Sony introduces AI for single-instrument accompaniment generation in music production

June 26 2024, by Ingrid Fadelli

Credit: S. Marino, S. Lattner, DALL-E

In recent decades, many engineers have started developing artificial intelligence (AI)-based tools that can support the work of creative professionals, speeding up or enhancing the production of different types of content. These include computational models that can generate musical tracks and facilitate some aspects of music production.

Researchers at Sony CSL have been working on various AI-powered solutions designed to help musicians, music producers and other music enthusiasts throughout their creative endeavors. In a recent paper posted to the *arXiv* preprint server, they introduced Diff-A-Riff, a promising computational model that can generate high-quality instrumental accompaniments for any music.

"Our recent paper builds on our previous research on generating bass accompaniments," the music team of Sony CSL Paris, told Tech Xplore. "While our earlier work focused on creating bass lines to complement existing tracks, Diff-A-Riff extends this concept to generate single-instrument accompaniments of any instrument type."

"This evolution was inspired by the practical needs of music producers and artists, who often seek tools to enhance their existing compositions by adding additional instruments, and by their desire to be flexible concerning instrument types/timbres."

The primary goal of the recent work by the music team at Sony CSL Paris was to create a versatile AI system that can generate high-quality instrumental accompaniments that seamlessly integrate with a given musical context, focusing on one instrument at a time. The tool they developed is based on two distinct and powerful deep-learning techniques: latent diffusion models and consistency autoencoders.

"Diff-A-Riff leverages the power of latent diffusion models and consistency autoencoders to generate instrumental accompaniments that match the style and tonality of a given musical context," they explained.

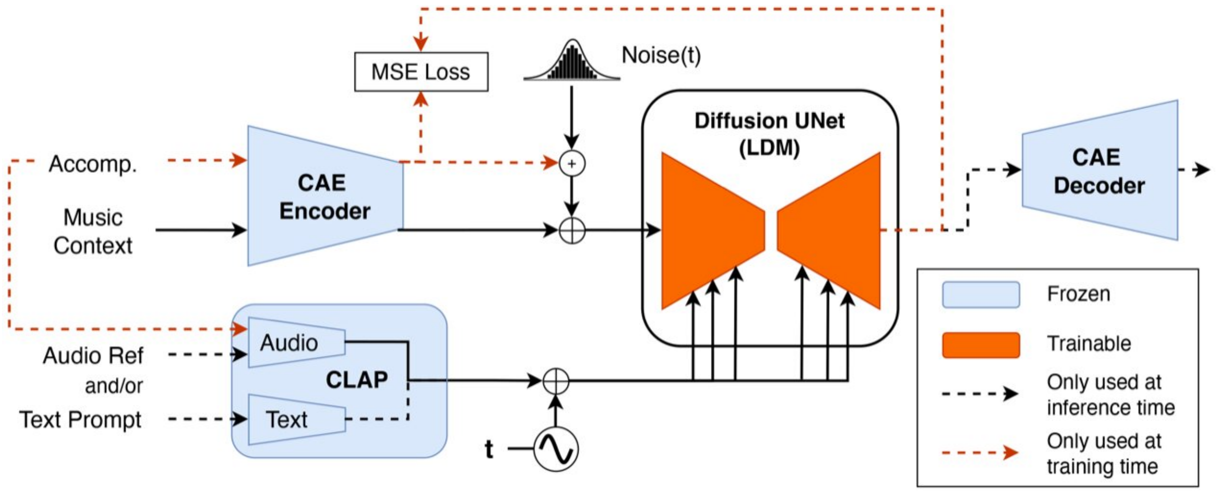
"The system first compresses the input audio into a latent representation using a pre-trained consistency autoencoder, a codec developed in-house, that guarantees high-quality decoding through a generative decoder. This compressed representation is then fed into our latent diffusion model, which generates new audio in the latent space, conditioned on the input context and optional style references from either text or audio embeddings."

Diff-A-Riff has numerous advantages over other tools for instrumental accompaniment generation. The first is its versatile control, which allows users to condition both audio and text prompts, offering them greater flexibility in guiding the generation of accompaniments. In addition, Diff-A-Riff produces high-quality outputs, with pseudo-stereo audio of 48kHz.

"Diff-A-Riff also significantly reduces inference time and memory usage compared to previous systems, as we are using a 64x compression ratio," the team explained. "We found that it can generate accompaniments for any musical context, making it a valuable tool for music producers and artists.

"Moreover, it features additional controls, such as the interpolation between instrument references and text prompts, the definition of stereo-width, and the possibility to create seamless transitions for loops."

The Sony CSL music team evaluated their model in a series of tests. Their findings were highly promising, as the model generated high-quality instrumental accompaniments for various music tracks that human listeners were unable to distinguish from recorded accompaniments played by human musicians.



Credit: C. Aouameur

"A generation speed of three seconds for one minute of audio is unprecedented and is achieved by the high compression ratio of the consistency autoencoder," they said. "In real-world scenarios, Diff-A-Riff can be applied to music production, creative collaboration and sound design."

The instrumental accompaniment generation tool developed at Sony CSL could soon be employed by music producers worldwide, allowing them to create instrumental tracks that complement their existing compositions. Diff-A-Riff could also be used by artists to easily explore new musical ideas or by sound designers to rapidly test different timbres or playing styles for their projects.

"Our future research plans include enhancing Diff-A-Riff's capabilities by improving the control mechanisms and exploring new ways to integrate the model into various stages of the music production process," the team added.

"We aim to provide even more intuitive inputs to make the model more accessible and useful for artists, including amateurs and professionals. Additionally, we plan to collaborate with musicians and composers to further refine and validate our models, ensuring they meet the practical needs of users in the music industry."

More information: Javier Nistal et al, Diff-A-Riff: Musical Accompaniment Co-creation via Latent Diffusion Models, *arXiv* (2024). DOI: 10.48550/arxiv.2406.08384

More images and audio available at: sonycslparis.github.io/diffariff-companion/

Citation: Sony introduces AI for single-instrument accompaniment generation in music production (2024, June 26) retrieved 29 June 2024 from https://techxplore.com/news/2024-06-sony-ai-generate-high-quality.html