Interspeech 2024
1-5 September 2024, Kos, Greece

Sound of Vision: Audio Generation from Visual Text Embedding through Training Domain Discriminator

Jaewon Kim[1], *Won-Gook Choi*[2], *Seyun Ahn*[2], *and Joon-Hyuk Chang*[1,2]

[1]Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea
[2]Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
{eingkim, onlyworld94, tpdbs0907, and jchang} @hanyang.ac.kr

Abstract

Recent advancements in text-to-audio (TTA) models have demonstrated their ability to generate sound that aligns with user intentions. Despite this advancement, a notable limitation arises from the models' inability to effectively synthesize audio from visual-domain texts. In this study, we address this challenge by utilizing a novel dataset that pairs visual and acoustic-domain texts, derived using ChatGPT-3.5, and encoding switch through a domain discriminator. This approach ensures not only computational efficiency but also enhances the model's generalization, adaptability, and flexibility. It addresses concerns that training exclusively with visual texts might compromise audio generation quality from audio texts. This study presents a novel methodology for enhancing text-to-audio synthesis, demonstrating significant improvements in audio output fidelity from visual-text inputs.

Index Terms: audio generation, text embedding, multi-modal

1. Introduction

Text-to-audio (TTA) generation task aims to synthesize accurate audio waveform from user-provided text. As long as users deliver the text information to TTA systems, these systems can leverage this input to synthesize user-intended audio content. This underscores the significant role of text encoders in TTA models, as they are instrumental in interpreting textual data to acoustic data. Diffsound [1] first deployed denoising diffusion probabilistic model (DDPM) [2] in latent space for audio creation, incorporating pre-trained contrastive language-image pre-training (CLIP) [3] as text encoders. Recently, T5 [4] and contrastive language-audio pre-training (CLAP) [5] have been preferred for text feature extractors due to their superior performance. Notably, AudioGen [6] and TANGO [7] also utilize diffusion in latent space, adopting T5 and Flan-T5-Large as text encoders, respectively. Similar to CLIP, a pre-trained contrastive learning text encoder, CLAP has become a popular choice for text representation in several studies [8, 9, 10, 11]. While CLIP is pre-trained to align text and image embeddings through similarity training, CLAP focuses on the correlation between text and audio embeddings. Despite T5-Large and CLAP achieving comparable results on several benchmarks, CLAP is recognized for its computational efficiency [9]. Pre-trained CLAP model architecture deploys robustly optimized bidirectional encoder representations from transformer (BERT) [12] pretraining approach (RoBERTa) [11, 13] as a text encoder and audio transformer with a hierarchical structure (HTS-AT) [14] as an audio encoder.

Despite the achievements of these studies, TTA systems still need to manage a severe problem: failing to generate audio from text not containing acoustic information-merely including visual information. For instance, text *"a small girl holding milk bottle and comb doing some actions"* is unable to synthesize proper audio waveform with the existing model. In this study, we address this failure with a novel dataset and manipulating text encoder. We name the new dataset V2A (Visual-to-Audio), which comprises pairs of visual and audio texts, utilizing MSR-VTT [15], Clotho [16], and AudioCaps [17] for its creation, while MSR-VTT is a dataset for the video captioning task, and Clotho and AudioCaps are audio captioning datasets. Therefore, providing visual-domain text and acoustic-domain text, respectively. To generate captions of one domain from the other, we employed large language models (LLMs). Specifically, we utilized ChatGPT3.5, Mistral [18], and Phi-2 [19] LLM models, among them.

In addition to the dataset, we propose an encoding switch method through a domain discriminator. This method employs a domain discriminator, an acoustic-domain text encoder (ATE), and a visual-domain text encoder (VTE). The ATE is the baseline text encoder since it is trained with acoustic-domain texts, while the VTE is newly trained using our novel dataset. Initially, the domain discriminator calculates the probability that the input text belongs to the visual domain. Subsequently, this probability determines whether to use ATE or VTE. We name this domain encoding switch (DES) since it switches the domain encoder according to domain of the input. We emphasize that we only train the text encoder, rather than the entire TTA model to take advantage of computational cost savings. While the baseline, AudioLDM-M-Full [8], the model has a total of 726 M parameters, our trained RoBERTa model only has 125 M parameters. This indicates that our method, despite needing to train two models (domain discriminator and VTE, totaling 250 M parameters), still significantly reduces the computational cost required for training. Additionally, this approach offers advantages in flexibility, scalability, and improved robustness.

2. Related Works

2.1. Pre-trained Transformer Models

Following the significant breakthrough of the Transformer [20] in natural language processing, primarily utilizing the attention mechanism [21, 22, 23], related pre-trained models have emerged. Models using the Transformer encoder are referred to as autoencoding models, notably including BERT [12], trained with 3.3 billion words, and RoBERTa [13], which significantly improved performance by adjusting training strategies and methods from BERT. These models are characterized by a masked language model that masks certain parts of sentences. Since only the Transformer encoder was used, they have been employed as text encoders in various multi-modal models [8, 11]. Additionally, sequence-to-sequence (seq2seq) mod-

10.21437/Interspeech.2024-1451

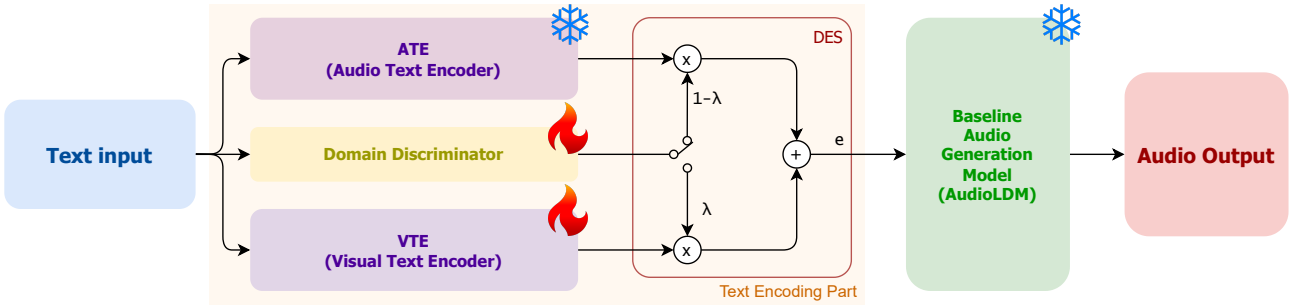

Figure 1: *The inference flow of the proposed model. Domain discriminator decides which text encoder should be used by multiplying zero. DES refers to the domain encoding switch. ATE is a pre-trained RoBERTa model from the baseline [8] without extra training, while the domain discriminator and VTE are trained by contrastive learning. The λ denotes the output of the domain discriminator and indicates if the input is a visual domain or acoustic domain.* ***e*** *refers to text encoding vector after the DES. The baseline generation model after the text encoding part, is used only for the inference process.*

els that process text inputs with both encoder and decoder from the Transformer exist in the TTA models field [6, 7, 10, 24]. The encoder is trained bidirectionally, and the decoder is trained autoregressively, with bidirectional auto-regressive transformer (BART) being trained with a text infilling method [25] beyond the conventional training methods. Moreover, T5 series models [4, 26], trained both with text including prefixes for various tasks as inputs, have been widely used recently among seq2seq models.

2.2. Contrastive Learning

Contrastive learning is a training method that encourages data forming positive pairs to have similar embeddings and those forming negative pairs to have opposite embeddings. Typically, in a batch, samples forming the same pair are considered positive pairs, and all other samples are considered negative pairs [3, 5, 27, 28, 29]. Recent studies in the multi-modal field use pre-trained models as feature extractors for inputs through contrastive learning [1, 8, 9, 10, 30, 31]. This approach, is effective not only for multi-modal tasks by increasing similarity between data related across different modalities but also for uni-modal tasks by enhancing similarity between similar data, outperforms traditional embedding methods. For instance, inspired by SimCSE [28], which increased text similarity, our study also performs contrastive learning between text encodings.

3. Proposed Methods

Our model consists of two text encoders: ATE and VTE. The role of ATE is encoding the acoustic-domain text, while VTE encodes the visual-domain texts. This is the reason why whether to use encoding from the ATE or VTE depends on whether the domain of input text is acoustic or visual. However, in the real world, it is often unclear if the text input belongs to the visual or acoustic domain. Therefore, our proposed model, as shown in Figure 1, comprises a domain discriminator and both ATE and VTE for text encoding, while ATE is a pre-trained RoBERTa text encoder from baseline [8]. The domain discriminator determines whether the text input belongs to the acoustic domain or the visual domain, represented by 0 and 1, respectively. This value is then used to decide whether to activate the ATE or the VTE. This method provides flexibility as the model can handle the input for both of the domains. Furthermore, if text domains beyond acoustic and visual are defined in the future, they can be additionally trained in the domain discriminator, offering scalability. Since domain-specific encoders are trained, high performance is maintained in single domains, and applying DES across all defined domains results in improved robustness. In this approach, It is not considerable the case of input text containing characteristics of both domains as this means the TTA system is easily able to generate audio if it contains a slight component of the acoustic domain.

Domain discriminator. We train the domain discriminator as a binary classification method to distinguish between acoustic-domain text and visual-domain text, representing the domain component of the input text as 1 if it is closer to the visual domain and 0 if it is closer to the acoustic domain. Therefore, we select BERT and RoBERTa, models known for their high performance in text classification, as candidates for the domain discriminator. Also, cross-entropy loss and means-square error (MSE) losses are opted for optimizing the model. The domain discriminator, trained to distinguish between visual-domain texts and acoustic-domain texts, calculates the probability that an input text belongs to the visual domain. This probability then determines the usage of VTE and ATE, represented as λ, and is mathematically expressed as follows:

$$\lambda = \mathrm{Dis}(\mathbf{x}_{\mathrm{text}}) \tag{1}$$

where $\mathrm{Dis}(\cdot)$ denotes the domain discriminator, $\mathbf{x}_{\mathrm{text}} = (x_n \in V | n = 1, ..., N)$ implies the input text, V denotes the vocabulary space, and $\lambda \in \{0, 1\}$

VTE. For training the VTE, we choose three candidates such as dense layers, a transformer encoder, and the RoBERTa from the baseline model [8]. To align with the output dimensions of the existing baseline model, all candidate models possess dimensions of $L = 768$. Each candidate model incorporates a projection layer structure similar to the AudioLDM baseline (dense layer - ReLU - dense layer), and during performance evaluation, models with dense layers were assessed both with and without the inclusion of the projection layer. The common word-level augmentation methods of natural language processing, such as RS, RD, and SS [32, 33, 34] are adopted for three-quarters of the input text with the same ratio of VTE to avoid overfitting and secure the generalization. We train the VTE with the contrastive learning loss between acoustic-domain text from the frozen ATE and visual-domain text from the training VTE. The loss function treats paired visual-domain text and acoustic-domain text samples as positive pairs, and other samples in batch as negative pairs. We use the normalized temperature-scaled cross entropy (NT-Xent) loss function from [27] and the formula of the loss function is as follows:

$$L = -\log \frac{\sum_{i=1}^{N} \exp(\mathbf{e}_i^r \cdot \mathbf{e}_i^t / \tau)}{\sum_{i=1}^{N} \sum_{j=1}^{N} (1 - \delta_{ij}) \exp(\mathbf{e}_i^r \cdot \mathbf{e}_j^t / \tau)} \tag{2}$$

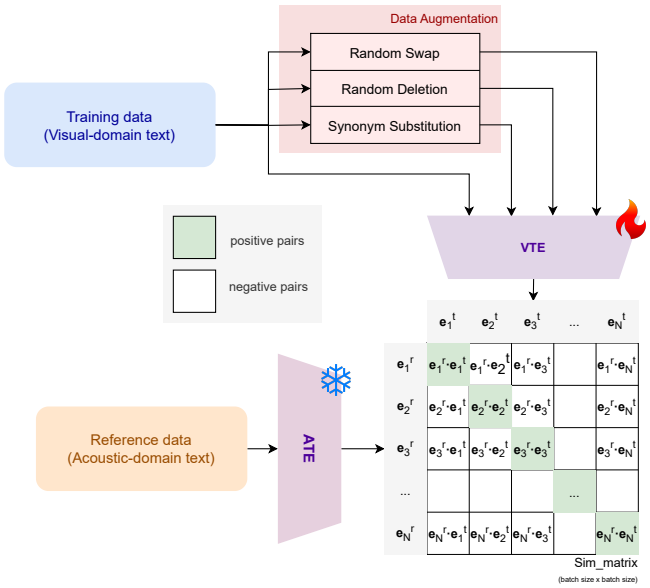

Figure 2: *The VTE training flow with contrastive learning.* $\mathbf{e}^r$ *refers to reference encoding vector, while* $\mathbf{e}^t$ *implies training encoding. The batch size is symbolized with* N*. The acoustic-domain captions become reference encoding vectors by passing through ATE. The training encoding vectors are derived from visual-domain texts after VTE with text augmentation.*

where $\mathbf{e}^r$, $\mathbf{e}^t$, τ, and N are the embedding of reference text encoding vector, training text encoding vector, temperature parameter, and batch size, while δ_{ij} denotes Dirac delta function, which results in 1 when $i = j$ and 0 for otherwise.

DES. During the inference, the DES mechanism switches encoding vectors from ATE and VTE. The VTE receives the value of λ from the domain discriminator, which is then directly applied to the results through usage of the output of the ATE. This process is formulated as follows:

$$\mathbf{e} = \lambda \cdot \mathrm{VTE}(\mathbf{x}_{\mathrm{text}}) + (1 - \lambda) \cdot \mathrm{ATE}(\mathbf{x}_{\mathrm{text}}) \quad (3)$$

where $\mathrm{VTE}(\cdot)$ and $\mathrm{ATE}(\cdot)$ represent the VTE and ATE, respectively, while $\mathbf{e}$ denotes the text encoding vector that becomes the input of the AudioLDM [8].

4. Experiments

4.1. Datasets

4.1.1. MSR-VTT

MSR-VTT [15] is a large-scale video captioning dataset collected from a commercial video search engine. It consists of over 10,000 web video clips, each divided into 20 categories and containing 20 natural language captions. It includes 200,000 clip-sentence pairs and encompasses a vocabulary of approximately 29,000 words. Since our dataset contains audio clips as well, we only handled the video data with audio clips. We select 6,176 clips for the training set, 1,370 clips for the validation set, 1,265 clips for the test set, and second captions from MSR-VTT for our dataset. The captions chosen from MSR-VTT are treated as visual-domain texts as these are derived from a video captioning dataset.

4.1.2. Clotho

Clotho v2.1 [16] is an automated audio captioning (AAC) dataset that contains 6,974 audio samples and 34,870 captions, ranging in length from 15 to 30 seconds, collected from the Freesound platform. Each audio sample has five captions ranging from 8 to 20 words, collected following a specific protocol of crowdsourced audio annotation for diversity and reduced grammatical errors. The dataset consists of a development set, a validation set, and an evaluation set. The development set contains 3,840 audio samples and 14,465 captions, the validation set contains 1046 audio samples and 5,225 captions, and the evaluation set contains 1,045 audio samples and 5,215 captions.

4.1.3. AudioCaps

Another AAC dataset we assembled is AudioCaps [17], which is a dataset containing 46,000 audio-caption pairs. The audio clips were collected through from the AudioSet [35] dataset, each 10 seconds long, and the captions were written by humans based on these clips. The datasets consist of training, validation, and testing. The training set contains 38,118 audio clips, the validation set contains 500 audio clips, and the testing set contains 979 audio clips. The training set provides one caption for each audio clip, while the validation and testing sets provide five captions for each audio clip. Both automated audio captioning and TTA were developed with AudioCaps.

4.1.4. V2A dataset

We used existing audio and video captioning datasets as input for the text generation model. We used visual captions as input from MSR-VTT the video captioning dataset and generated audio captions as output. On the other hand, we took audio captions as the input from Clotho and AudioCaps the audio caption datasets, and generated visual captions as output. We compared three LLMs. (ChatGPT3.5, Mistral, Phi-2) [18, 19] for creating high-quality visual-domain texts. We monitored how the values of the NT-Xent loss function were dropped from each dataset. As a result, the captions generated by ChatGPT3.5 achieved the lowest value of NT-Xent, consequently, we applied ChatGPT-3.5 generated data for the V2A dataset. Each data consisted of three parts: waveform, visual-domain text, and acoustic-domain text. We split the dataset into 59,280 training datasets, 3,358 validation datasets, and 3,607 testing datasets. We also deleted the audio-related vocabulary in visual-domain texts, in case LLMs generate acoustic words for generating visual-domain texts.

4.2. Experimental Setup

We nominated several models for domain discriminator and VTE. We selected BERT and RoBERTa as candidates for domain discriminator. Dense layers, untrained transformer encoder layers, and RoBERTa were also chosen for VTE. We deployed cosine annealing learning rate from 10^{-6} to 10^{-2} for every 4 epochs and trained 20 epochs in total with a batch size of sixteen for the training domain discriminator. In terms of VTE, we augmented three-quarters of the text data, then trained with contrastive learning as shown in Figure 2. Frozen ATE was treated as reference values, that results of VTE should follow. The 10^{-4} learning rate was chosen for training, and 400 epochs in total with a batch size of 512 since a result increases dramatically with the larger batch size [27]. We combined the ATE, VTE, and domain discriminator through both encoding mix-up and encoder switching methods. The mix-up [36] approach is a popular method for augmentation in several modalities [37, 38]. However, we tried a mix-up in the encoding realm instead of augmentation by combining the outputs of VTE and ATE through a weighted sum. The probability calculated by the

Table 1: *The results of experiments. DD refers to domain discriminator. MSE and CE denote that the model was trained with mean square error loss and cross-entropy loss, respectively. The evaluation metrics of domain discriminators were calculated with the RoBERTa model with the switching method. In DD + VTE rows, each of the best-performing models for FD and FAD scores was implemented for evaluation. The evaluation was executed with visual-domain texts of the V2A dataset.*

Models		Evaluation scores				
		Acc (DD only)	FD↓	FAD↓	IS↑	KL↓
Baseline [8]		-	51.080	9.637	**4.676**	3.678
Domain Discriminator	BERT (MSE)	0.933	71.971	16.626	3.445	5.184
	BERT (CE)	0.950	71.678	16.625	3.438	5.185
	RoBERTa (MSE)	0.968	**48.320**	**9.337**	3.965	**3.655**
	RoBERTa (CE)	**0.974**	48.331	9.339	3.967	3.655
VTE	Dense layer + ReLU	-	69.110	16.166	3.512	5.221
	Transformer Encoder	-	111.278	21.969	2.543	5.297
	RoBERTa	-	48.332	9.364	3.960	3.656
DD (RoBERTa-MSE) + VTE (RoBERTa)	Mix-up	-	85.029	18.919	1.892	4.168
	Switching	-	**48.320**	**9.337**	3.965	**3.655**

domain discriminator, which indicates whether the input text belongs to the visual domain, is used to weight the output of VTE, and vice versa. In the switching method, the probability calculated by the domain discriminator is used to immediately determine the domain of the input, subsequently employing either VTE or ATE exclusively for text encoding.

4.3. Evaluation Methods

We evaluated the generated waveforms by feeding our model the visual-domain texts from the V2A dataset as input. We measured frechet distance (FD), frechet audio distance (FAD) [39], inception score (IS), and kullback-leibler (KL) divergence. FD score measures the distribution of the generated audio using the pre-trained audio neural networks (PANNs) [40] audio encoder to calculate the distance of the generated audio from the target audio. FAD is similar, except that it uses the VGGish [41] encoder instead of the PANNs encoder. The higher the IS, the more diverse the audio can be generated. For KL divergence, it computes the difference in entropy between the two distributions. In this way, FD, FAD, and KL-divergence can be used to evaluate audio quality, and IS can be used to evaluate the diversity of the model.

5. Results and Analysis

As shown in Table 1, the comparative performance of domain discriminators reveals intriguing insights into the complexity of domain-specific TTA synthesis. RoBERTa, usually rated more fine-trained than BERT, surpassed the BERT on both accuracy and evaluation metrics. Even with the same RoBERTa model, the model trained with MSE, despite achieving slightly lower accuracy than its cross-entropy counterpart, attained superior evaluation scores across the board. This indicates the accuracy of domain discriminators does not always lead to the high performance of the overall system. Since MSE loss focuses on minimizing the squared differences in probability estimations, this may generalize varied textual inputs. Also, the crucial aspect is not the binary accuracy of the domain prediction but how well the predicted probabilities guide the optimal encoder selection for a given input. Even with marginally lower accuracy, a more finely tuned probability estimate results in more effective use of ATE and VTE, thereby indirectly influencing the quality of the generated audio. VTE with dense layers was unable to catch complicated relations between the visual and acoustic-domain texts. Considering untrained transformers usually require a larger amount of data for training, the performance of the transformer encoder model is less than the baseline model. On the other hand, we observed the RoBERTa VTE outperforms the baseline in its ability to handle the complexity of visual-domain texts. This comparison emphasizes the importance of selecting an encoder that is flexible and adaptable to the specific requirements of TTA task. We could also analyze the DES strategy that the switching method has improved not only more than the baseline, but also than the VTE-only model. Since we only trained text encoders, the parameters of the rest of the baseline model have not been affected during training. This explains why the encoding mix-up method dropped the performance than the RoBERTa-MSE model. It is uncertain whether the mix-upped encoding values match the existing baseline model. However, as VTE itself was trained to refer to ATE, baseline text encoder, switching the encoding values without manipulation increases its performance.

6. Conclusion

Encompassing domain discriminators, VTE, and the DES mechanism in this study represents a proceed in the field of TTA synthesis. The understanding and application of the domain discriminator model manage the intricate balance between classification accuracy and functional utility within the TTA system. The high-level performance of RoBERTa-based VTEs in transforming visual-domain texts into coherent audio outputs highlights the importance of encoding strategies in overcoming the challenges of TTA generation. Moreover, the successful integration of the DES mechanism underscores the feasibility and effectiveness of domain-adaptive approaches in enhancing system adaptability and output quality. This study not only broadens the field of TTA systems but also sets a precedent for future research to research deeper into domain adaptation strategies, aiming to bridge the gap between visual and acoustic information. Lastly, without training the whole pre-trained TTA model, we discovered a way to effectively generate audio from visual-domain text.

7. Acknowledgement

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.RS-2023-00302424) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)).

8. References

[1] D. Yang *et al.*, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Advances in Neural Information Processing Systems* (NeurIPS), vol. 33, 2020, pp. 6840–6851.

[3] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning* (ICML). PMLR, 2021, pp. 8748–8763.

[4] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[5] Y. Wu *et al.*, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2023, pp. 1–5.

[6] F. Kreuk *et al.*, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[7] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proc. ACM International Conference on MultiMedia* (MM), 2023, pp. 3590–3598.

[8] H. Liu *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[9] R. Huang *et al.*, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[10] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *arXiv preprint arXiv:2401.01044*, 2024.

[11] Y. Yuan *et al.*, "Retrieval-augmented text-to-audio generation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2024.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] Y. Liu *et al.*, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] K. Chen *et al.*, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2022, pp. 646–650.

[15] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conference on computer Vision and Pattern Recognition* (CVPR), 2016, pp. 5288–5296.

[16] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2020, pp. 736–740.

[17] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[18] A. Q. Jiang *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[19] Y. Li *et al.*, "Textbooks are all you need ii: phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[20] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems* (NeurIPS), vol. 30, 2017.

[21] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *arXiv preprint arXiv:1702.00887*, 2017.

[22] Z. Lin *et al.*, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[23] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conference of Empirical Methods in Natural Language Proceeding* (EMNLP). ACL, 2015, pp. 1412–1421.

[24] H. Liu *et al.*, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[25] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[26] S. Shen *et al.*, "Flan-MoE: Scaling instruction-finetuned language models with sparse mixture of experts," *arXiv preprint arXiv:2305.14705*, 2023.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. International Conference on Machine Learning* (ICML). PMLR, 2020, pp. 1597–1607.

[28] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conference on Empirical Methods in Natural Language Processing* (EMNLP). ACL, 2021, pp. 6894–6910.

[29] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," *arXiv preprint arXiv:2309.05767*, 2023.

[30] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[31] M. Tang *et al.*, "Clip4caption: Clip for video caption," in *Proc. ACM International Conference on MultiMedia* (MM), 2021, pp. 4858–4862.

[32] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP). ACL, 2019.

[33] J. Kim, Y.-A. Park, J.-H. Cho, and J.-H. Chang, "Improving automated audio captioning fluency through data augmentation and ensemble selection," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (DCASE), 2023, pp. 86–90.

[34] J.-H. Cho, Y.-A. Park, J. Kim, and J.-H. Chang, "HYU submission for the dcase 2023 task 6a: Automated audio captioning model using al-mixgen and synonyms substitution," in *Proc. IEEE Detection and Classification of Acoustic Scenes and Events* (DCASE), 2023.

[35] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. 2017 IEEE international conference on acoustics, speech and signal processing* (ICASSP), 2017, pp. 776–780.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. International Conference on Learning Representations* (ICLR), 2018.

[37] X. Hao *et al.*, "Mixgen: A new multi-modal data augmentation," in *Proc. IEEE/CVF Winter conference on Applications of Computer Vision* (WACV), 2023, pp. 379–389.

[38] E. Kim *et al.*, "Improving audio-language learning with mixgen and multi-level test-time augmentation," *arXiv preprint arXiv:2210.17143*, 2022.

[39] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Frechet audio distance: A metric for evaluating music enhancement algorithms," in *Proc. INTERSPEECH*, 2019.

[40] Q. Kong *et al.*, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[41] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017, pp. 131–135.