High Level Summary of Sora: Video generation models as world simulators

Isaac Akintaro

24/02/2024

Explanation of Key Terminologies

- **Patches:** In the context of images, a patch would be a small square of pixels. In videos, which add the dimension of time, a patch extends this concept to include a sequence of frames.
- **Tokens:** A token is a commonly occurring sequences of characters. For example, you have the word 'prompting', which is quite new in its widespread usage. It would most likely be broken into these three tokens: Prom-pt-ing

High Level Overview

The fundamental belief behind machine learning is this:

There is an underlying phenomenon or model in nature that generated our data. Our quest is to find it.

If our data is not totally random and has been generated...then despite it being in a complex format such as image, video or text there must be an underlying pattern to it. This pattern is the simplified way of expressing that complex data.

In Sora, a raw stack of video frames are taken and then transformed to a simplified representation. It is then broken down into smaller, manageable pieces known as patches.

For large language models such as ChatGPT, they are trained using tokens, whilst Sora is trained using those patches.

Key Capabilities

Sora generates videos and images but can also:

- Extend videos going backwards and forward in time.
- Animate images by turning them into videos.
- Be used to edit videos e.g. provide new environments and definitions.
- Interpolate between two videos, causing a smooth transition.

Issues

Some physics are a bit odd such as broken glass or food quantities after it has been eaten.