The Mathematical Principles Behind OpenAI's SORA

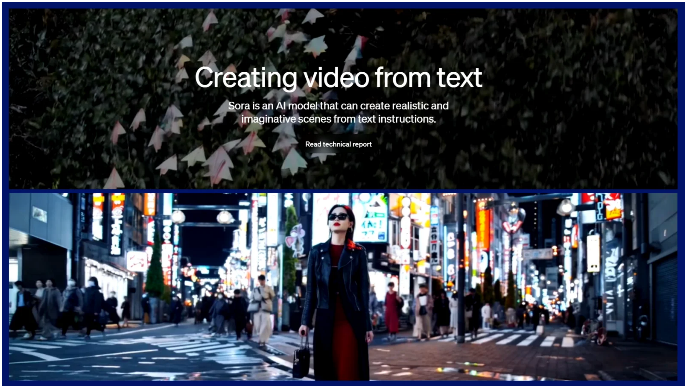
Emmanuel Adjei
University of California, Irvine
Statistics Department (DHB 2011)

June 1, 2024

OUTLINE

- Introduction
- Methodology
  1. Diffusion Probabilistic Modeling
  2. Scalable Diffusion Models with Transformer
  3. Tubelet embedding
  4. Factorized dot-product attention mechanism
- Reference

Introduction



Creating video from text

Sora is an AI model that can create realistic and imaginative scenes from text instructions.

Figure: Source:https://images.app.goo.gl/3cC34WCaGwAApw9o9

Everyone has witnessed the unveiling of Sora, OpenAI's remarkable video generation tool.

Introduction

- Advancing from DALL · E's static imagery, this innovation adds a vibrant, dynamic layer, reshaping our engagement with AI-generated material. The inevitable inquiry emerges: how did OpenAI accomplish this breakthrough?
- In this presentation we will delve into the mathematics and framework that form the background for this model.

Methodology

Diffusion Probabilistic Modeling.

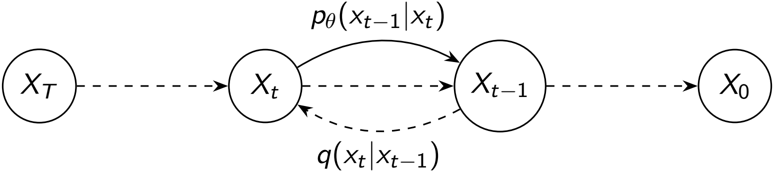


Figure: The directed graphical model

A diffusion probabilistic model is a Markov chain with parameters that is trained through variational inference to generate samples that resemble the data within a finite time frame.

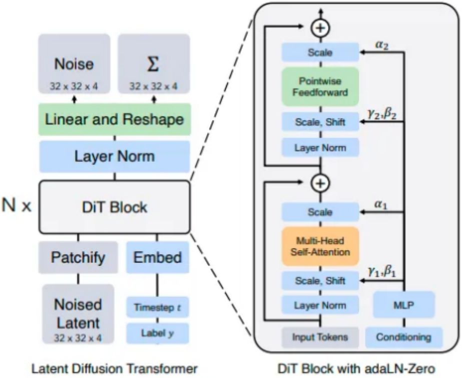
Diffusion Probabilistic Modeling.

The given model defines $p_\theta(\mathbf{x}_0)$ as the integral of $p_\theta(\mathbf{x}_{0:T})$ over $\mathbf{x}_{1:T}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latent variables with the same dimensionality as the data $\mathbf{x}_0$, which is distributed according to $q(\mathbf{x}_0)$. The joint distribution $p_\theta(\mathbf{x}_{0:T})$ is denoted as the reverse process, characterized as a Markov chain with learned Gaussian transitions, beginning from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$:

$$p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right), \quad p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right),\right.$$

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), \quad q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$$

Scalable Diffusion Models with Transformer



Diffusion Transformer Architecture

Figure: Source:https://images.app.goo.gl/3cC34

Scalable Diffusion Models with Transformer

The architecture is designed to handle the diffusion process in a transformer-based framework, likely aiming to leverage the transformer's ability to model long-range dependencies for generating or denoising data in a manner similar to denoising autoencoders but with the benefits of attention mechanisms. Nevertheless, transformers are highly resource-intensive, making them impractical for video scaling if we merely extend the temporal dimension.

Tubelet embedding

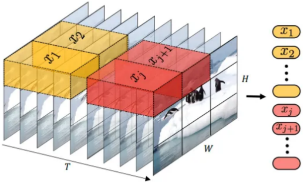


Figure: Source:https://images.app.goo.gl/3cC34WCa

The key advantage of tubelet embedding is that it allows a neural network, such as a transformer, to process videos by taking into account the dynamic changes across frames, which is essential for tasks such as action recognition, video classification, or any other application requiring an understanding of both spatial and temporal dimensions of the video. This extract non-overlapping, spatio-temporal "tubes" from the input volume, and to linearly project this to $\mathbb{R}^d$ [1].

Factorized dot-product attention mechanism

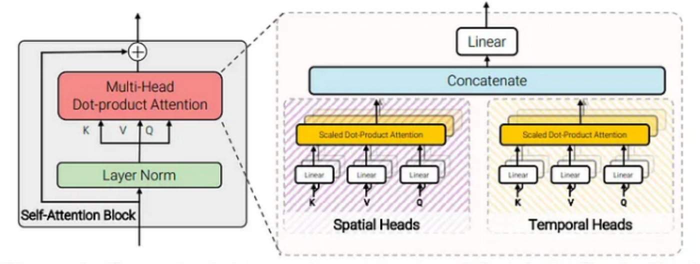


Figure: Source:https://images.app.goo.gl

Factorized dot-product attention mechanism

By factorizing the attention this way, the model can efficiently process video data by treating space and time separately, which is computationally less demanding than the traditional method that would consider all spatial and temporal elements together. This also allows the model to specialize in extracting spatial features (like shapes and textures) and temporal features (like movement and change) from the video data, possibly leading to a more nuanced understanding and better performance on video-related tasks.

- Vivit: A video vision transformer, Arnab, Anurag and Dehghani, Mostafa and Heigold, Georg and Sun, Chen and Lučić, Mario and Schmid, Cordelia, Proceedings of the IEEE/CVF international conference on computer vision, 6836–6846, 2021

- Denoising diffusion probabilistic models, Ho, Jonathan and Jain, Ajay and Abbeel, Pieter, Advances in neural information processing systems, 33, 6840–6851, 2020

Thank You ...