Outline:

1. Why has the progress for scaling world models on robotic applications such as autonomous driving been somewhat less rapid than scaling language models with Generative Pre-trained Transformers (GPT)?

dealing with complex and unstructured observation space, and having a

scalable generative model.

1. What is a world model?

World models explicitly represent the knowledge of an autonomous agent about its environment. They are defined as a generative model that predicts the next observation in an environment given past observations and the current action.

1. What is Copilot4D?

a novel world modeling approach that first tokenizes sensor observations with VQVAE, then predicts the future via discrete diffusion.

(For better efficiency of decoding and denoising🡺 some changes with Masked Generative Image Transformer as discrete diffusion.)

一張含有 文字, 字型, 數字, 行 的圖片

自動產生的描述

* Can be used on LiDAR.
* the world model uses a Transformer that interleaves spatial (Liu et al., 2021) and temporal blocks in Bird-Eye View (BEV). After tokenization, our world model operates entirely on discrete token indices.
* Flow: Tokenizer (like VQVAE:encoder🡪tokens🡪decoder🡪render🡪reconstruction)🡪discrete diffusion operated on BEV tokens.

1. Motivation of this work:

GPT learns to understand languages via next token prediction. If a

world model can predict unlabeled future observations really well, it must have developed a general understanding of the scene including geometry and dynamic.

1. What is a diffusion model?

a class of generative models that define a forward process from data distribution to noise distribution in closed-form, and then learn the reverse process from noise distribution to data distribution.

Formula used:

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

🡺for training, it masks a part of the input tokens (with an aggressive masking schedule) and then predicts the masked tokens from the rest; for sampling, it iteratively decodes tokens in parallel based on predicted confidence.

1. Training of the world model:

a. 50% of the time, condition on the past, denoise the future.

b. 40% of the time, denoise the past and the future jointly.

c. 10% of the time, denoise each frame individually, regardless of past or future. (to learn an unconditional generative model, which is necessary for applying classifier-free diffusion guidance during inference.)

World model: a spatio-temporal Transformer that simply interleaves spatial attention(Swim Transformer) and temporal (GPT2 blocks) attention.

1. Marks:
2. Classifier-free diffusion guidance (CFG) is important: by the class of c in training.
3. Improvement upon MaskGIT.
4. Overall results:

for 1s prediction, we are able to see a 65%-75% reduction in Chamfer Distance compared to prior SOTA across all three datasets; for 3s prediction, we are able to see more than 50% reduction in Chamfer.

Ideas:

This paper shows that building up an unsupervised World model could improve the one with classifiers. However, I’m not sure how good this model is to use in the real life: e.g. could we use it to predict the future on LiDAR within the deadline and with high rate of success?

Be that as it may, this World model could still shed light on the possibility of being used to give more precise guidance to drivers. Besides, from the perspectives of security, this prediction also can be used to detect malicious attack because serious LiDAR-spoofing should correspond to huge difference in a short-tern. Furthermore, if this can be used to detect the attacks on LiDAR, it also means that we can simulate a model based on Copilot to abstract the attacks out of the scenery in LiDAR. Then, we could utilize this model to analysis and quantize the attack on LiDAR to easily merge into the domain-specific language. The only extra thing we need to do is to expand the original domain-specific language, like SCENIC, to also simulate how it will render during the modelling, which could be done via Copilot.