Review-46

Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations

Modeling scene geometry requires explicit 3D supervision in the case of modern geometric deep learning models. The work aims to address this challenge by proposing Scene Representation Networks (SRNs). SRNs make use of continuous 3D structure-aware scene representations by encoding geometry and appearance. These networks represent scenes as continuous functions which map world coordinates to feature representations with local properties. SRNs are trained end-to-end from 2D images and camera poses using a differential ray-marching algorithm. This demonstrates the efficacy of SRNs to complex geometric tasks such as novel view synthesis, few-shot reconstruction, joint shape and appearance interpolation.

The SRN structure relies on the mapping of world coordinates to local feature representations by utilizing a continuous differential function. This is achieved by formulating a scene representation using MLP which maps 3D cartesian coordinates to features at final coordinates. Scene representations obtained from the MLP undergo a neural rendering algorithm which maps the representation to an image $\mathcal{I}$ based on the camera parameters. The neural rendering algorithm first finds the coordinates of intersections of respective camera rays. Following the determination of world locations, the rendering process maps feature vectors to spatial coordinates of color. The first step is achieved using a novel differentiable ray-marching algorithm. An LSTM maps feature vectors denoting the current estimate of ray intersections to the length of the next marching step. This is called steplength prediction. The second step of the process consists of a Pixel Generator Architecture which employs a per-pixel MLP mapping a feature vector to RGB vector. This is akin to 1x1 convolutions and preserved the global world coordinates of objects in the scene.

SRNs demonstrate suitable generation of novel views and reconstructions of prior scenes in comparison to prior representation methods. For instance, generated objects have sharp edges and appropriate geometry which is learnt implicitly without supervision. Additionally, the latent codes learnt by SRNs present a smooth transition depicting suitable interpolation of geometry and unsupervised representations. While SRNs present a novel viewpoint towards geometric deep learning and implicit representations of scenes, their formulation may be improved towards their components. Firstly, the per-pixel generator does not make use of convolutions and build a feature representation for each pixel. This hinders scalability of the scheme towards high-resolution scenes which consist of intricate details of objects and multi-view geometry. Secondly, the work aptly notes a limitation of ray marching algorithm. The algorithm gets stuck in cases of occluded portions of objects or pores on surfaces. This restricts the neural renderer to correctly generate complex objects such as chairs and tables consisting of 3D occluders.

SRNs present two novel directions for future research. Firstly, SRNs motivate the implicit learning of geometry which can be extended towards sophisticated architectures and room-scale scenes. Secondly, the use of a differentiable ray-marching algorithm presents the challenging problem of occluder marching. A potential solution to this end could be to encode parts of objects from different viewpoints and reconstruct them based on a likelihood function.

The work introduced SRNs, which make use of a continuous scene representation function mapping world coordinates to feature representations. These are utilized by a neural renderer consisting of a differential ray-marching algorithm in conjunction with pixel generator. SRNs demonstrate suitable generation of multi-view scenes which are learnt without supervision.