Reaction Report for "SceneFormer: Indoor Scene Generation with Transformers"

Nishant Raj

What I like about this paper?

3D scene generation by traditional methods generally use internal representation of scenes like graph, 2D images etc. and make assumptions about the possible relationships between different objects. These assumptions may require heavy optimization as well as post-processing steps. Since this scene information is sequential in nature, the idea that a series of transformers can learn these object relations (location and orientation) without requiring any additional information in sequences in 3D using their multi-headed self-attention capabilities is very interesting. This avoids any manual bias involved in comparison to traditional methods. Another idea that I liked was the use of text description of scenes which can also be leveraged to build conditional models using the cross-attention mechanism of decoder in transformers. The paper also highlighted the fact that transformer architecture is extremely fast compared to other existing baselines which proves that this could be an active area of research for different downstream tasks in vision tasks like visual common-sense reasoning, autonomous car parking in a parking lot, augmented reality etc.). The concept of using 3D IOU for 3D CAD object insertion to avoid collision also seems to be a good approach. Authors have also performed sufficient ablation studies to give a solid basis to their research work. One of the major things that they were able to accomplish in comparison to previous works was that they object relations were learned very effectively like TV on walls, bed stands on side etc.

What I do not like about this paper? One of the issues that I felt while reading the paper was lack of a good evaluation metric. While it is difficult to come up with a robust metric for generative tasks, using perceptual study with a sample population size of 30 users through Amazon Mechanical Turk is simply unreliable even statistically. Second issue that I felt while reading the paper was that the since the model architecture followed a layer approach i.e., output of next step is conditioned upon the previous step. Though this would be helpful in most cases, still there are high chances of error amplification effect from one step to another (errors from a single step would flow into the next step) even though conditioning is used to guide the outcome only (a highly misinformed prior can lead to a poor posterior).

Future Directions: One of the points mentioned in the paper was that a single transformer with similar capacity fails to generate equivalent good scenes. This seems a bit counterintuitive to me as using a single transformer with similar capacity and shared weights across different tasks can help to minimize error amplification effect. This is something that might be worth investigating. The two user inputs in the paper have been conditioned separately. A joint conditioning on both the room layout and text (shared weights point mentioned in previous line) might be able to produce better scenes than the current architecture. One of the other ideas that has been mentioned in the paper as well is to make use of 3D mesh of each object. This can help to maintain a global style consistency.

---------- "I completed the Forward Focus survey." ----------