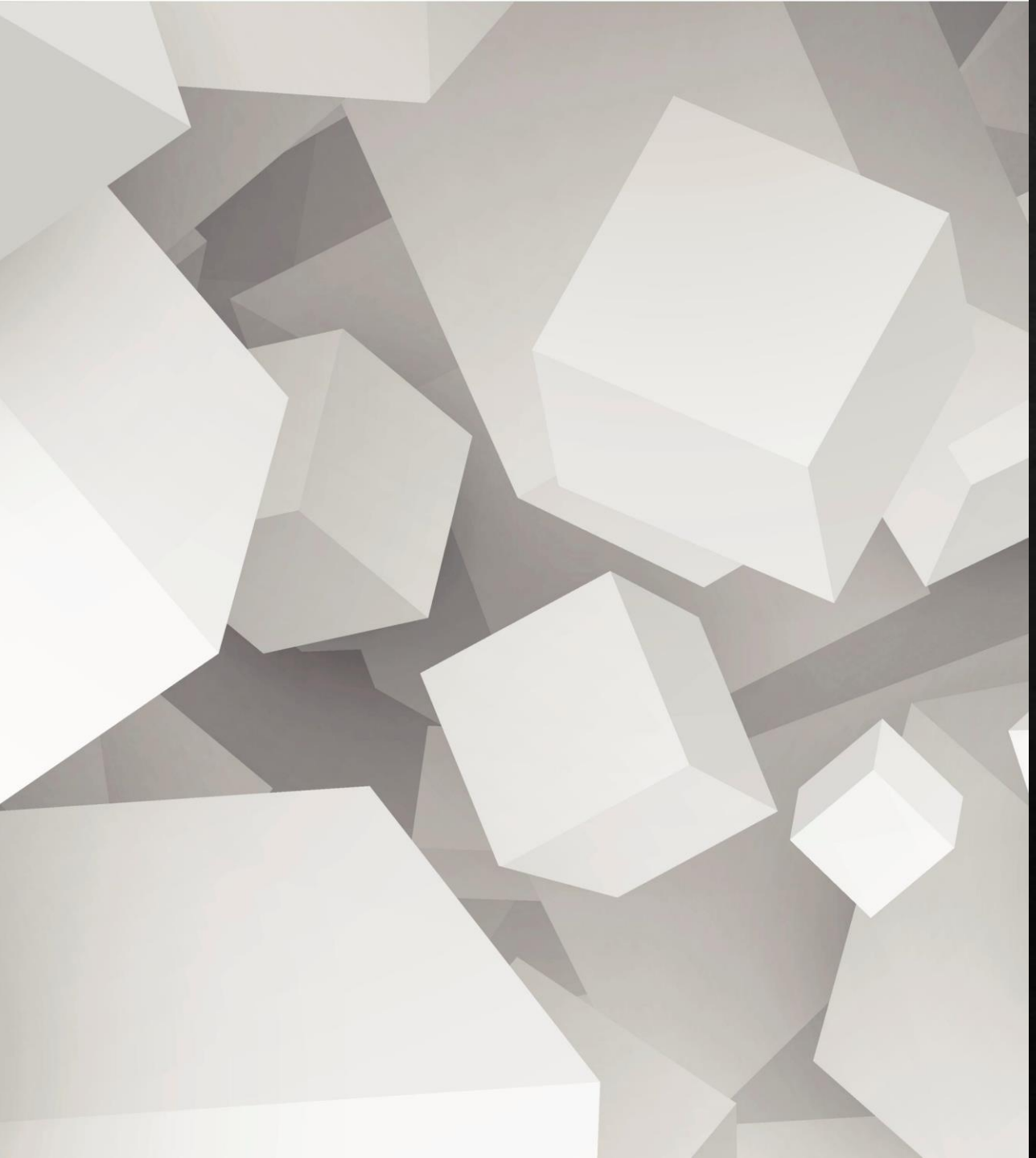
Syntactic Aware Cross Modality Alignment for Vision Language Navigation

Sihui Wang

April 6th, 2023



Overview

Problem Statement

Task and Data

Approach

Results

Analysis

Conclusion

Problem Statement (I)

- What is Visual Language Navigation?
  - An embodied AI task which requires an agent to navigate 3D environments through following language instructions

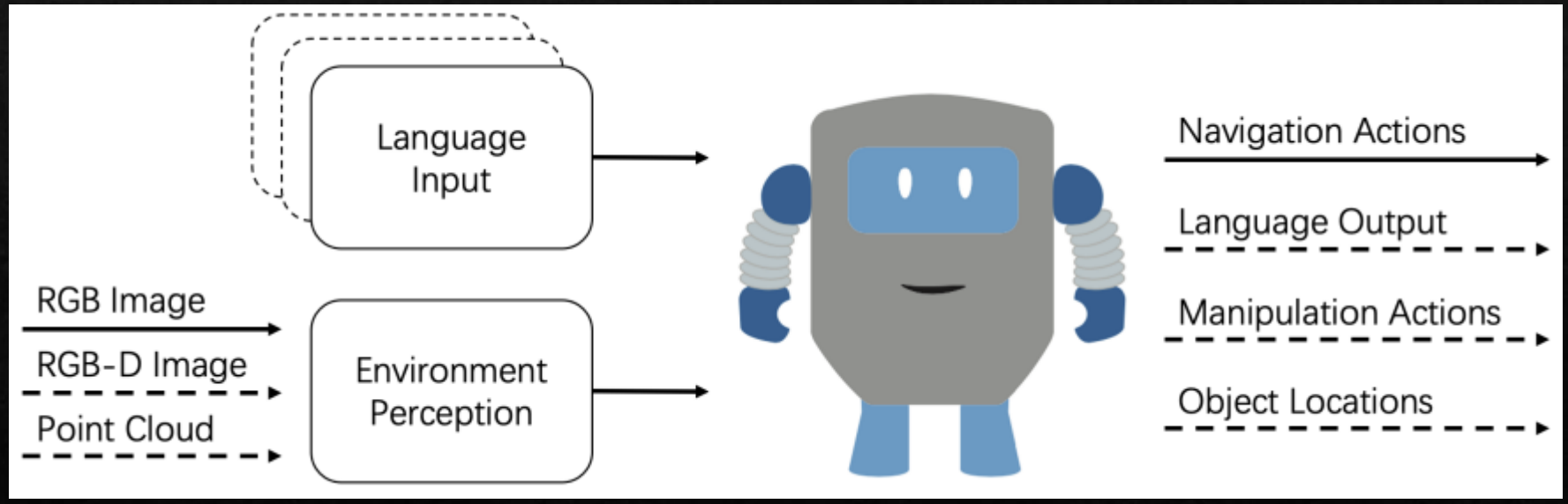


Image from: Wu, Wansen, Tao Chang, and Xinmeng Li. "Vision-Language Navigation: A Survey and Taxonomy." *arXiv e-prints* (2021): arXiv-2108.

Problem Statement (II)

- Our Goal: Improve cross modality alignment by incorporation of syntactic information
- Motivation:
  - Identify important words
  - Extract sub-instructions
  - Better alignment with visual clues and actions

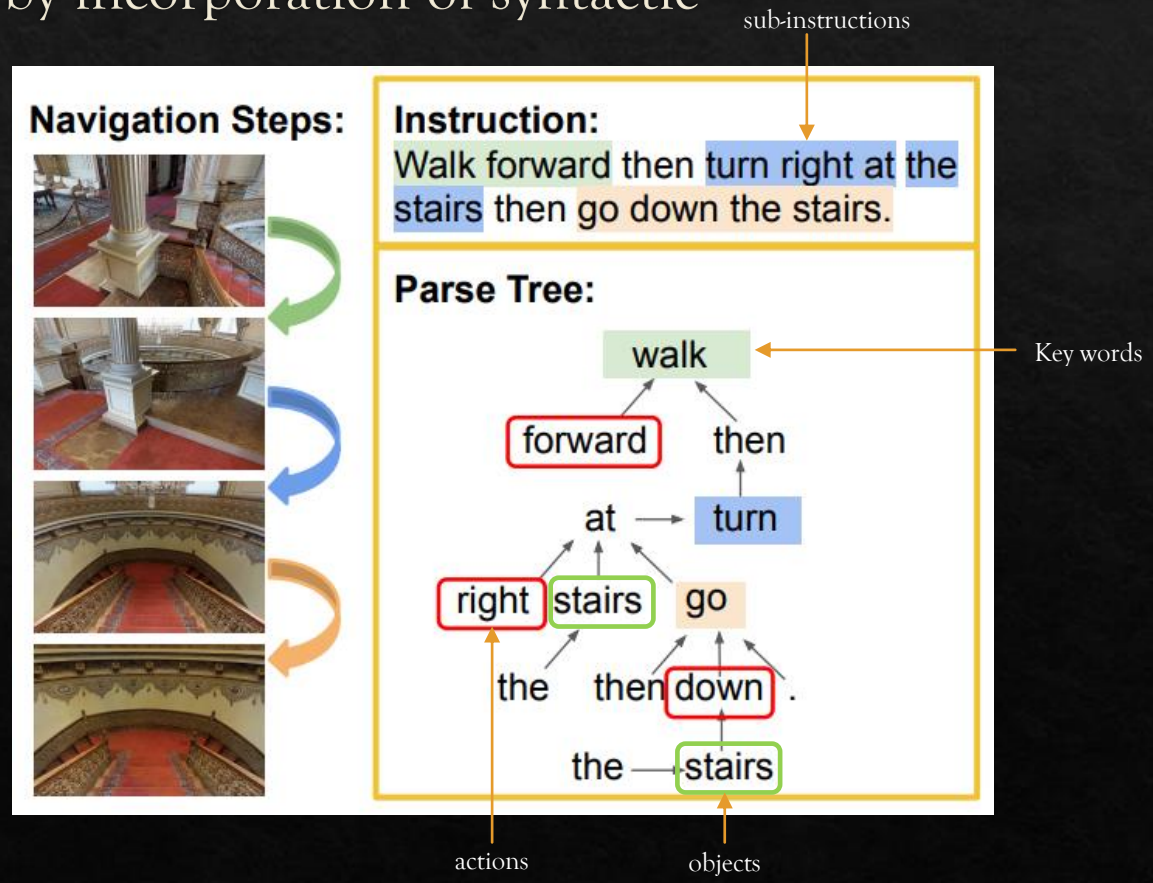
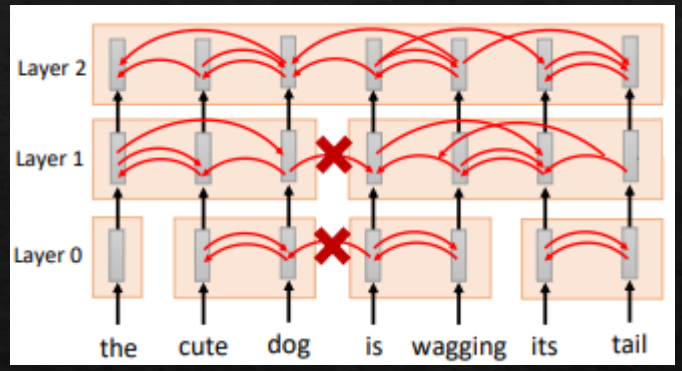


Image adapted from: Li, Jialu, Hao Tan, and Mohit Bansal. "Improving Cross-Modal Alignment in Vision Language Navigation via Syntactic Information." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1041-1050. 2021.

Problem Statement (III)

- Previous Approach: (Syntactic LSTM)
  - LSTM-based, slow to train, low performance
  - Learn syntactic information by supervised learning
- Our Approach: (Syntactic Transformer)
  - Transformer-based, faster to train, high performance
  - Learn syntactic information by un-supervised learning
  - Key Insight: Use local-to-global attention to provide the hierarchical structure for better representation of language



Local to global attention in Transformers

Image adapted from: Wang, Yaushian, Hung-Yi Lee, and Yun-Nung Chen. "Tree Transformer: Integrating Tree Structures into Self-Attention." 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)*, pp. 1061-1070. 2019.

Task

- Input: instruction texts (and their parse trees), visual features, connectivity graph
- Output: agent's actions and path
- Task: Given observations of the environment, generate a sequence of actions (navigation) following the instructions.

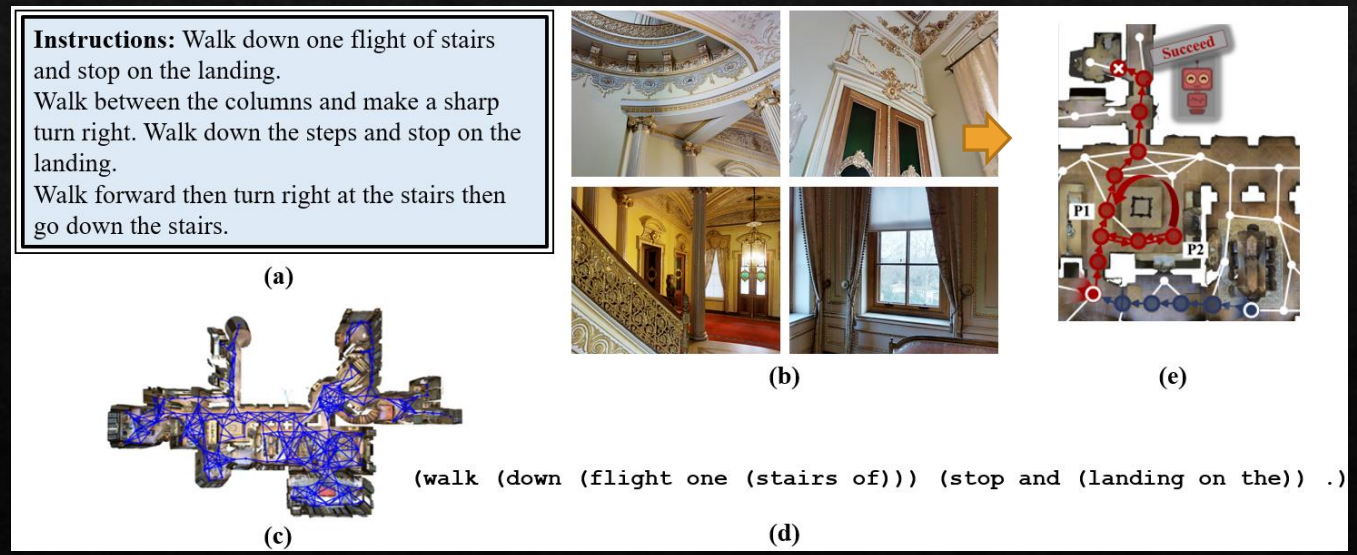


Image (c) from: Anderson, Peter, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674-3683. 2018.
Image (e) from: Wang, Hanqing, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. "Structured scene memory for vision-language navigation." In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 8455-8464. 2021.

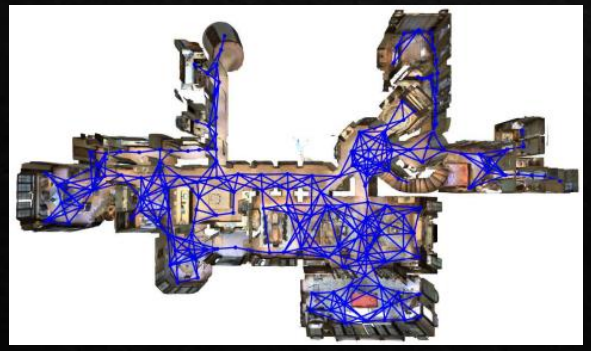
Data (I)

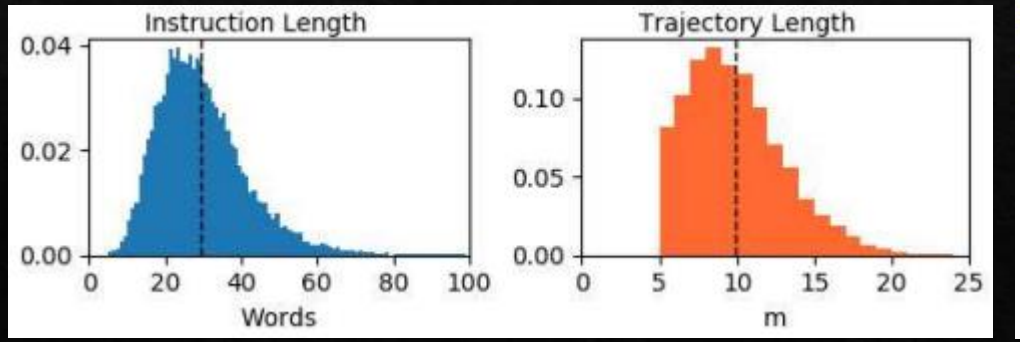
- ImageNet and Places365 Image Features from Matterport3D dataset
  - Matterport3D: 10,800 panoramic views from 194,000 RGB-D images; 90 building - scale scenes (61 – training, 11 – validation, 18 – testing)

Image Examples from Matterport3D

Data (II)

- R2R (Room-to-Room) Dataset: instructions, connectivity graphs, reference navigation trajectories
  - 21,567 navigation instructions (14,025 – training, 1,020 – validate seen, 2,349 – validate unseen, 4,173 – testing)
  - Instruction Average Length: 29 words
- Augmented Instructions: 1,069,620 instructions





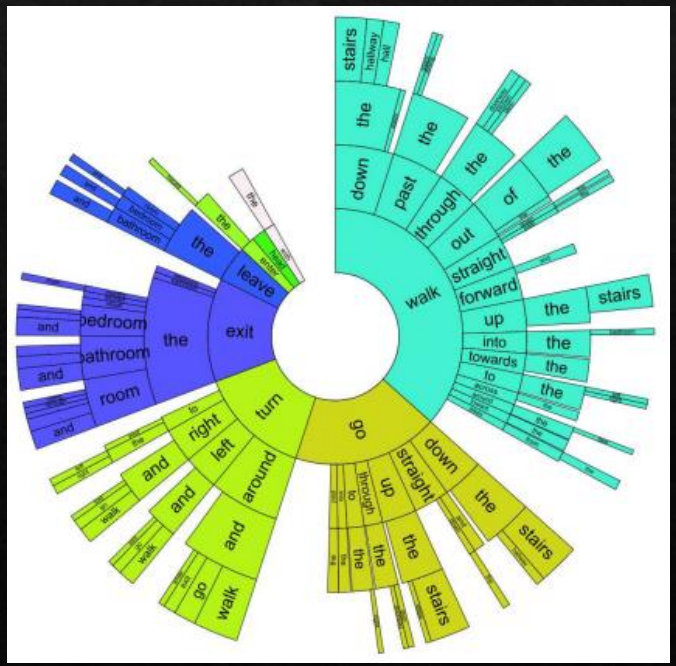


Image from: Anderson, Peter, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674-3683. 2018.

Approach (I)

- LSTM with Syntactic Information

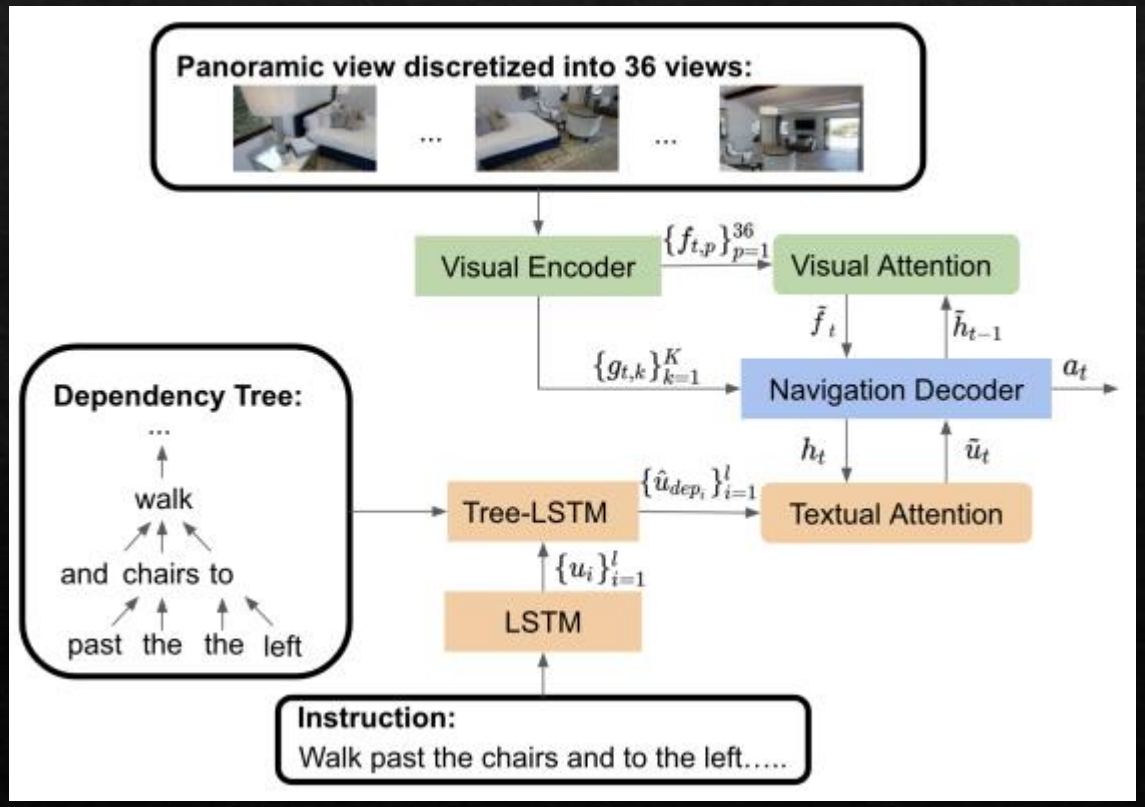


Image from: Li, Jialu, Hao Tan, and Mohit Bansal. "Improving Cross-Modal Alignment in Vision Language Navigation via Syntactic Information." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1041-1050. 2021.

Approach (II)

- Recurrent VLN Bert with Syntactic Information
  - Neighbor Attention
  - Constituent Prior
  - Hierarchical Constraints

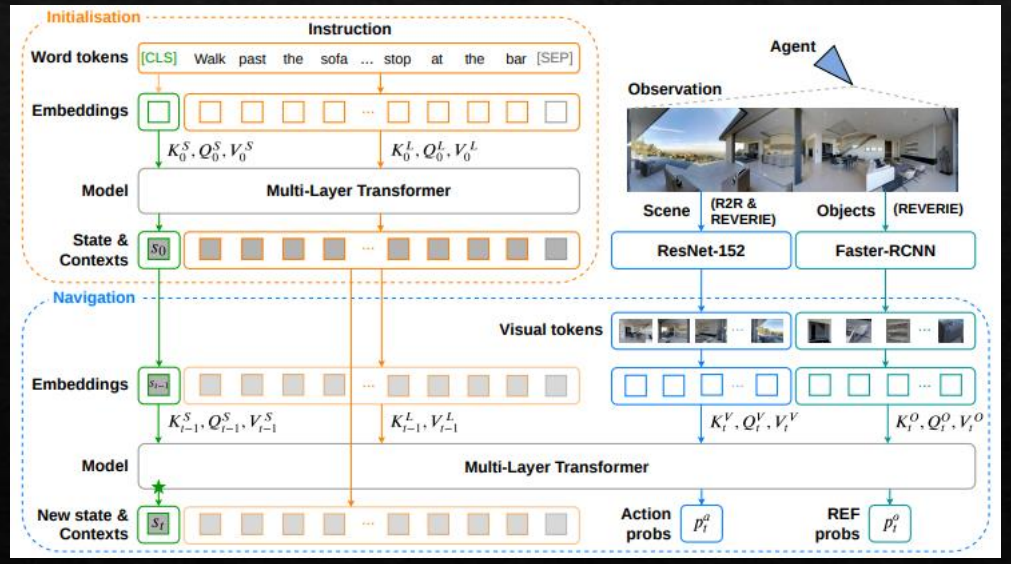


Image from: Hong, Yicong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. "Vln bert: A recurrent vision-and-language bert for navigation." In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 1643-1653. 2021.

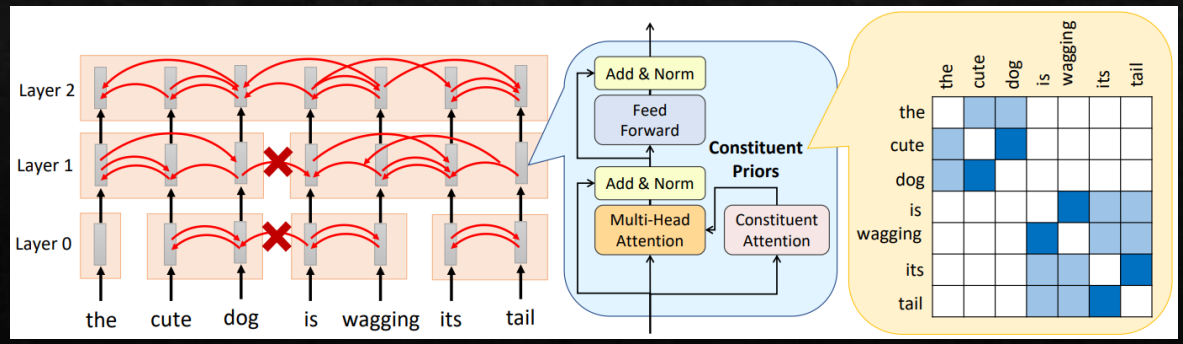


Image from: Wang, Yaushian, Hung-Yi Lee, and Yun-Nung Chen. "Tree Transformer: Integrating Tree Structures into Self-Attention." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1061-1070. 2019.

Experiments & Evaluation

- Experiments:
  - i) Compare LSTM – Baseline Model with LSTM – Syntactic Model
  - ii) Compare Transformer – Baseline Model with Transformer – Syntactic Model
- Evaluation Metrics:
  - Success Rate (SR): if the agent stops within 3m from the target location
  - Success Rate weighted by Path Length (SPL): SR with length penalty
  - Normalized Dynamic Time Warping (nDTW): penalize deviation from ground truth path; higher is better
  - Success Rate Weighted by Dynamic Time Warping (sDTW): nDTW computed for successful navigation examples; higher is better
  - Coverage Weighted by Length Score (CLS): encourage path fidelity; higher is better

Results (I)

- LSTM-Baseline VS LSTM-Syntactic

On R2R Validation – Seen Dataset

Models	SR(%)	SPL	nDTW	sDTW	CLS
LSTM-Baseline	**0.610**	**0.583**	**0.696**	**0.548**	**0.692**
LSTM-Syntactic	0.578	0.547	0.679	0.522	0.677

On R2R Validation – Unseen Dataset

Models	SR(%)	SPL	nDTW	sDTW	CLS
LSTM-Baseline	**0.476**	**0.443**	**0.587**	**0.414**	**0.587**
LSTM-Syntactic	0.463	0.430	0.586	0.403	0.586

Contrary to the original paper, we observed that syntactic – LSTM performs worse than the baseline on validation datasets.

Results (II)

- Transformer-Baseline VS Transformer – Syntactic

On R2R Validation – Seen Dataset

Models	SR(%)	SPL
Transformer-Baseline	0.746	**0.700**
Transformer-Syntactic	**0.750**	0.699

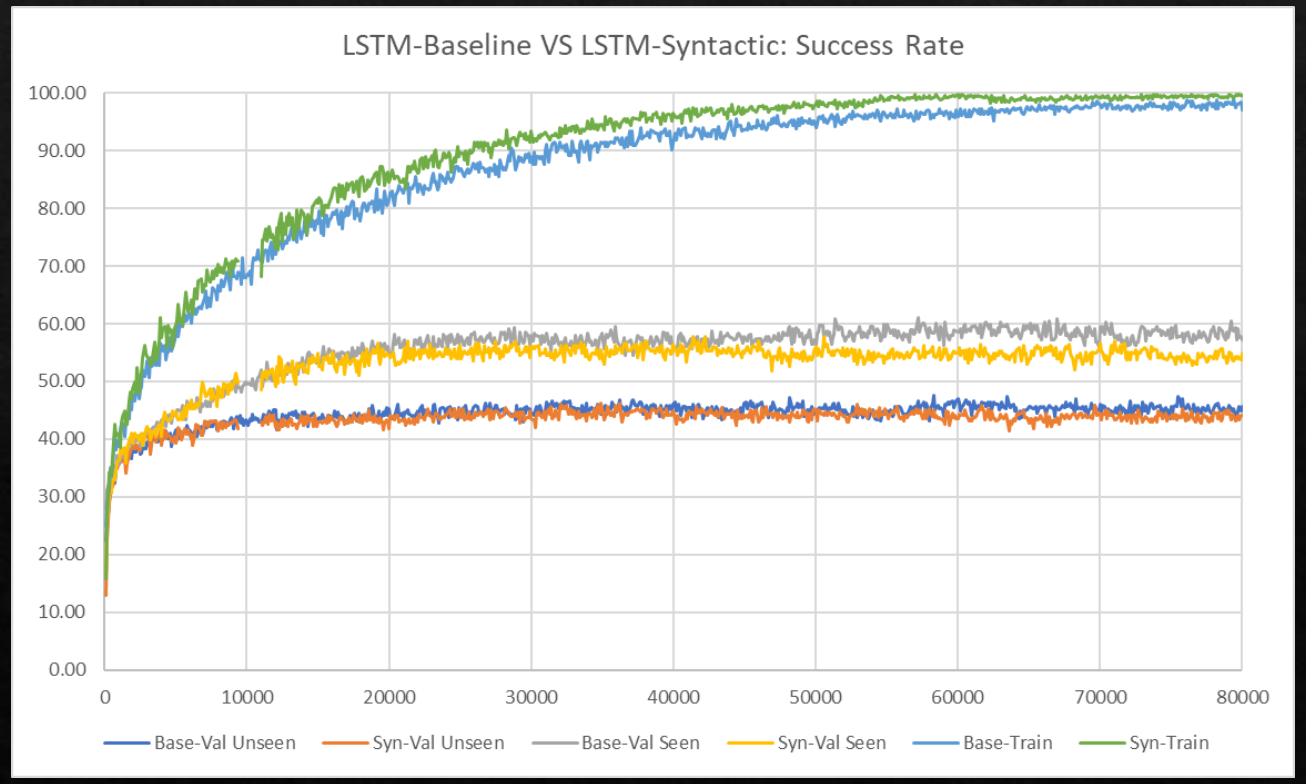
On R2R Validation – Unseen Dataset

Models	SR(%)	SPL
Transformer-Baseline	0.613	**0.556**
Transformer-Syntactic	**0.622**	0.549

Syntactic – Transformer tends to outperform the baseline transformer model in success rate.
However, syntactic – transformer generates longer paths.

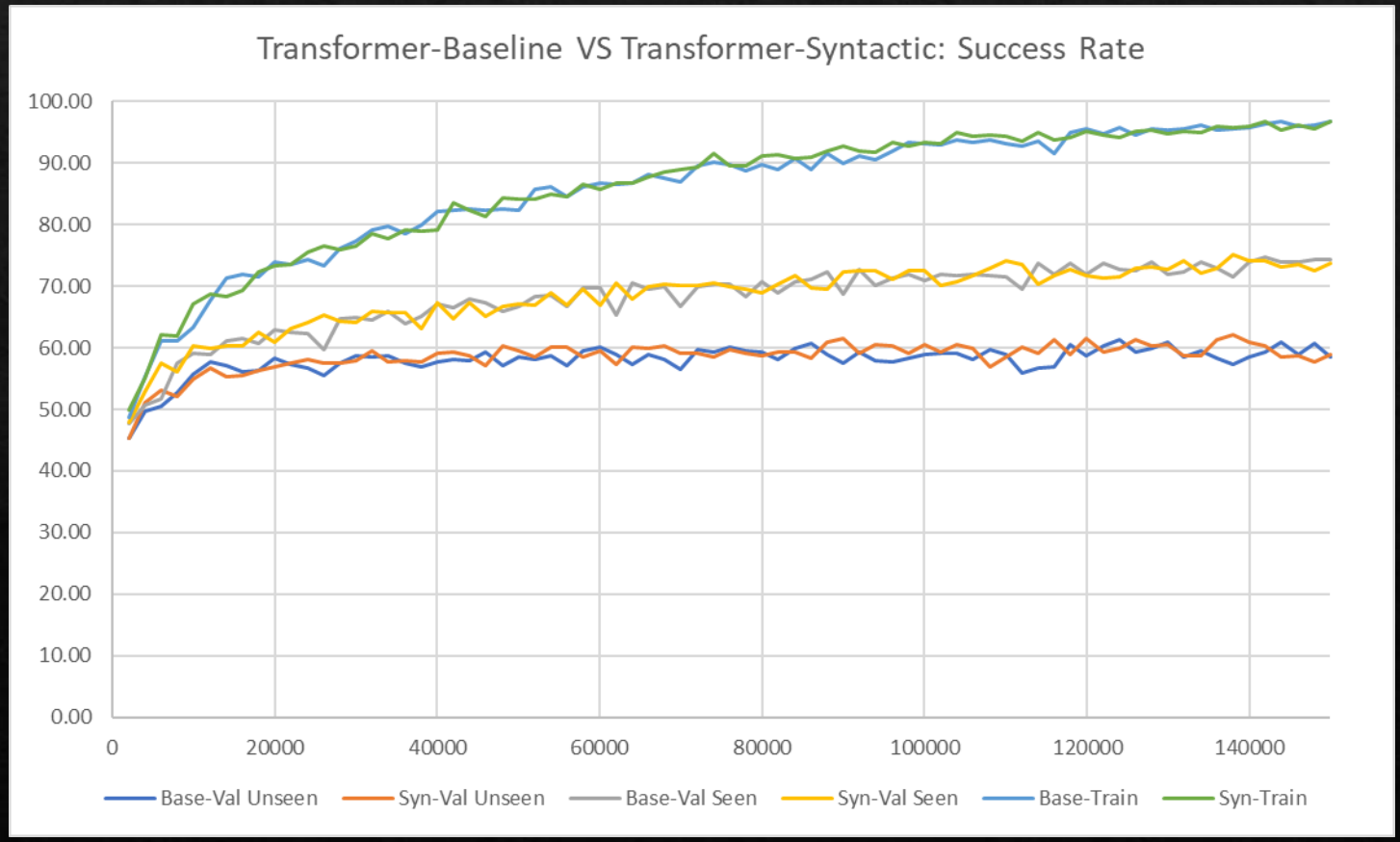
Analysis (I)

Tree – LSTM leads to larger train – validation gap: indication of poor generalization abilities



Analysis (II)

Syntactic – Transformer might be slightly better than the baseline model ? Need to replicate the experiments for multiple times.



Conclusion

- **Contrary to** the original paper, we find that LSTM-syntactic model performs worse than the baseline model; Larger train-validation gap indicates that LSTM-syntactic model doesn't generalize well to new data;
- Syntactic-Transformer model outperforms the baseline by +0.9% of success rate on unseen validation dataset; however, syntactic-transformer model tends to generate longer paths.

Future Work

- Replication of Experiments
- Tree LSTM on model generalization
- Further Analysis on Self Attention and Cross-Modality Attention
- Tree Transformer's Potential Impact:
  - Higher Dropout Rate
  - Local Attention

Reference

- Anderson, Peter, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka et al. "On evaluation of embodied navigation agents." arXiv preprint arXiv:1807.06757 (2018).
- Anderson, Peter, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian 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, pp. 3674-3683. 2018.
- Chang, Angel, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. "Matterport3d: Learning from rgb-d data in indoor environments." arXiv preprint arXiv:1709.06158 (2017).
- Hong, Yicong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. "Vln bert: A recurrent vision-and-language bert for navigation." In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 1643-1653. 2021.
- Ilharco, Gabriel, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. "General evaluation for instruction conditioned navigation using dynamic time warping." arXiv preprint arXiv:1907.05446 (2019).
- Li, Jialu, Hao Tan, and Mohit Bansal. "Improving Cross-Modal Alignment in Vision Language Navigation via Syntactic Information." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1041-1050. 2021.
- Tan, Hao, Licheng Yu, and Mohit Bansal. "Learning to navigate unseen environments: Back translation with environmental dropout." arXiv preprint arXiv:1904.04195 (2019).
- Wang, Yaushian, Hung-Yi Lee, and Yun-Nung Chen. "Tree Transformer: Integrating Tree Structures into Self-Attention." 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), pp. 1061-1070. 2019.