Multimodal Text Style Transfer fo Outdoor Vision-and-Language Navigation

Wanrong Zhu[¶], Xin Eric Wang[‡], Tsu-Jui Fu[¶], An Yan[§], Pradyumna Narayana[*], Kazoo Sone[*], Sugato Basu[*], William Yang Wang[¶]

[¶]University of California, Santa Barbara, [‡]University of California, Santa Cruz, [§]University of California, San Diego, [*]Google

Motivation

Outdoor vision-and-language navigation (VLN) is a challenging task that requires informative instructions to address the complex navigation environment. We introduces a multimodal text style transfer (MTST) learning approach and leverages external multi-modal resources to mitigate data scarcity in outdoor navigation tasks.

Multimodal Text Style Transfer Framework Overview

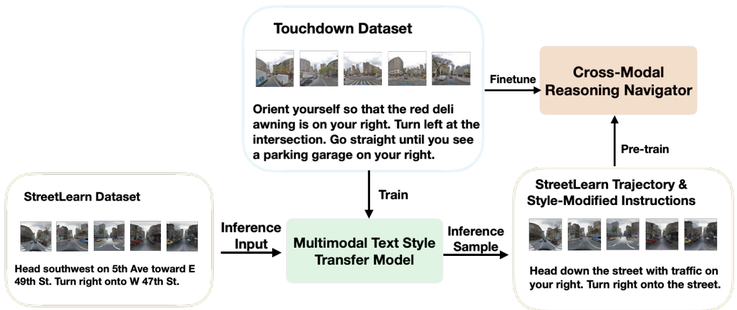

Figure 1:Our MTST learning framework mainly consists of two modules, namely the *multimodal text style transfer model* and the *VLN Transformer*.

Cross Modal Reasoning Navigator

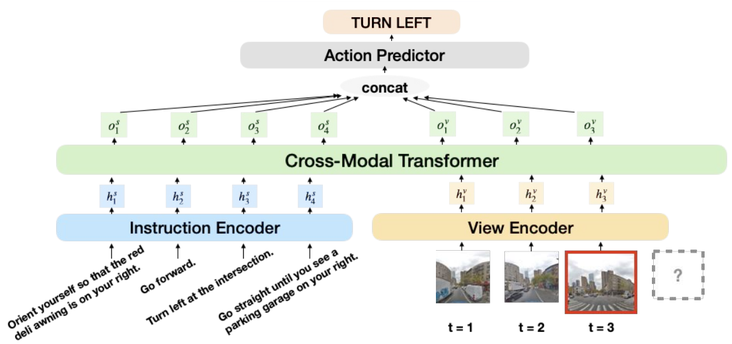

Figure 2:Overview of the VLN Transformer agent. In this example, the agent predicts to take a left turn for the visual scene at $t = 3$.

Multimodal Text Style Transfer

Training

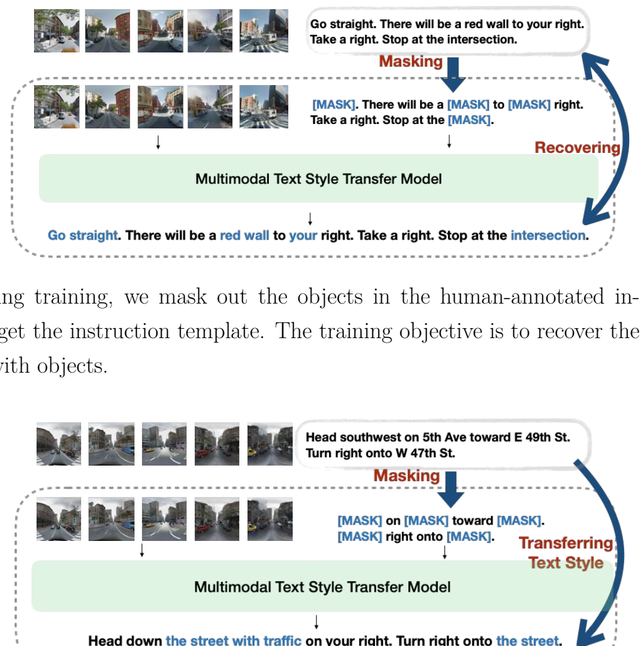

Figure 3:During training, we mask out the objects in the human-annotated instructions to get the instruction template. The training objective is to recover the instructions with objects.

Inference

Figure 4:When inferring new instructions for external trajectories, we mask the street names in the original instructions and prompt the model to generate new object-grounded instructions.

Results

In Figure 5, we compare our VLN-Transformer model with the RCONCAT and GA model as baseline, and show the evaluation results on the task completion rate (TC), success weighted by edit distance (SED) and coverage weighted by length score (CLS).
Results show that pre-training the navigation models on external data with machine-generated instructions, can partially improve navigation performance. Pre-training the navigation models on external data with our style-modfied instructions can further improve the agents' performance on metrics related to successful cases, such as TC and SED.

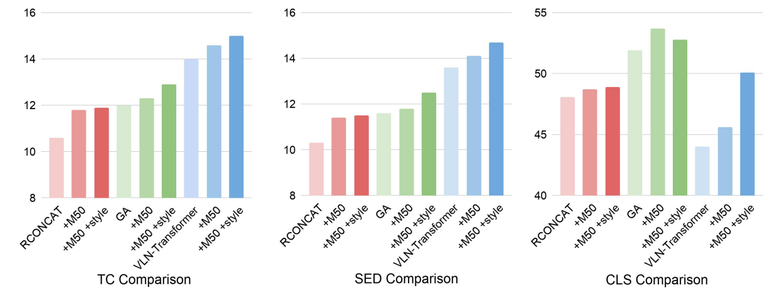

Figure 5:*+M50* denotes pre-training on a StreetLearn subset Manh-50 with machine-generated instructions, while *+M50 +style* denotes pre-training on Manh-50 with style-modified instructions. Applying our MTST approach leads to model-agnostic improvement on ourdoor VLN performance.

StreetLearn	**Turn right** onto W 36th St. **Turn right** onto Dyer Ave.
Original Speaker	Go to the next intersection and turn left again. There will be a building with a red awning on your right. Go straight through the next intersection.
Speaker with Textual Attention	Turn right at the next intersection. Stop just before the next intersection.
Multimodal Text Style Transfer	Turn right again at the next intersection. On your right will be scaffolding on your right. Turn right.

Figure 6:A showcase of the instruction generation results. Blue tokens: alignments with the ground truth. Red tokens: contradictions. Orange bounding box: the objects in the surrounding environment have been successfully injected into the style-modified instruction.

Contributions

- We present a new Multimodal Text Style Transfer learning approach to generate style-modified instructions for external resources and tackle the data scarcity issue for outdoor VLN.
- We provide the Manh-50 dataset with style-modified instructions as an auxiliary dataset for outdoor VLN training.
- We propose a novel VLN Transformer model as the navigation agent for outdoor VLN and validate its effectiveness.
- We improve the task completion rate by 8.7% relatively on Touchdown test set with the VLN Transformer model pre-trained on the external resources processed by our MTST approach.

UC SANTA BARBARA UC SANTA CRUZ UC San Diego Google