as cars, bicycles, and couches, as illustrated in Figures 2 and 3. A panorama from a different time period will not contain these objects, so the instructions are not stable across time periods.
- Spatial description resolution requires coverage of multiple points-of-view for those specific panoramas. Figure 3 shows an example SDR description and the corresponding views from which it can be answered.

In all, the Touchdown tasks encompass 29,641 panoramas. All of these went through extensive manual review by annotators to check for personally identifiable information (PII), such as faces and license plates. Regions containing PII were marked as bounding boxes by annotators, and we blurred all of these regions for the final images.

3 Experiments

We re-implement the best-reported models on the navigation and spatial description resolution tasks from Chen et al. (2019) to compare performance with our data release to the original Touchdown paper. The key difference between the two settings is that our released panoramas contain additional blurred patches (Section 2). Another minor difference is that we use a word-piece tokenizer (Devlin et al., 2019) instead of a full-word tokenizer.

Spatial Description Resolution. SDR results are given in Table 1. Following Chen et al. (2019), we report mean distance error and accuracy with different thresholds (40px, 80px, and 120px), which measures the proportion of evaluation items where the pixel chosen by the model is within the specified pixel distance. Our Retouchdown reimplementation of LINGUNET obtains better performance on the accuracy measures, but worse performance on mean distance error. To check whether this is a consequence of the blurring, we ran our model with features retrieved from original panoramas and obtained similar results as those listed in Table 1. Given this, the performance difference between our model and the original paper are likely not due to the additional blurring. As

Figure 3: Actual example taken from the dataset with multiple SDR panorama viewpoints for the same instruction: *Two parked bicycles, and a discarded couch, all on the left. Walk just past this couch, and stop before you pass another parked bicycle. This bike will be white and red, with a white seat. Touchdown is sitting on top of the bike seat.*

Method	A@40px ↑	A@80px ↑	A@120px ↑	Dist ↓
Development				
Chen et al. (2019)	24.81	32.83	36.44	729
Retouch-LINGUNET	29.79	35.28	38.14	800
Test				
Chen et al. (2019)	26.11	34.59	37.81	708
Retouch-LINGUNET	30.32	36.73	39.27	793

Table 1: SDR development and test results using the LINGUNET architecture, which Chen et al. (2019) reported as the best performing system.

Method	TC ↑	SPD ↓	SED ↑	NDTW ↑	SDTW ↑
Development					
Chen et al. (2019)	9.8	19.1	0.094		
Retouch-RCONCAT	13.4	17.1	0.124	4.9	1.3
Test					
Chen et al. (2019)	10.7	19.5	0.104		
Retouch-RCONCAT	12.8	17.1	0.131	5.0	1.4

Table 2: Navigation development and test results. We use the RCONCAT architecture, which Chen et al. (2019) reported as the best performing.

such, the Touchdown panoramas available through StreetLearn can be reliably used as direct replacement for those used in Chen et al. (2019).

Vision-and-Language Navigation. We use the following metrics to evaluate VLN performance:

- Task Completion (TC): the accuracy of navigating to the correct location. The correct location is defined as the exact goal panorama or one of its neighboring panoramas. This is the equivalent of the success rate metric (SR) used commonly in VLN for R2R.
- Shortest-path distance (SPD): the mean of the distances over all executions of the agent's final panorama position and the goal panorama.
- Success weighted by Edit Distance (SED): normalized graph edit distance between the agent path and true path, with points only awarded for successful paths.
- Normalized Dynamic Time Warping (NDTW): a minimized cumulative distance between the agent path and true path, normalized by path length.
- Success weighted Dynamic Time Warping (SDTW): NDTW, with points awarded only for successful paths.

TC, SPD, and SED are defined in Chen et al. (2019) and NDTW and SDTW are defined in Ilharco et al. (2019).

VLN results are given in Table 2. Our Retouchdown reimplementation of the RCONCAT model improves over the results given in Chen et al. (2019) for all metrics. We also establish benchmark scores for NDTW and SDTW. As with SDR, the panoramas now available via StreetLearn thus do not remove information critical for the VLN task. In our implementation, we use imitation learning on top of a scalable framework based on the Actor-Learner architecture (Lansing et al., 2019), instead of supervised learning using Hogwild! (Recht et al., 2011). These differences likely explain the observed differences with the original results.

Compared to interior navigation in the Room-to-Room (R2R) task, the Touchdown task is much harder: e.g. the current state-of-the-art success rate (equivalent to TC) for R2R on the validation unseen dataset is 55% (Zhu et al., 2019). It is even considerably harder than Room-across-Room dataset, which has longer, more challenging paths than R2R and success rates of 26% to 30% for three different languages (Ku et al., 2020). The same holds for DTW measures: Ilharco et al. (2019) report a success rate of 44% and corresponding SDTW of

38.3% for a fidelity-oriented version of the Reinforced Cross-modal Matching agent (Wang et al., 2019). Ku et al. (2020) reports lower SDTW scores of 21% to 24%. Given this, the TC of 12.8% and SDTW of 1.4% obtained by Retouch-RCONCAT and current best results from Xiang et al. (2020) (TC: 19.0%; SDTW: 16.3%), amply demonstrates the challenge of the outdoor navigation problem defined by Touchdown. The greater diversity of the visual environments and the far greater degrees-of-freedom for navigation thus provide plenty of headroom for future research.

4 Conclusion

The research community is interested in using large-scale resources such as Street View for work on computer vision and navigation. In order to comply with Street View's terms-of-service (which allow for only limited use of its data and APIs) and with its data restrictions, we have enriched StreetLearn with panoramas from the Touchdown study. Takedown requests that respect individuals' privacy preferences can be managed through the StreetLearn package. We encourage the research community to use only vetted and approved resources like StreetLearn, including our new release of the Touchdown panoramas, for their Street View oriented work.

The addition of Touchdown to StreetLearn (a.k.a. *Retouchdown*) boosts the total panorama count for the StreetLearn dataset from 114k to 144k. Furthermore, it contains multiple panoramas from the same neighborhoods, which supports work on learning to navigate in a region and testing in that same region using panoramas from a different time. Our code for training and evaluating vision-and-language navigation agents and spatial description resolution models are publicly available as part of the VALAN framework (Lansing et al., 2019).[4]

[4]https://github.com/google-research/valan

References

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

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*.

Howard Chen, Alane Suhr, Dipendra Misra, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Conference on Computer Vision and Pattern Recognition*.

Volkan Cirik, Yuan Zhang, and Jason Baldridge. 2018. Following formulaic map instructions in a street simulation environment. *NeurIPS Visually Grounded Interaction and Language Workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. *arXiv*.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. Association for the Advancement of Artificial Intelligence (AAAI).

Felix Hill, Karl Moritz Hermann, Phil Blunsom, and Stephen Clark. 2017. Understanding grounded language learning agents. *arXiv*.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. Effective and general evaluation for instruction conditioned navigation using dynamic time warping. *NeurIPS Visually Grounded Interaction and Language Workshop*.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of EMNLP 2020*.

Larry Lansing, Vihan Jain, Harsh Mehta, Haoshuo Huang, and Eugene Ie. 2019. Valan: Vision and language agent navigation.

Matt Macmahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, action in route instructions. In *Proceedings of AAAI*, pages 1475–1482.

Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2019. The streetlearn environment and dataset. *CoRR*, abs/1903.01292.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2018. Learning to navigate in cities without a map. *Advances in Neural Information Processing Systems*.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3D environments with visual goal prediction. In *Proceedings of EMNLP 2018*, pages 2667–2678, Brussels, Belgium.

Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of EMNLP 2017*, pages 1004–1015.

Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of the Conference on Neural Information Processing Systems*.

Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The HCRC Map Task corpus: Natural dialogue for speech recognition. In *Proceedings of the Human Language Technology Workshop*.

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.

Xin Wang, Qiuyuan Huang, Asli Çelikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin Weiss, Simon Chamorro, Roger Girgis, Margaux Luck, Samira E. Kahou, Joseph P. Cohen, Derek Nowrouzezahrai, Doina Precup, Florian Golemo, and Chris Pal. 2019. Navigation agents for the visually impaired: A sidewalk simulator and experiments. In *Proceedings of the Conference on Robot Learning*.

Jiannan Xiang, Xin Eric Wang, and William Yang Wang. 2020. Learning to stop: A simple yet effective approach to urban vision-language navigation. In *Findings of EMNLP 2020*.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2019. Vision-language navigation with self-supervised auxiliary reasoning tasks.