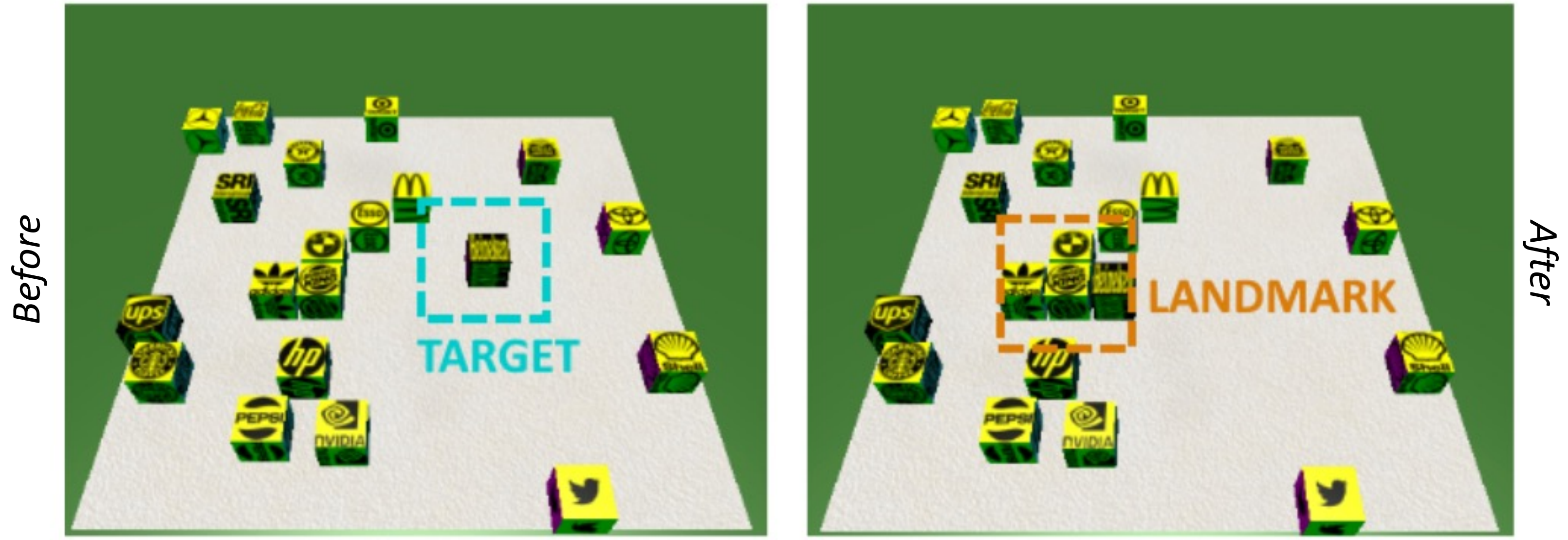
Generating Landmark-based Manipulation Instructions from Image Pairs

Sina Zarrieß[1], Henrik Voigt[1], David Schlangen[2] and Philipp Sadler[2]

[1]University of Bielefeld first.last@uni-bielefeld.de [2]University of Potsdam first.last@uni-potsdam.de

Introduction

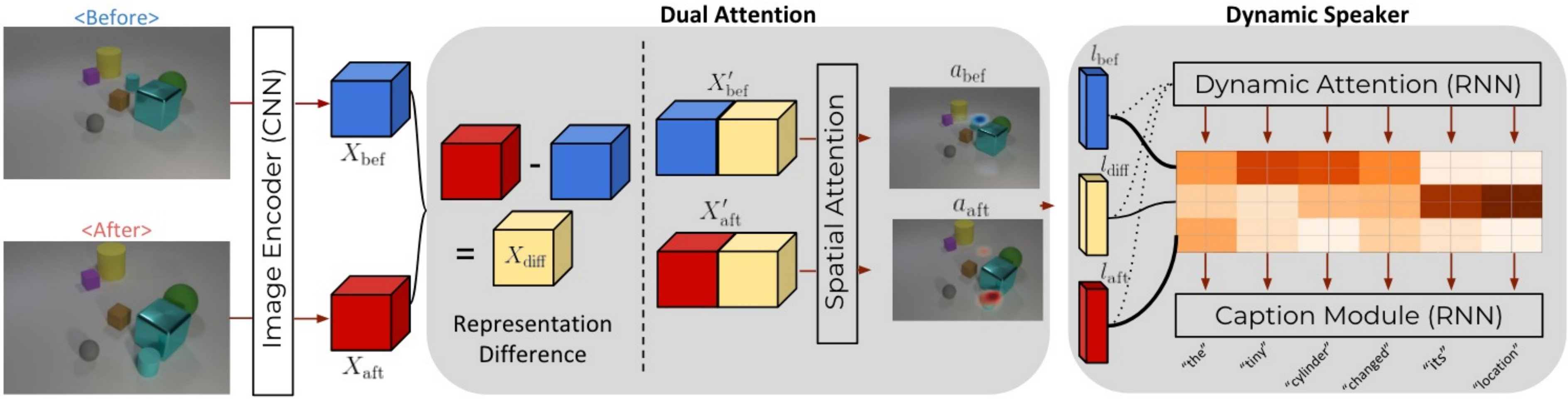
How to generate correct landmark references in manipulation instructions from image pairs?



GT: "Place the **Heineken** box so that it touches the **Burger King** box on the right side"

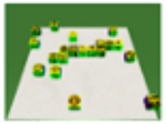
Models: DUDA

➢ Park et al. (2019) used a Dual Dynamic Attention Model (DUDA) to articulate changes in image pairs

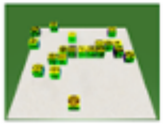


Models: Self-Attention

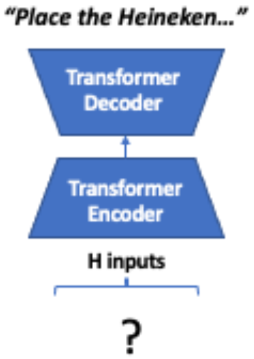
How can we feed the images as useful inputs to a transformer?



BEFORE

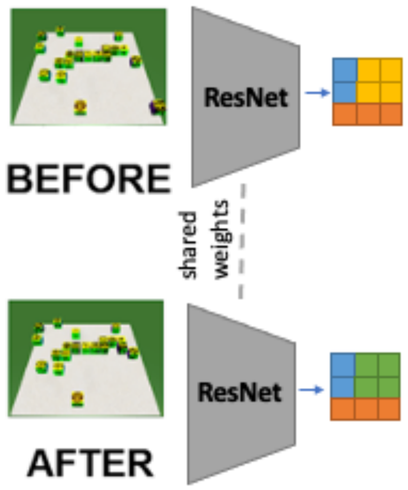


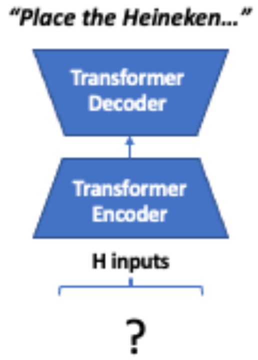
AFTER



Models: Self-Attention

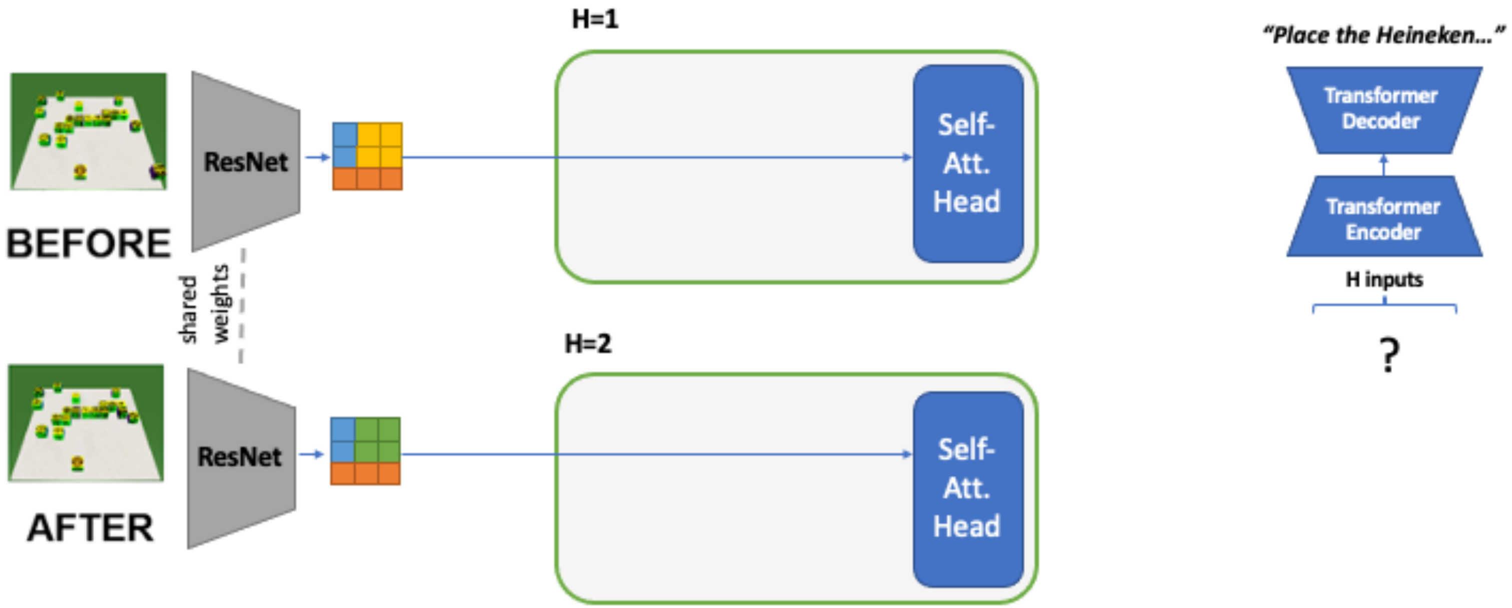
How can we feed the images as useful inputs to a transformer?





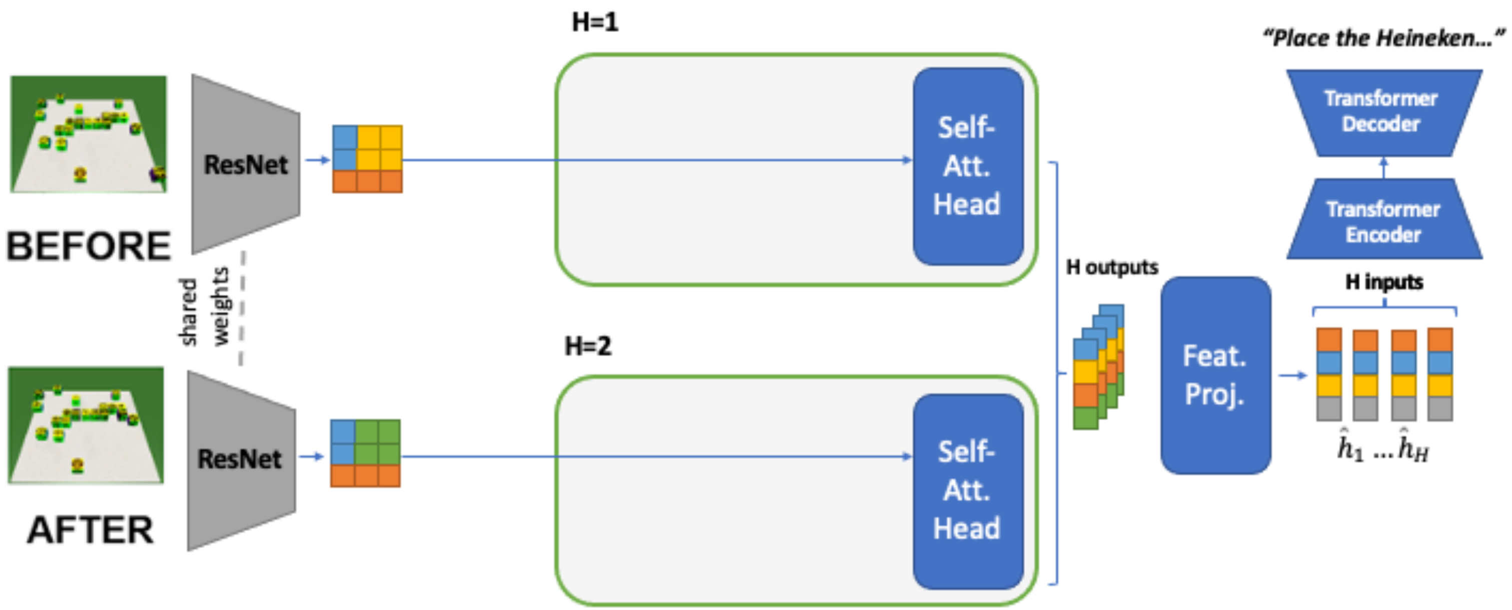
Models: Self-Attention

How can we feed the images as useful inputs to a transformer?



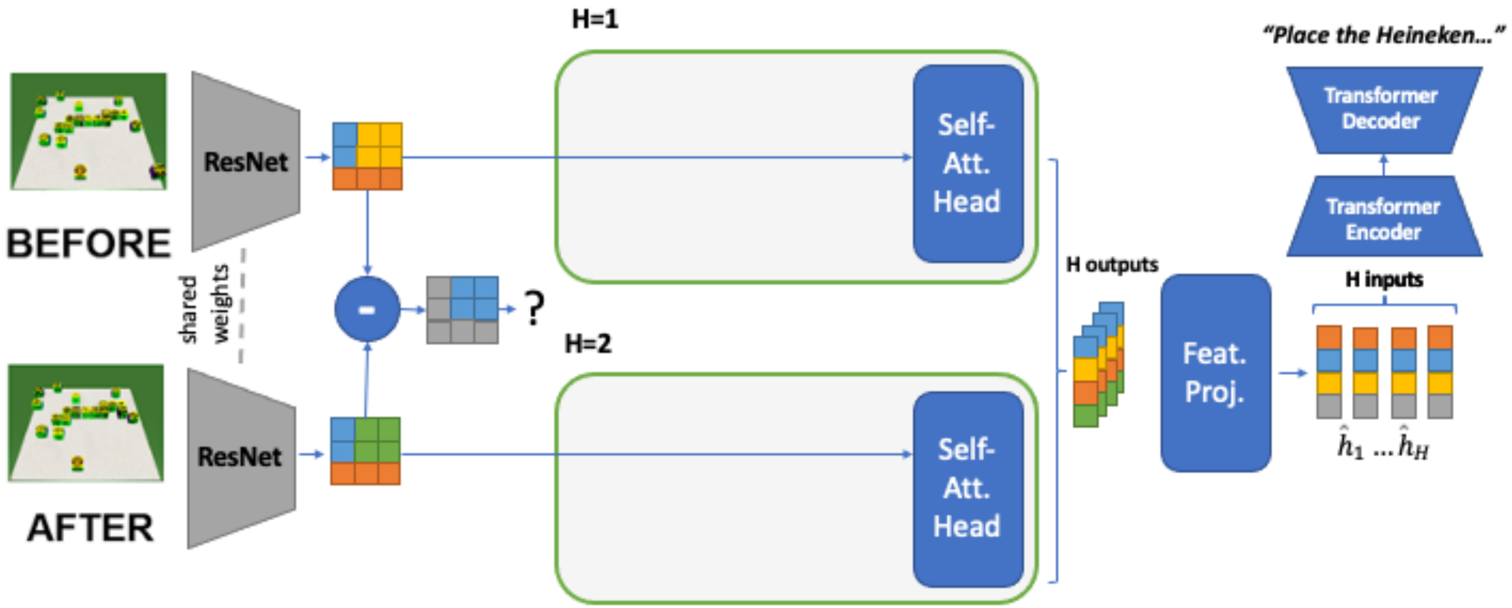
Models: Self-Attention

How can we feed the images as useful inputs to a transformer?



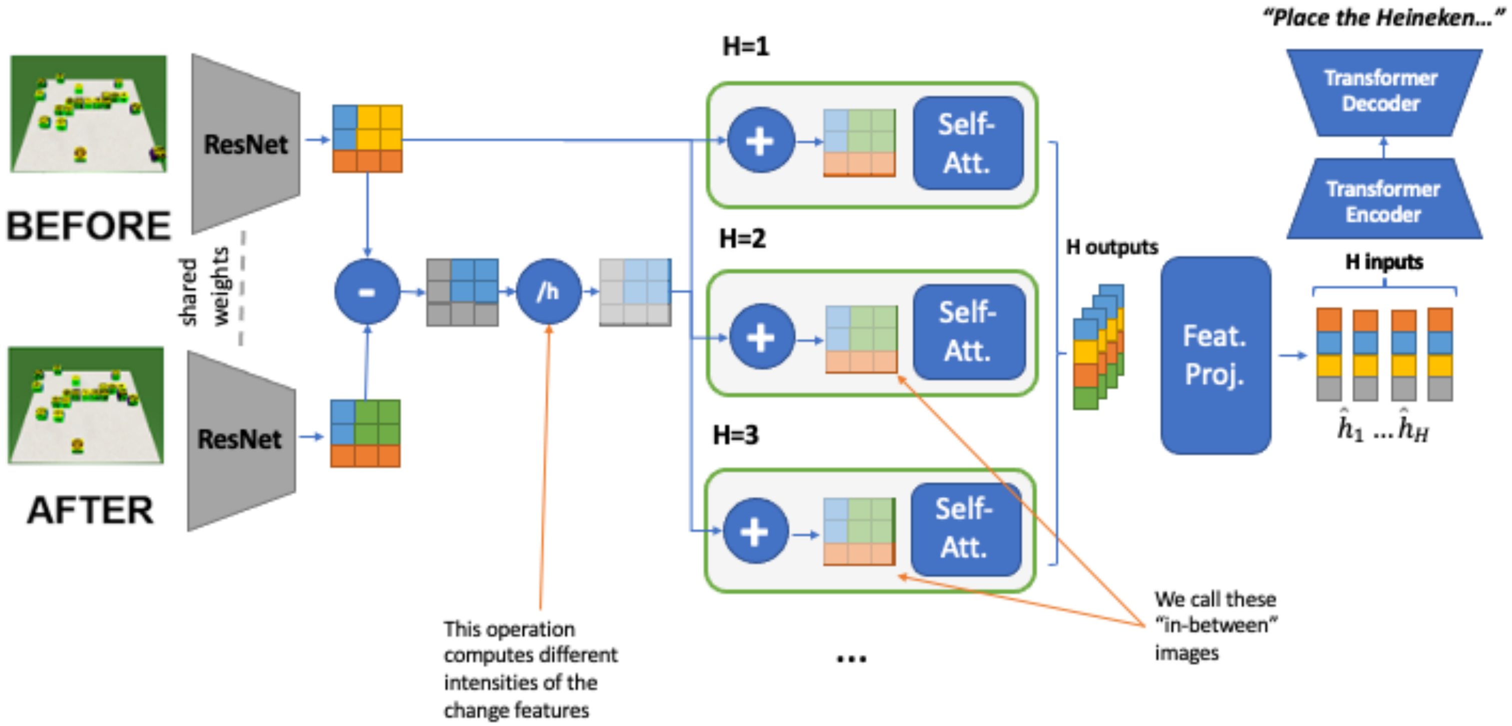
Models: Self-Attention

How can we use the „change" features for self-attention heads?



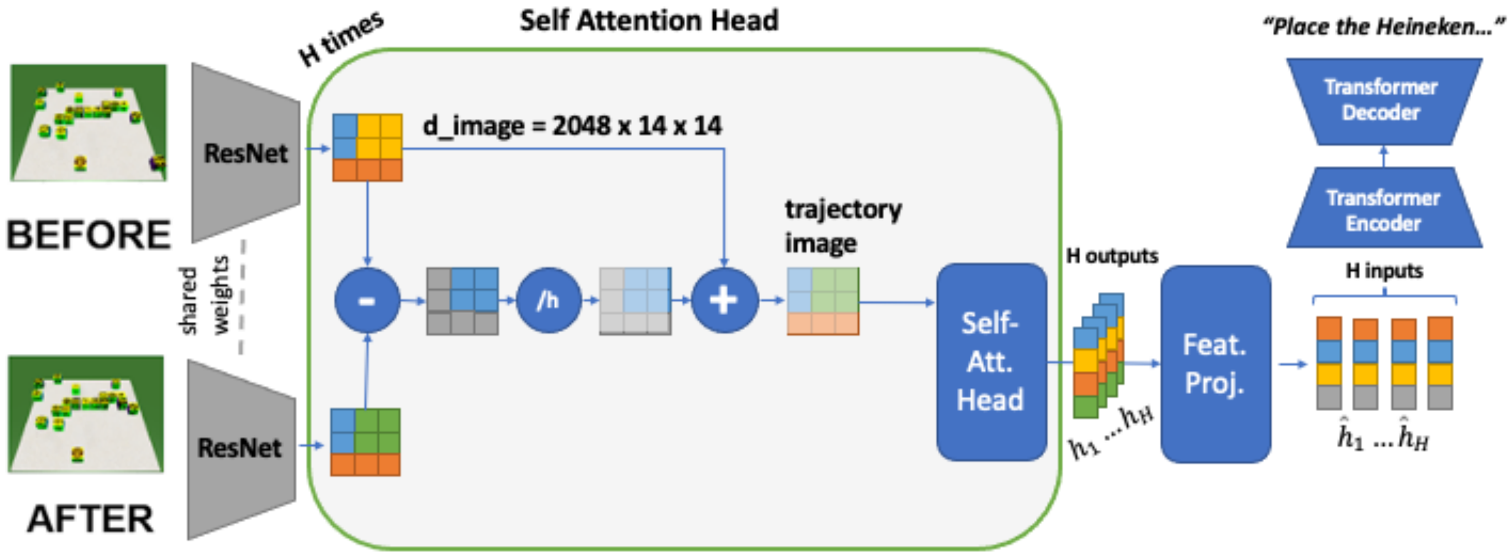
Models: Self-Attention

How can we use the „change" features for self-attention heads?



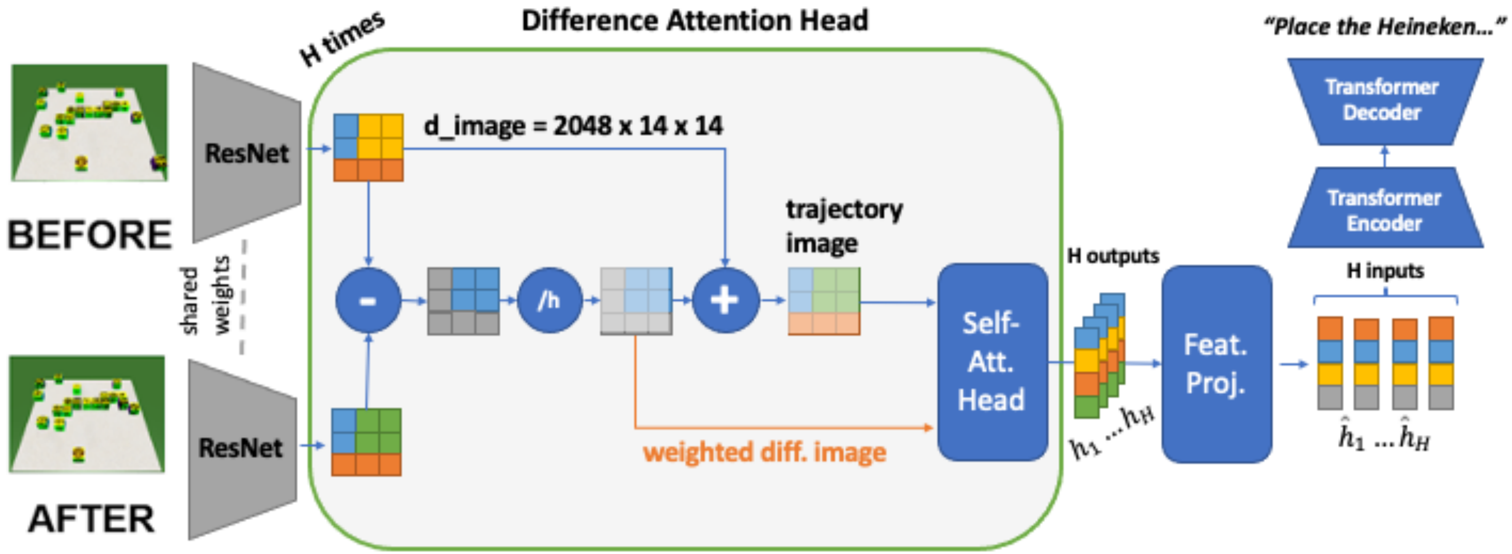
Models: Self-Attention

How can we use the „change“ features for self-attention heads?



Models: Difference(-guided)-Attention

Is cross-attention a powerful application here?



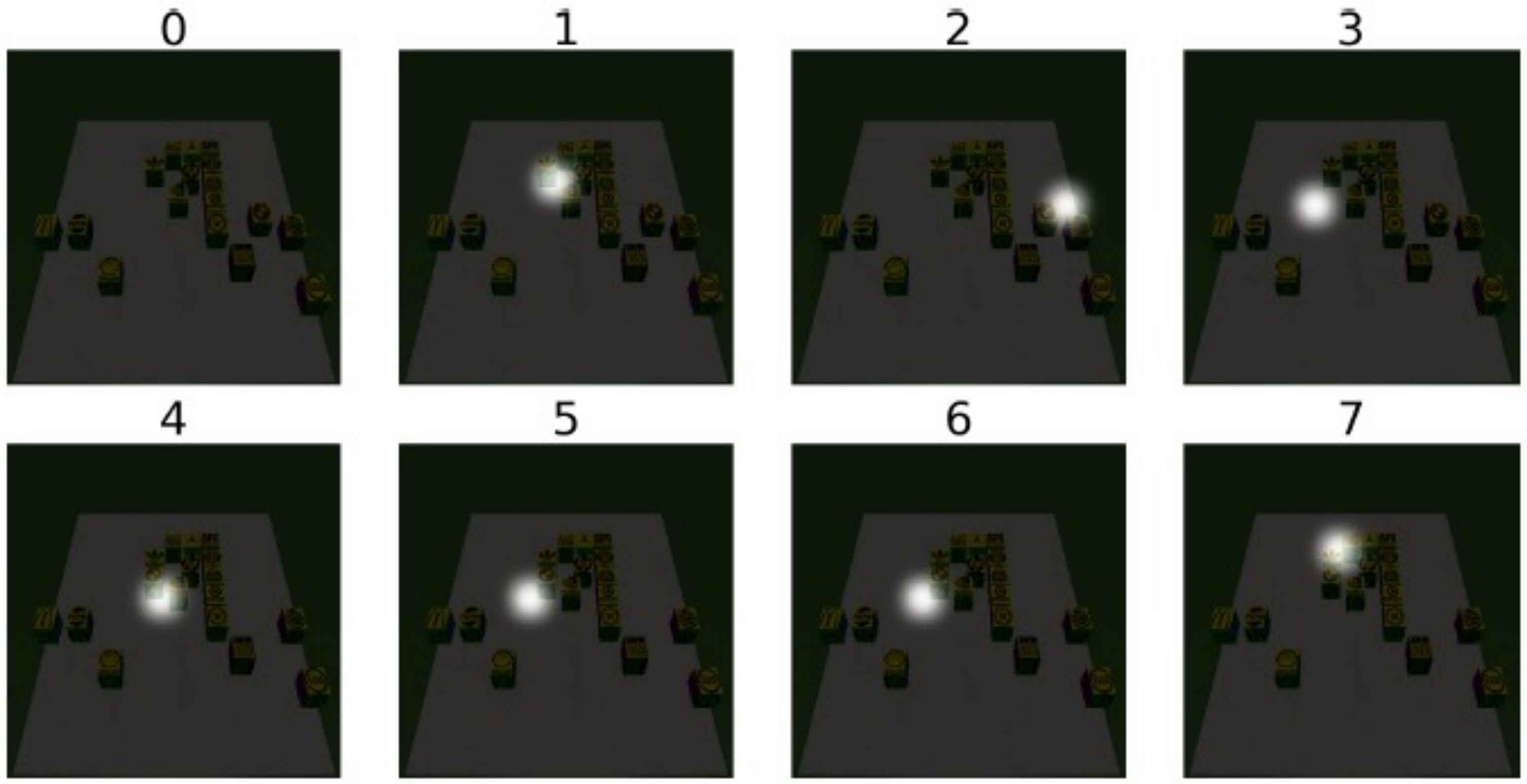
Results and Discussion

- We observe that difference attention with "in-between images" gives a very clear performance boost for the realization of landmark references

Model	B	M	C	Target	Landm	Spatial
LSTM+Att*	0.38	0.28	0.27	0.11	0.28	-
DUDA	0.53	0.37	0.96	0.59	0.42	0.66
TF-self-att-2	0.34	0.28	0.35	0.19	0.26	0.76
TF-self-att-8	0.44	0.32	0.66	0.37	0.45	0.72
TF-diff-att-2	0.55	0.38	1.06	0.73	0.40	0.80
TF-diff-att-8	**0.68**	**0.43**	**1.52**	**0.86**	**0.73**	**0.83**

Table 1: BLOCKS results: B(LEU-4), M(eteor), C(ider) and word accuracies (see Section 3.3), LSTM+Att* as reported in Rojowiec et al. (2020).

Example Attention for TF-diff-att-8



Additional Results

- Jhamtani and Berg-Kirkpatrick (2018) took surveillance images to detect and articulate changes in images

„4 additional people are present in after photo"

Additional Results

- the differences between models on Spot-the-diff are generally much smaller but our model performs best

Model	B	M	C	S
DUDA*	0.081	0.115	0.34	-
FCC*	0.099	0.129	0.368	-
SDCM*	0.098	0.127	0.363	-
DDLA*	0.085	0.12	0.328	-
M-VAM + RAF*	0.111	0.129	0.425	0.171
TF-self-att-2	0.109	0.135	0.777	0.197
TF-self-att-8	0.110	0.136	0.786	0.191
TF-diff-att-2	**0.117**	**0.137**	**0.843**	**0.205**
TF-diff-att-8	0.113	0.136	0.842	0.202

Table 2: Spot-the-diff results: B(LEU-4), M(eteor), C(IDEr), S(PICE). *Models as reported in Shi et al. (2020)

Conclusion

- difference attention heads help transformers greatly to produce landmark based manipulation instructions
- the results are in line with other approaches (Herdade et al. 2019, Park et al. 2019, Cornia et al. 2020)
- n-gram overlap metrics can be only an auxiliary measure for instruction generation

Thanks for listening!

References

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. *Natural language communication with robots.* Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. *Learning to describe differences between pairs of similar images.* In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. *Meshed-memory transformer for image captioning.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10578–10587.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. *Image captioning: Transforming objects into words.* In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Robin Rojowiec, Jana Götze, Philipp Sadler, Henrik Voigt, Sina Zarrieß, and David Schlangen. 2020. *From "before" to "after": Generating natural language instructions from image pairs in a simple visual domain.* In Proceedings of the 13th International Conference on Natural Language Generation, pages 316–326, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need.* In Advances in neural information processing systems, pages 5998–6008.