Vision-and-Language navigation (VLN) (Anderson et al. 2018): the task of following navigation instructions to traverse a path in a photorealistic environment.

VLN Instruction Generation: the task of generating natural language navigation instructions for a given path in a photorealistic environment.

- Generated Instructions have been widely adopted for data augmentation in VLN tasks and have been shown to be very effective.
- Outstanding issues and our motivation
  - If human following performance of generated instructions has never been evaluated.
  - Efficacy of automated evaluation metrics for instruction generators has not been established.

Our objectives
- Address the gaps mentioned above.
- Establish an effective metric for grounded instruction generation.

Efficacy of automated evaluation metrics for instruction generators

We also compare Speaker-Follower and EnvDrop with the following:

- Speaker-Follower (SF)
- EnvDrop
- Crafty (template-based)
- Human Instructions (newly labeled)

Perturbations:

- Direction Swap
- Phrase Swap
- Result:
  - Speaker-Follower and EnvDrop are noticeably worse than Crafty (template-based)
  - Human Instructions (newly labeled)

Compatibility Model

To build better Instruction Generators, we first need accurate automatic evaluation metrics. We propose a trajectory-instruction compatibility model to learn the alignment in a shared latent space.

The independence between the two encoders facilitates learning using both contrastive and classification losses.

Comparison of Correlation with Human Wayfinders

Our compatibility models performs the best among all the compared approaches.

<table>
<thead>
<tr>
<th>System-level (evaluating a model against many instances):</th>
</tr>
</thead>
<tbody>
<tr>
<td>○ SPICE performs the best, while other automatic metrics do not show any correlation with human wayfinder performance.</td>
</tr>
<tr>
<td>Instruction-level (evaluating an individual instruction):</td>
</tr>
<tr>
<td>○ Our compatibility models performs the best</td>
</tr>
<tr>
<td>○ Almost as good: the SPL/STWD score averaged over three VLN Agents (Followers)</td>
</tr>
<tr>
<td>○ Additional advantage: Unlike SPICE, these methods don't require reference captions!</td>
</tr>
</tbody>
</table>

Data Augmentation for VLN

- Dashed lines (green and red): use only augmented paths in training
- Dotted lines (blue and orange): use both augmented and R2Train paths.
- Each point is the mean of 3 runs and the error bars represent the standard deviation.
- Result: The model-ranked fractions show consistent improvement over random samples of the same percentage.

Conclusion

- Almost all recent VLN papers use data augmentation from an Instruction Generator (Speaker).
  - These generators have substantial headroom for improvement.
- Progress may have been hindered by a lack of suitable evaluation metrics.
- Textual evaluation metrics should not be trusted in new domains without validation.
- For navigation instructions - don't use BLEU, CIDER, METEOR or ROUGE to evaluate!
- Use SPICE for model-level evaluation.
- Use our learned compatibility model or VLN Agents for instruction-level evaluation.

Acknowledgments: We thank the Google Data Compute team, in particular Ashwin Kakarla and Priyanka Rachapally, for their tooling and annotation support for this project.