CAVLab Weekly Meeting

20/12/02 - 17:30 - Room 814, Building No. 2

WÜRSTCHEN: An Efficient Architecture for Large-Scale Text-to-Image DIFFUSION Models

Abstract

Diffusion models have emerged as powerful tools for generating realistic images from text, but state-of-the-art approaches are computationally expensive. The paper proposes Würstchen, a novel three-stage architecture that enables efficient text-to-image synthesis while maintaining competitive performance and unprecedented cost-effectiveness. A key innovation is learning a compact semantic representation of the image to guide the diffusion process. This compressed representation provides more detailed guidance than language representations alone, significantly reducing computational requirements for achieving state-of-the-art results. Notably, Würstchen also improves the quality of text-conditioned image generation, as evidenced by user preference studies. While current state-of-the-art diffusion models produce impressive results, they are computationally demanding. Existing alternatives that reduce computational costs tend to sacrifice image quality, resolution, and aesthetic appeal. Increasing resolution exacerbates visual complexity and training computation demands. Latent diffusion models partially address this by operating in a compressed latent space but are ultimately limited. Würstchen circumvents these limitations by training a diffusion model on an extremely low-dimensional 42:1 compressed latent space, which then conditions a second generative model in a higher 4:1 compressed VQGAN latent space. This innovative three-stage process consists of text-to-low-latent, low-to-high-latent, and high-latent-to-image decoding stages. Würstchen represents a significant step towards democratizing high-quality text-to-image synthesis models.[1]

Speaker

Fatemeh Nadi : graduated from the Isfahan University of technology in 2021. Currently, she is pursuing her masters degree in the field of Artificial Intelligence at the University of Tehran. Her research interests primarily focus on the speed up in diffusion models. Her research is conducted under the supervision of Dr. Reshad Hosseini and Dr. Mostafa Tavassolipour.

References

[1] P. Pernias, D. Rampas, M. L. Richter, C. J. Pal, and M. Aubreville, Wuerstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models, Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.00637