Neural Network Project Report- Exploring TANGO in the Process of Text-to-Audio Generation

Braum Russell & Griffin Williams

MIT 5432 - Machine Learning

Dr. Shuang Zhai

April 29, 2024

Abstract

In this project, our group focuses on finding an area in machine learning that we deem to be both interesting and have beneficial use cases. Specifically, we were tasked to focus on the subset deep learning for our project. Deep learning is a section of machine learning, where deep learning specifically focuses on algorithms that mimic (loosley) the structure of our own brains, neural networks. This is why deep learning algorithms are also known as artificial neural networks. These algorithms are made up of many layers of connected nodes (which are mimicking the neurons in the brain) that allow them to learn representations of data.

Specifically in our study, we decided to use the paper "Text to Audio Generation using Instruction Tuned LLM and Latent Diffusion Model" which follows the development of the model TANGO. TANGO uses a large language model for text encoding, a latent diffusion model to create latent representations of audio, a variational autoencoder to create a mel-spectrogram from the latent audio representation to be then fed to a vocoder to generate the final audio requested from the prompt.

Domain introduction

Just over the past few years, we have seen the huge boom of popularity in Artificial Intelligence. Perhaps what started it for the general population would be the release of Chat-GPT back in November of 2022. Since then, AI has become a baseline in the production in many industries, as my partner and I have seen it been used in in many industries/domains. One specific area of AI that is becoming very popular is generative AI. We look specifically at the domain of text-to-audio generative AI due to the use cases it can provide.

One side of TTA AI could benefit the entertainment industry. For example, a movie producer might need the noise of a Fox's howl, which would be pretty hard to obtain. These producers could use a TTA model like TANGO to produce this specific noise instead of allocating resource time and money to obtain the noise.

Another side TTA AI could be used for could be in malicious ways. Threat actors could possibly obtain the voice of a specific person, say a Vice President of a Fortune 500 company and train such a model to produce a voice that exactly replicates them. They could use this to gain sensitive information that no one should have any access to except those who have the authorization.

Due to all of these interesting use cases that TTA systems can provide, we predict that AI companies will be investing in this technology in the upcoming future, where the models will just increasingly better perform to evaluation metrics.

Task introduction

Text-to-audio involves the conversion of a textual prompt into an audio representation, basically allowing the model to speak out the description of the text you give it. The sound you want produced might be layered, where you could possibly want a specific tone, for example a man speaking in a soft voice compared to someone yelling. You also might want emotion added, where in this same example you want the man to sound sad instead of happy. Achieving these goals require machine learning models capable of both understanding the prompt and generating natural-sounding audios.

To do this, TANGO first allows textual encoding using a LLM. Specifically, TANGO uses FLAN-T5 which is an instruction tuned large language model. Being instruction tuned allows the model to understand tasks, like creating the sound of a wave hitting a bank. The model can break this down to understand what the input wants it to perform. TANGO then uses a diffusion process involving adding noises and reducing noise level until the desired audio signal is made into a latent representation. From there, the VAE will translate the latent audio representation to a mel-spectrogram, where this is then fed to a vocoder which creates the audio.

Dataset introduction

The dataset used to train the Latent Diffusion Model in the Tango system was the AudioCaps dataset. The AudioCaps dataset includes about 46,000 audio clips, its AudioCap ID, Youtube ID, start time, and caption. Each one of these clips come from YouTube and are paired with human written prompts that are a description of what the audio clip is saying/sounds like to allow for training. Each clip is around 10 seconds long and the LDM is only trained with the audio and the caption. Also prevalent in the TANGO model is a VAE model and a LLM. Both of these are gathered pretrained.

Problem formulation

The problem TANGO approaches to solve is improving text-to-audio generation, which can provide a multitude of benefits. For example, TTA systems can serve as a content generation to allow those who are visually impaired to hear a prompt instead of having to read it. While existing models have proven to provide valuable TTA generation, semantics and deep representation of the sounds are still not far into the TTA life cycle. TANGO specifically claims to enhance audio quality, enhance semantic and ensure correctness, and improve efficiency by changing model architectures than previously developed models. By reaching these goals, TANGO can become the next state of the art TTA model and become the next stepping stone for such models.

Related work

The paper continuously compares its model TANGO to the text-to-audio model AudioLDM. AudioLDM main's novelty was that it was the first to translate a latent diffusion model to text-to-audio generation. Where the two differ was that they used audio embeddings instead of text embeddings during the process of backwards diffusion. TANGO was developed after AudioLDM and proved better results. The original TANGO model was trained on a dataset 63 times smaller than that of AudioLDM and still out performed AudioLDM.

Model details

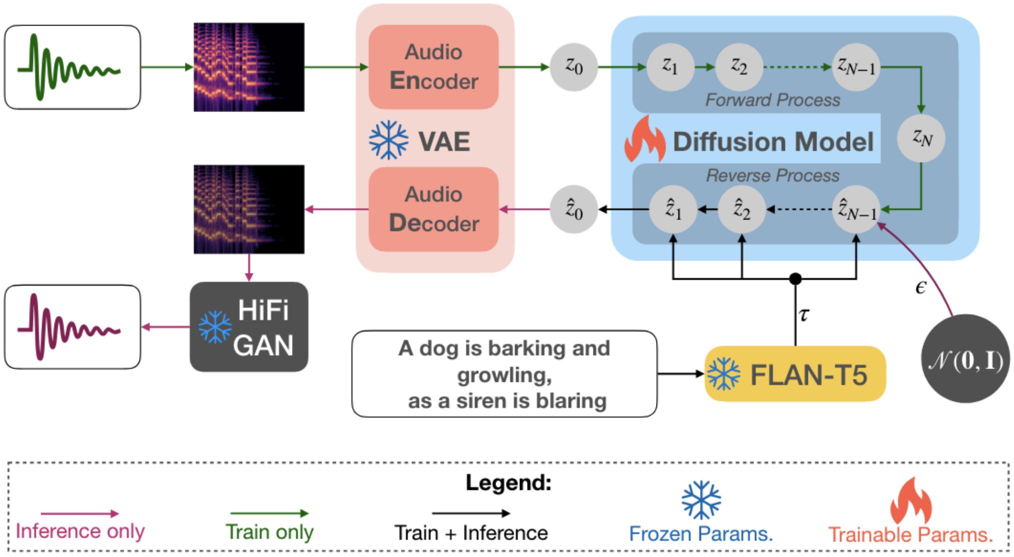


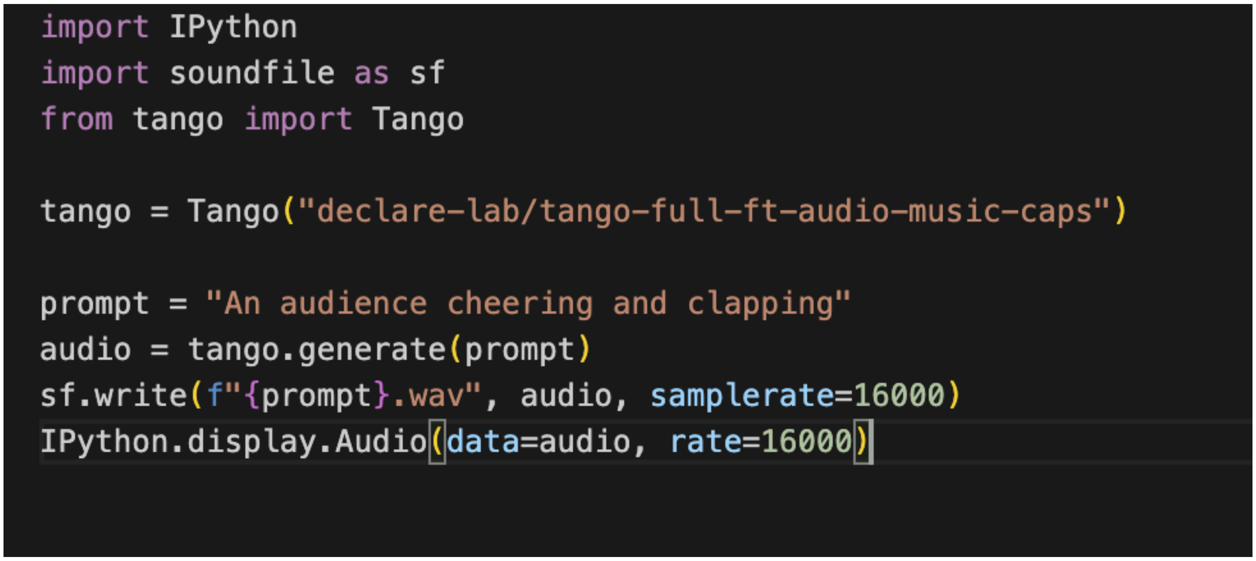
Figure 1: Overall architecture of TANGO.

Shown above is the architecture of the TANGO model. The model has 3 main parts: the textual prompt encoder, the latent diffusion model, and the mel spectrogram/audio variational encoder.

The three main parts of the TANGO system are traditional neural network components. First is the text encoder, FLAN-T5, being responsible for converting the text prompts to a latent representation to be used by the diffusion process. Second is the latent diffusion model, which is used to create a latent representation of the prompt. Then the variational autoencoder is used to generate mel-spectrogram tokens from the latent representations produced by the latent diffusion model. All three of these parts use deep learning algorithms, which at the end are fed to a vocoder to create the actual sound

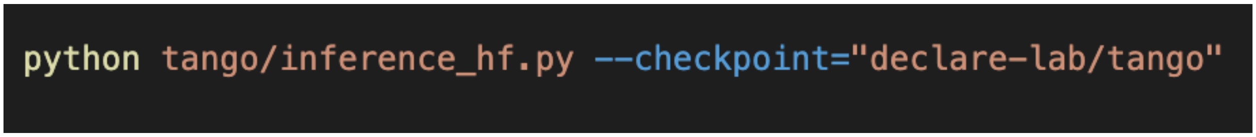
The main novel idea of TANGO is their use of FLAN-T5, which is the large language model they used that is fine/instruction-tuned to understand tasks in prompts, used as the text encoder. By leveraging FLAN-T5's understanding of textual semantics, TANGO is able to produce better relationships between prompts and audio content, compared to using the T5 base model. The researchers that created this model states that due to this, their model outperforms similar models that use non-instruction tuned LLMs, even after only training their model on a dataset 63 times smaller.

Implementation code snippets



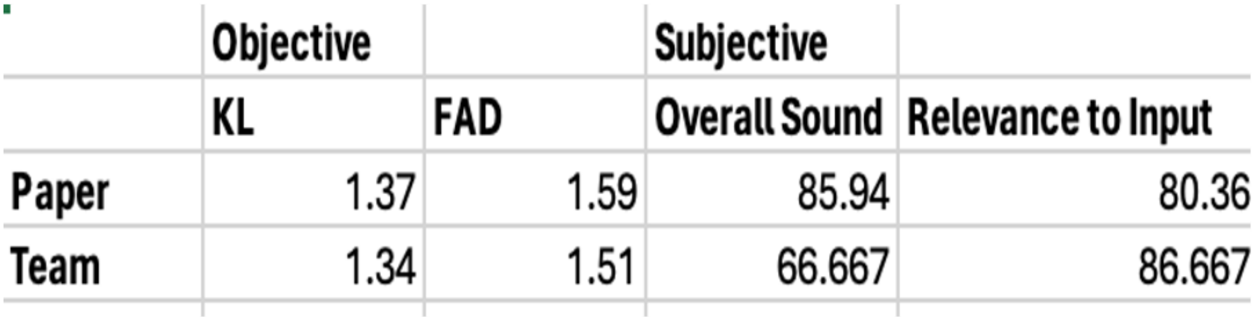
```python
import IPython
import soundfile as sf
from tango import Tango

tango = Tango("declare-lab/tango-full-ft-audio-music-caps")

prompt = "An audience cheering and clapping"
audio = tango.generate(prompt)
sf.write(f"{prompt}.wav", audio, samplerate=16000)
IPython.display.Audio(data=audio, rate=16000)
```

Model to Generate Sound File



```
python tango/inference_hf.py --checkpoint="declare-lab/tango"
```

Script to Run Python Test from Checkpoint



	Objective		Subjective	
	KL	FAD	Overall Sound	Relevance to Input
Paper	1.37	1.59	85.94	80.36
Team	1.34	1.51	66.667	86.667

Comparison Results

Paper vs Our Team’s Testing Results

As we can see from the table above, our model scores better than the original on both the KL (Kullback-Leibler Divergence) and FAD (Frechet Audio Distance) metrics by a small margin. The KL metric quantifies how much information is lost when using the generated distribution to approximate the real distribution. If the KL divergence is high, it suggests that the generated samples differ significantly from the real ones, indicating poor performance of the model in capturing the real data distribution. The FAD metric measures the similarity between two sets of audio signals by considering their underlying distributions in a feature space. It's essentially a way to quantitatively evaluate the similarity between two sets of audio samples. These are the objective metrics we tested our model with and those are more accurate in those areas compared to the original model.

However, the overall sound in the subjective metric was scored rather poorly at a score of 66.667 and this is due to one of the prompts being “A older man speaking the word hello” which ended up generating an audio sample of complete gibberish. I believe this is due to the AudioCaps dataset being primarily made up of general sounds and not actual speech samples. Therefore, when the Tango model is prompted to generate people speaking specific words, it may just sound like the background noise (or in our case, unintelligible words) that the dataset is mainly composed of. For both the objective and subjective testing methods we focused on these three prompts:

1. A dog barking
2. Rolling thunder with lightning strikes
3. An old man speaking the word hello

Comparison (on above two)

Our model surpasses the original model in both the KL and FAD metrics by a narrow margin. This suggests that our model is slightly more effective at capturing the underlying distributions of the real audio data compared to the original. This improvement is crucial as it signifies progress in generating audio that closely resembles authentic samples. This is crucial for various applications such as speech synthesis, music composition, and sound design.

However, despite these advancements, our model's overall sound score lags behind the original, scoring only 66.667 compared to the original's 85.94. This significant disparity can be attributed to specific challenges, such as generating coherent speech from prompts like "A older man speaking the word hello." In such cases, our model struggles to produce intelligible audio, potentially due to the dataset's bias towards general sounds rather than speech samples. This limitation highlights the importance of diverse and representative datasets in training AI models, ensuring they can effectively handle various prompts and tasks.

Discussion

Our project centered on the paper "Text to Audio Generation using Instruction Tuned LLM and Latent Diffusion Model," which introduces the TANGO model. TANGO employs a large language model for text encoding, a latent diffusion model for creating audio representations, a variational autoencoder for generating mel-spectrograms, and a vocoder for producing the final audio output.

The surge in artificial intelligence, particularly evident since the debut of Chat-GPT in 2022, has permeated various industries. Generative AI, in particular, holds promise, especially in the domain of text-to-audio generation. On one hand, it offers benefits to sectors like entertainment, where producers can effortlessly create specific sounds without the need for elaborate resources. On the other hand, there's a looming concern regarding potential misuse. Malicious actors could exploit text-to-audio models to impersonate individuals, leading to identity theft or spreading misinformation for personal gain.

With the burgeoning interest in text-to-audio systems, it's anticipated that AI companies will heavily invest in this technology. As models evolve, their performance metrics will likely improve, catering to diverse user needs and application scenarios.

Text-to-audio conversion, as a task, involves translating textual prompts into natural-sounding audio representations. This process entails understanding the nuances of the prompt and generating appropriate auditory outputs. TANGO achieves this by employing FLAN-T5, an instruction-tuned large language model, for textual encoding, followed by a diffusion process and vocoder for audio synthesis.

The dataset used to train TANGO's Latent Diffusion Model is the AudioCaps dataset, consisting of audio clips paired with human-written prompts. While objective metrics like KL divergence and FAD indicate improvements in model performance compared to the original, subjective testing revealed shortcomings, particularly in generating intelligible speech from specific prompts.

Despite advancements, our model's performance in subjective testing lags behind the original, highlighting the challenges associated with dataset bias and the importance of diverse training data. This underscores the need for continuous refinement and validation of AI models to ensure their effectiveness across various tasks and scenarios.

The comparison between these models highlights the ongoing need for advancements in AI audio generation but also reveals some possible negative consequences that come with it. As technology continues to progress, the ability to generate high-quality audio will greatly impact numerous domains. For instance, in entertainment, AI-generated music and voiceovers can streamline production processes and unlock creative possibilities. Furthermore, in entertainment, costs can be cut in sound design where audio does not have to be physically created and recorded, but just generated on a computer. In healthcare, speech synthesis technologies can assist individuals with speech impairments. In education and accessibility, AI-generated audio content can enhance learning experiences and make information more accessible to larger audiences.

However, it's important to acknowledge the potential risks associated with advancing text-to-audio generation technology. Beyond its positive applications, such as aiding the visually impaired or enhancing entertainment experiences, this technology could also be exploited for malicious purposes.

One concerning application is the creation of convincing audio forgeries. With increasingly sophisticated text-to-audio models, malicious actors could fabricate audio recordings of individuals saying things they never actually said. These forgeries could be used to spread misinformation, manipulate public opinion, or even frame individuals for crimes they didn't commit. Another concern is the potential for audio phishing attacks. By generating lifelike audio messages, attackers could trick individuals into divulging sensitive information or performing harmful actions. For example, a convincingly crafted voice message from a purported bank representative could deceive someone into providing their account details, leading to financial loss or identity theft.

Conclusion

Text-to-audio has been utilized for years, but only now are we seeing the real-world applications of it and how it can change industries and save money. From entertainment to healthcare, the possibilities for text-to-audio generation are vast and with our model being based on an open-sourced dataset that is constantly growing, it is easy to see how text-to-audio will improve in the near future. However, it is also crucial to acknowledge the negative consequences and possibly malicious actions that can take place with such impressive tools as this text-to-audio generation model. Malicious actors can generate fraudulent audios to achieve their goals and since this technology is constantly improving, so are the resources for malicious actors. Another important note is how objective testing metrics do not tell the whole story, especially in our case. It is easy to look over subjective testing and human intervention when there are numerical metrics that are generated from objective metrics, but that is exactly where our model fell short of the original model. In conclusion, it was very insightful working with the Tango model as our introduction to machine learning foundations as the results we found indicated the ever-growing use for AI and the promising future for deep learning technologies.

****** Our paper is based on the works of "Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model" which can be referenced with this link https://arxiv.org/pdf/2304.13731 ******