Reaction Report – 5
Rushikesh Dudhat

What I like about this paper: Based on the idea that a scene can be treated as a sequence of objects, authors have proposed SceneFormer, a series of transformers that autoregressively predict the class category, location, orientation, and size of each object in a scene. The authors have shown that such a model generates realistic and diverse scenes while requiring little domain knowledge or data preparation and without the use of any visual information. This paper uses the cross-attention mechanism of the Transformer decoder to build conditional models, which differ from the traditional translation task as it just guides the scene generation, thus making it independent of annotators. This avoids the model to avoid any bias introduced by the manual selection of the relations or the heuristics used to create these relations. SceneFormer can generate high-quality complex scenes with a large number of objects without requiring user input, but the model is flexible enough to accept the user input if desired. The authors have made the baseline comparison robustly, and it is impressive to see that SceneFormer is more accurate and complex in generating output and the model also outperforms the previous state-of-the-art models like FastSynth in terms of computation time.

What I don't like about the paper: We see that 3D world modeling requires the priors to be very strong since this paper extracts the information from the dataset without taking explicit annotations. The input dataset needs to be robust enough for making good predictions. Additionally, the authors have not mentioned the subcategories within the classes. For instance, there can be multiple types of lights, like the one placed at the center of the ceiling or on walls. Thus, there will be constraints on the placement of such objects depending on the subclasses which need to be learned by the model. Additionally, greedily soring the classes based on the observed frequency in the training dataset might lead to a sub-optimal solution. Also, the complexity analysis of the input text is not done, i.e., how long the user statements can be? What is the loss between input text and transformer output?

Future works: In this paper, the room layout and text conditioning are done separately. It would be interesting to see the scene generation jointly done on both inputs. Also, in the paper, the model leverages the text data in the sentences that only describe object class and their spatial relations and generates the scenes using the category and location model. It would be interesting to see the performance by adding the orientation model to the cross-attention network for the sentences that have orientation defined in them. For instance, "The table should be in the northwest direction." Also, it would be great if we can also consider the 3D mesh of every object to maintain global consistency. Currently, the model can serve as a general framework for scene generation, and a different task can also be further solved by changing the set of objects properties or conditioning inputs.

I completed the Forward Focus Survey