Audio-to-Image Project

Askhat Sametov, Galymzhan Zharas, Kenzhebek Taniyev, Marat Serikbayev, Temirtas Sapargaliyev
Computer Science, SEDS
Nazarbayev University
Astana, Kazakhstan
Email: askhat.sametov@nu.edu.kz, galymzhan.zharas@nu.edu.kz, kenzhebek.taniyev@nu.edu.kz,
marat.serikbayev@nu.edu.kz, temirtas.sapargaliyev@nu.edu.kz

***Abstract*—This project presents an approach to multimedia content generation by integrating an audio captioning system with an image generation model. We propose an encoder-decoder architecture for generating descriptive captions from audio inputs. The model's architecture consists of an encoder that processes spectrogram features of audio clips, and a decoder that generates captions sequentially. This automated captioning of audio data serves as a bridge to a subsequent image generation process, where the textual descriptions are converted into corresponding visual representations. Our integrated system not only enhances the accessibility of audio content by providing visual representations but also opens new avenues for creative multimedia applications.**

I. Introduction

In recent years, the field of artificial intelligence has witnessed remarkable advancements in both visual and audio understanding. Audio captioning, the task of generating textual descriptions from audio signals, is a complex challenge that combines elements of audio signal processing, natural language understanding, vision transformers, and multimodal learning. Traditional approaches to audio processing have focused on tasks such as speech recognition, music classification, and environmental sound detection. However, the potential to extend these capabilities to generate descriptive textual captions and corresponding visual images from audio inputs remains relatively unexplored and highly promising. Such a multimodal system could transform the way users interact with audio content, making it more accessible and interpretable, especially for those with sensory impairments.

Our proposed encoder-decoder framework uniquely integrates advanced machine learning models to first transcribe audio into text and then transform this text into images. This involves capturing audio signals in a manner that preserves their contextual and thematic details. By leveraging state-of-the-art models in both natural language processing and computer vision, our system interprets audio data in a way that can be extended with caption generative models. We use already pre-trained models in each phase of our system, in order to keep the knowledge of available models. We fine-tune the audio captioning model, and keep the image generation model as it is since our dataset only consists of audio-to-text pairs and not audio-to-image pairs.

The structure of this paper is organized as follows: We begin by detailing the architecture of our system, describing the specific technologies and methods employed at each stage of the audio to text to image transformation. This includes a discussion on the adaptation and integration of the Audio Spectrogram Transformer (AST) [1] for audio encoding and the GPT-2 [2] model enhanced with cross-attention for text decoding. Following this, we outline the procedure for preparing and processing data to suit the needs of each model component. We then delve into the training regimen, highlighting the hyperparameters and techniques used to optimize our model's performance.

II. Related Work

Audio captioning is a rapidly growing field that involves the creation of natural language descriptions for audio clips, thanks to the contributions of deep learning methodologies. The seminal work by Mei et al. [3] introduces a new approach in the use of Convolutional and Recurrent Neural Network architectures — termed Audio Captioning Transformer (ACT), which significantly differs from the traditional CNNs and RNNs previously used for audio captioning. Unlike conventional models, whose major downside lies in capturing the long-range dependencies of audio signals, the model presented in that paper leverages a full Transformer architecture — renowned for its success in natural language processing, to enhance global information modeling and temporal relationships in audio data.

Prior research in audio captioning has mainly used encoder-decoder frameworks with CNNs and RNNs. There were limitations associated with these methods — the local focus of CNNs and gradient issues of RNNs, which hindered their ability to manage long-term dependencies [4] [5]. At this point, there was a remarkable shift toward addressing these challenges with the introduction of the Transformer-based decoder, as the models became more context-aware and well-performing by understanding and generating contextually accurate audio descriptions [6].

The ACT model by Mei et al. builds atop this innovation by discarding convolutions entirely in favor of the self-attention mechanism, which allows for a more nuanced understanding and generation of audio captions. Authors of original paper conducted experiments with this approach on the AudioCaps dataset, which is the largest available for audio captioning, where ACT turns out to be competitive with the state of the art. Moreover, the ability of the architecture to function without

convolution layers makes the model simpler but still very robust in performance metrics during multiple evaluations.

That work definitely set a new benchmark in the landscape of audio captioning and, in effect, also suggests a pathway for future research, particularly on optimizing Transformer architectures for audio-specific tasks. Success of ACT just underscores the ability of Transformers to be trained for cross-modal applications and sets the stage for further explorations into their capabilities and adaptations in the audio domain.

III. Dataset

In our experiments we use the freely available audio captioning dataset Clotho [7], with 4981 audio samples and 24 905 captions. All audio samples are of duration from 15 to 30 seconds. This dataset was chosen mainly due to the fact that it consists of five different captions of 8 to 20 words length for each audio file, which is one of the most important benefits of training the model from the original pager and the model from our approach. Such captions improve the model's ability to comprehend and generate diverse linguistic expressions relating to one and the same audio context, therefore furthering a better understanding of the audio content of wild sounds.

IV. Methodology

The model architecture consists of two main components:

1) Audio Encoder: The encoder is based on the Audio Spectrogram Transformer (AST) [1] model, specifically pre-trained on the AudioSet dataset for classification task. Since AST has classification layer on top of the encoder layer, we removed it as we are interested in last hidden states of the encoder rather than class predictions. AST's encoder is responsible for processing raw audio inputs and encoding them into a feature-rich representation. The encoder effectively captures various audio characteristics essential for understanding the context and content of the sound.

2) Text Decoder: For the decoder component, we used the GPT2-base [2] model, configured with cross-attention layers to enable it to attend to the output of the audio encoder. This setup allows the GPT-2 model to generate textual descriptions based on the encoded audio features. The integration of cross-attention is critical for aligning the audio features with the corresponding textual output, facilitating more accurate and coherent caption generation.

3) Encoder-Decoder Model: The final model is just a combination of AST's encoder with GPT2's decoder via cross-attention. This model operates at two phases:

- Encoding phase: The audio input is first converted into a spectrogram and then fed into the audio encoder. The encoder processes the spectrogram to produce a set of latent features.
- Decoding phase: These latent features are then passed to the GPT-2 decoder. The decoder, using its cross-attention mechanisms, generates captions by contextualizing the encoded audio features, translating them into coherent text.

A. Data Preparation

In order to feed the data into our encoder-decoder model we need to prepare it, so it is compatible with both AST's encoder and GPT2 decoder.

1) Resampling: The first step in preparing the audio data involves resampling the audio files to a consistent sampling rate. This is crucial because the audio encoder expects inputs at a specific sampling rate to accurately process and encode them. Audios were converted from an original frequency of 44100 Hz to a new frequency of 16000 Hz. This step ensures uniformity across all audio samples, which is essential for consistent feature extraction.
2) Feature Extraction: Once all audio samples are resampled, they are passed to a feature extractor provided by the AST. The feature extractor computes Log-Mel Spectrogram features at the specified sampling rate of 16000 Hz. These features are critical as they represent the audio in a form that the neural network can process. Padding is applied to ensure all feature tensors are of the same length.
3) Caption Encoding: The raw text captions are then processed using a GPT2's tokenizer. This tokenizer is designed to convert text into a numerical format that the model can understand. After tokenization, the sequences are split into input IDs and labels for the training process. Input IDs are the token indices that will be fed into the model as input during the forward pass. The input IDs for each caption are obtained by slicing off the last token from the tokenized tensor, which effectively shifts the sequence to the left. Labels are used for calculating the loss during training and are typically the target outputs the model needs to predict. They are created by removing the first token from the tokenized sequences, shifting the sequence to the right. This setup means each label corresponds to the "next token" that the model should predict given the input IDs.

B. Model Training

Hyperparameters:

- Optimizer: Adam
- Learning rate: 5e-5
- Epochs: 50
- Batch Size: 8
- Loss: Cross Entropy Loss

V. Results

A. Caption Generation

Evaluation Metrics: BERT-score [8] is a metric for evaluating the quality of text generated by models using the contextual embeddings from BERT, a pre-trained deep bidirectional transformer model. Unlike traditional metrics such as BLEU or ROUGE, which primarily focus on n-gram overlap and are thus limited to exact lexical matches, BERT-score leverages the semantic richness captured in BERT's embeddings. This

makes it particularly suitable for tasks like captioning audio data in the Clotho dataset, where each audio clip can be interpreted in multiple ways, resulting in various acceptable captions. For instance, the Clotho dataset includes five different captions for a single audio file, highlighting the diversity of possible interpretations. BERT-score excels in such scenarios because it evaluates the semantic similarity rather than just the lexical similarity, allowing for a more nuanced assessment of the captions' quality relative to the content of the audio. This capacity to capture the depth of semantic relations between varied but valid captions makes BERT-score a more robust and appropriate choice for evaluating generative models in complex, multi-interpretative contexts.

In the validation phase of our project, an analysis of the first words generated in the captions for the audio samples revealed a significant challenge. The first words of generated captions showed a prevalence of incomplete or phonetically fragmented tokens such as 'ells', 'ient', and 'irds', along with a few correctly formed words like 'steps' and 'door'. A deeper inspection indicated that most of these malformed tokens were not standard English words. This issue predominantly affected the initial words of the captions, while the subsequent text typically maintained proper structure and coherence.

The primary reason for these initial anomalies can be attributed to the model's adaptation process during the transition from audio features to textual representation. The beginning of a caption often sets the context and is crucial for interpreting the audio clip's overall content. Thus, it may encounter more variability and complexity in its attempt to condense a broad auditory context into a concise textual start. This complexity can lead to inaccuracies, particularly when the model attempts to generate text from less distinct or more ambiguous audio cues. Also, GPT2's tokenization method is Byte-level tokenization, so that is why we get incomplete first tokens.

To address this, we implemented a cleaning step in our data processing pipeline, where we removed non-English words and corrected obvious misinterpretations when confident about the context (e.g., changing 'irds' to 'birds' due to the presence of multiple bird sounds in the dataset). This refinement was crucial for enhancing the readability and accuracy of the generated captions.

Following these adjustments, we applied the BERT-score on the cleaned predictions. The results were promising, with an average precision of 0.7604, an average recall of 0.7617, and an average F1 score of 0.7609. These scores reflect a high degree of semantic accuracy and consistency in the captions relative to the audio content, indicating that despite the initial word formation challenges, the captions successfully captured the essence of the audio clips. This validation not only underscores the efficacy of our model in handling complex audio-to-text transformations but also highlights the importance of targeted post-processing to enhance output quality.

From the Table I, we can see that generated captions can properly identify one or even two sources of audio data, and construct a proper caption out of the predicted audio signals. It is pretty good at identifying animal sounds, human speech and music, but sometimes struggles with other natural or environmental sounds. We believe that longer training time and some hyperparameter tuning can enhance the performance of our model. Additionally, we could train the model using all of 5 captions for each audio and not 1.

Fig. 1. Predicted Caption: Birds are chirping and a strong wind blows. Original Caption: Crickets and locusts chirp while a bird sings.

Fig. 2. Predicted Caption: eesbling and buzzing its computeroshing as time goes on. Original Caption: The vibrations of wings on a swarm of bees is causing a sound of buzzing.

B. Image Generation

In our project, we used this "runwayml/stable-diffusion-v1-5" [9] stable diffusion model for the task of image generation. Stable diffusion is a type of generative model that employs a diffusion process to gradually transform noise into a coherent image, guided by a deep learning model that has been trained on a large dataset of images. This approach allows for the generation of high-quality, detailed images that can be controlled with textual descriptions, making it highly suitable for creative and diverse image synthesis. The model leverages conditioning techniques to align the generated images with specific attributes or themes described in the input, enabling precise and contextually relevant visual outputs. By using this stable diffusion model, we were able to achieve impressive results in generating images that are both visually appealing and thematically aligned with the captions provided.

Examples of Image Generation: can be seen in figures 1 and 2.

VI. Conclusion

In this paper, we introduced an audio based image generation system using our own audio-to-caption generation encoder-decoder model with existing image generation model. Our approach leverages the robust capabilities of an Audio Spectrogram Transformer (AST) for audio encoding and a modified GPT-2 model for text decoding, integrated through cross-attention mechanisms. This design enables our system to handle the complexities of audio data, translating it into rich, contextually appropriate textual captions, which are then used to generate images using Stable Diffusion model.

Original Caption	Generated Caption
Various birds are chirping at a short distance over a hum in the background.	birds are chirping and a strong wind blows.
A chant is being song and a group of men answers.	ring loudly, a group of people speak out.
A few cars drive in the distance as the wind blows.	speak as cars drive by on a busy road.

TABLE I: Captioning examples.

The successful integration of audio, text, and image modalities in our system not only demonstrates the feasibility of such an approach but also sets a foundation for future research. Potential directions include refining the encoder-decoder interface, expanding the training dataset to include a broader range of audio types, and exploring more sophisticated image generation techniques to further enhance the visual representation of audio data.

Ultimately, our work contributes to the ongoing exploration of AI-driven multimodal content generation, offering insights and methodologies that could be applied in various domains, including digital media, education, accessibility, and entertainment. By continuing to develop and refine these technologies, we can enhance our ability to create more engaging, accessible, and interpretable multimedia experiences.

REFERENCES

[1] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," 2021.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[3] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," 2021.

[4] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 374–378, IEEE, 2017.

[5] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 905–909, IEEE, 2021.

[6] K. Chen, Y. Wu, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, "Audio captioning based on transformer and pre-trained cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, pp. 21–25, 2020.

[7] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," 2019.

[8] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.