動画データと画像キャプション生成を用いた音とテキストペアの自動生成

石川裕地 [1,2]　齋藤主裕 [1]　青木義満 [2]
[1]LINE ヤフー株式会社　[2] 慶應義塾大学
{yuchi.ishikawa, kasaito}@lycorp.co.jp　aoki@elec.keio.ac.jp

概要

本研究では, 動画データと画像キャプション生成モデルを活用した, 音とテキストのペアデータの自動生成手法を提案する. 提案手法は 3 段階で構成される. まず, 画像キャプション生成モデルを用いて動画のフレームごとにキャプションを生成する. 次に, これらのキャプションの選択/統合により動画全体を説明するキャプションを作成する. 最後に, 生成されたキャプションと音データのペアに対して, CLAP による類似度計算を用いてフィルタリングを行うことで, 高品質な音とテキストのペアを自動的に生成する. ClothoV2 と AudioCaps を用いた language-based audio retrieval タスクでの評価実験では, 提案手法で生成したデータによる学習が, 人手でアノテーションされたデータと同等の性能を達成することを確認した.

1　はじめに

音とテキストのマルチモーダル学習は, language-based audio retrieval [1, 2, 3], audio captioning [4, 5], text-to-audio generation [6, 7] など, 様々なタスクの基盤となる重要な研究分野である. このような音とテキストのマルチモーダル学習の実現には, 大規模かつ質の高い音とテキストのペアデータが必要不可欠である.

しかしながら, 高品質な音とテキストのペアデータの作成には, 多大なコストを要する. また, 公開されている音とテキストのペアデータセット [8, 9, 10] は, 画像とテキストのペアデータと比較して, その規模が大幅に劣るのが現状である. 例えば,language-based audio retrieval などで広く使用される ClothoV2 [8] には, 約 5,000 の音と 25,000 のテキストが含まれているが, これはマルチモーダルモデルの学習において必ずしも十分な規模とは言えない.

この問題を解決するために, 既存研究では, 一から音とテキストのペアデータを収集するのではなく, タグが付与された音データと, 大規模言語モデル (LLM; Large Language Model) [11, 12, 13, 14] を用いて, 音に対応するテキストを自動生成する手法が提案されている [15, 16, 17]. これらの研究は language-based audio retrieval などの下流タスクでの性能向上に大きく寄与した一方で, ペアデータの生成にはタグ付きの音データが必要である. またタグに含まれる情報以上の詳細な説明をテキストとして生成できない, という課題が存在する.

そこで本研究では, 音データを含む動画データと画像キャプション生成モデルを用いて, ラベルやタグのない動画データから, 音とテキストのペアを自動生成する手法を提案する. 提案手法の概要を図 1 に示す. 本手法では, まず音声を含む動画をフレームごとに分割し, 画像キャプション生成モデルを用いて, 各フレームのキャプションを生成する. 次に, 生成されたフレームごとのキャプションに対して, キャプションの選択や LLM などによるキャプションの統合処理を行い, 動画全体を説明するキャプションへと変換する. その後,CLAP (Contrastive Language-Audio Pretraining) [18] を用いて, 動画に含まれる音データと生成されたキャプションの類似度を計算し, 類似度の高いペアを音とテキストのペアとして採用する. 本手法は, タグやラベルを用いずに, 動画データだけから高品質な音とテキストのペアを生成でき, かつその生成プロセスが自動化されているため, データ作成のコストが抑えることができる.

実験では, 提案手法の有効性を示すために, 生成されたデータを用いて音とテキストの共通埋め込み空間を学習し, ClothoV2 [8] と, AudioCaps [9] の二つのデータセットを用いて language-based audio retrieval のタスクの性能評価を実施した. 実験を通して, 提案

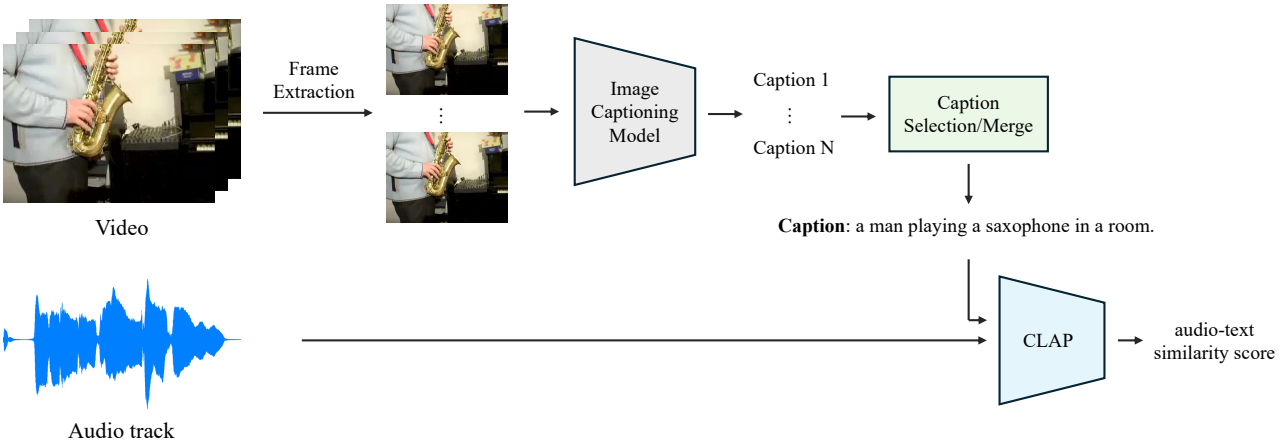

図 1　提案する音とテキストのペアの自動生成手法の概要図. 提案手法では, まず音データを含む動画に対して, 画像キャプション生成モデルを用いて, フレームごとの説明文を生成する. 次に, キャプションの選択・統合によって、動画を説明するテキストを作成する. このプロセスで得られたテキストを, 音データに対するテキストとみなし,CLAP の類似度スコアが高いものを, 質の高い音とテキストのペアとして採用する.

手法によって自動生成されたデータを用いた学習は, 人手でアノテーションされたデータセットを用いた学習と比較して, 同等の検索性能を達成することを確認した. また, 提案法で生成されたデータと, 人手でアノテーションされたデータを組み合わせて学習することで, 一部の指標で性能が向上することを確認した.

2　提案手法

本研究では, 高品質な音とテキストのペアを低コストで生成するために, 音データを含む動画データと画像キャプション生成モデルを用いた自動生成手法を提案する（図 1）. 本手法では, 動画から生成したキャプションを音データに対するテキストとして活用することで, 効率的な音とテキストのペアの生成を実現する.

提案手法は, 以下の 3 つのステップから成り立つ.

1. **Frame caption generation**: 音データを含む動画をフレームに分割し, フレームごとにキャプションを付与する
2. **Caption selection/merge** フレームごとのキャプションから,1 つのキャプションを選択, もしくは複数のキャプションを統合することで, 動画全体を説明するキャプションを生成する
3. **Audio-text filtering** 生成されたキャプションを音データの説明文として扱い, 音とキャプションの類似度に基づいて高品質なペアを選定する

なお, 動画へのキャプション付与（Frame caption generation）とキャプションの選択・統合（Caption selection/merge）のプロセスは, InternVid データセット [19] の構築手法を参考にしている. 次節からは, 各ステップの詳細について説明する.

2.1　Frame Caption Generation

まず, 動画データを一定間隔でサンプリングしてフレームに分割し, 各フレームに対して画像キャプション生成モデルを用いて, キャプションを生成する. 本研究では, 画像キャプション生成モデルとして,BLIP-2 [20] を採用する.

2.2　Caption selection/merge

次に, 生成されたフレームごとのキャプションから,1 つのキャプションを選択, もしくは複数のキャプションを LLM を用いて 1 つのキャプションに統合する. このキャプションを, 元の動画に含まれている音データのキャプションとして使用する. 本稿では, ベースラインの構築を目的として, 動画の先頭のフレームのキャプションを選択する手法を採用した.

2.3　Audio-text filtering

最後に, 生成された音とテキストのペアから高品質なものを選別するため, CLAP [18] を用いたフィルタリングを実施する. 具体的には, 音とテキスト間の類似度スコアを算出し, スコアの高いペアを高品質なペアとして採用する. 実験では, 全ペアを使用した場合, 上位 30%のペアを使用した場合, 上位 10%のペアを使用した場合, 上位 5%のペアを使用した場合の

図 2　提案手法で生成キャプションのうち, CLAP の類似度スコアが上位 10%のキャプション例. 類似度スコアが高いキャプションは, 楽器の演奏など音響イベントを含む動画から得られていることが確認できる.

4 条件で性能を検証する.

提案手法によって生成されたキャプションのうち, CLAP の類似度スコアが上位 10%に含まれるキャプション例を図 2 に示す. これらの例から,CLAP の類似度スコアが高いキャプションは, 楽器演奏をはじめとする音響イベントに関連する動画から生成されており, 音とテキストのペアとして高品質であることが確認できる.

3　実験と考察

3.1　実験設定とデータセット

提案手法の有効性を示すために, 生成された音とテキストのペアを用いて, 音とテキストの共通の埋め込み空間を CLAP で学習し, language-based audio retrieval のタスクで評価を行った. 実験は, [3] の設定に従い, audio encoder には,PaSST [21] を, text encoder には,RoBERTa [22] を使用した.

音とテキストのペアデータを生成するための動画データセットとして, Kinetics700 (K700) [23] を使用した. K700 は, 行動認識タスクのデータセットで, 700 種類の行動ラベルが付与された約 65 万動画から成る. 各動画は,YouTube から収集されており, 行動区間を含むように 10 秒程度の動画としてトリミングされている. 本実験では, 学習データと検証データに対して提案手法を適用し, 音とテキストのペアを生成した.

Language-based audio retrieval の性能評価には, ClothoV2 [8] と AudioCaps [9] を使用した. ClothoV2 は, 約 5,000 の音が含まれており, それぞれに対して人手で 5 つのキャプションが付与されたデータセットである. AudioCaps は, 約 48,000 の音を含み, それぞれに対して人手で 1 つのテキストが付与されたデータセットである.

評価指標は, 検索タスクで用いられる Recall at k (R@k), および mean Average Precision at top 10 (mAP@10) を使用した. R@k は, クエリに対して検索された上位 k 件の結果に正解が含まれる割合を表し, 本実験では $k = \{1, 5, 10\}$ とした. mAP@10 は, 上位 10 件の結果における平均適合率 (average precision) の平均値を表す指標である.

3.2　実験結果

表 1 に,language-based audio retrieval のタスクにおける性能比較の結果を示す. 実験結果から, 提案手法の有効性が複数の観点で確認された. まず,K700 から生成した音とテキストのペア (10%) を用いた学習は, 人手でアノテーションされたデータセット (AudioCaps + ClothoV2) での学習と同等の性能を達成した.

次に, 自動生成データのフィルタリングの割合を変えていった時に, 性能が大きく変化することが確認された. K700 データセットは全体の 10%を使用した場合に最も高い性能を示し, それ以上のデータ量（30%,100%）ではむしろ性能が低下する傾向が観察された. この結果は, 提案手法におけるフィルタリングの重要性を示唆している. 音データにはノイズが含まれていることが多く (例: 画面外から音が鳴っている, など), 動画から生成したキャプションと音データが, 綺麗に対応することは多くない. そのため, 単純にデータ量を増やすのではなく, CLAP による類似度に基づいて質の高いデータを選択するこ

表 1　AudioCaps と ClothoV2 における language-based audio retrieval の性能評価. 表の上段は, 人手でアノテーションされた音とテキストのペアで学習したモデルの性能を, 中段は, 提案手法によって自動生成された音とテキストペアで学習したモデルの性能を, 下段は, 両者を組み合わせて学習したモデルの性能を示している. 括弧内の数字は,CLAP によるフィルタリングによって選定されたデータの割合を示しており, 例えば K700(10%) の場合は,K700 から生成されたデータのうち, 類似度スコアの高い上位 10%のデータを指している.

Training Dataset	#data	AudioCaps R@1	AudioCaps R@5	AudioCaps R@10	AudioCaps mAP@10	ClothoV2 R@1	ClothoV2 R@5	ClothoV2 R@10	ClothoV2 mAP@10
AudioCaps	46k	**38.3**	**69.9**	**80.3**	**51.6**	13.9	35.7	48.6	23.2
ClothoV2	19k	12.9	36.2	50.3	23.0	15.6	42.2	57.7	26.9
AudioCaps + ClothoV2	65k	36.2	68.8	79.7	50.0	19.2	**47.0**	61.0	30.5
K700 (100%)	569k	4.0	13.7	21.3	8.2	4.4	14.3	22.1	8.9
K700 (30%)	170k	4.5	15.8	24.5	9.5	4.2	14.3	22.2	8.7
K700 (10%)	56k	35.6	67.0	78.3	49.0	16.2	42.0	55.2	27.3
K700 (5%)	28k	3.6	13.4	21.3	8.0	3.6	12.2	19.5	7.4
AudioCaps + ClothoV2 + K700 (30%)	235k	35.2	67.1	78.2	48.5	17.6	43.8	57.8	29.0
AudioCaps + ClothoV2 + K700 (10%)	121k	36.8	68.6	79.7	50.3	18.6	45.0	59.1	30.0
AudioCaps + ClothoV2 + K700 (5%)	93k	37.9	68.6	79.9	50.9	**19.3**	46.6	**61.8**	**31.2**

とが, モデルの性能向上に寄与することが明らかとなった.

一方で, 使用するデータ数を 10%よりも減らしていった場合も, 性能が下がっていくことが確認された. これは, 使用した動画データセットの特性に影響を受けていると考えられる. K700 は, 行動認識タスクのベンチマークであるが,「ギターを弾く」,「ハープを演奏する」など楽器演奏に関する行動が多く含まれている. 我々の分析では, これらに分類される動画から生成された音とテキストのペアは,CLAP の類似度スコアが高くなる傾向があることが確かめられた. そのため, スコアが上位のデータの多くは, 楽器演奏に関連する音響イベントが占めることになり, 結果データの多様性が損なわれ, 性能が低下したと考えられる. これを改善する方法として, より多様な動画を使用する, 頻度の高い単語を含むペアはサンプリングする, などの対処法が考えられる.

また, 提案手法で生成されたデータと, 既存のデータセットを組み合わせて学習を行った場合 (AudioCaps + ClothoV2 + K700) では, 既存のデータセットだけを用いた場合 (AudioCaps + ClothoV2) と比べて, ClothoV2 での性能が一部向上することが確認できた. これより, 提案手法によって生成されたペアデータが, データセットの多様性の向上に貢献していると考えられる.

4　おわりに

本研究では, 動画データと画像キャプション生成モデルを活用して, 音とテキストのペアデータを自動で生成する手法を提案した. Language-based audio retrieval での実験によって, 提案手法で生成されたデータセットで学習したモデルが, 人手でアノテーションされたデータセットで学習したモデルと同等の性能を達成することを確認した. また, 生成されたデータと既存のデータセットを組み合わせることで, 一部の指標においてモデルの性能が向上することも示された.

本稿では, ベースラインの構築を目的として, 動画データセットして Kinetics700 のみを使用し, 動画の 1 フレーム目だけを使用してキャプションを生成するなど, シンプルな枠組みでの検証を行った. 今後の課題として, より多様な動画データの活用や, LLM などの高度な言語モデルの導入により, 高品質かつ多種多様な音とテキストペアを生成を目指す.

参考文献

[1] A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. **IEEE Transactions on Multimedia**, Vol. 25, pp. 2675–2685, 2022.

[2] Paul Primus, Florian Schmid, and Gerhard Widmer. Estimated audio–caption correspondences improve language-based audio retrieval. In **Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)**, pp. 121–125, Tokyo, Japan, October 2024.

[3] Hokuto Munakata, Taichi Nishimura, Shota Nakada, and Tatsuya Komatsu. Pre-trained models, datasets, data augmentation for language-based audio retrieval. In **Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)**, pp. 86–90, Tokyo, Japan, October 2024.

[4] Xinhao Mei, Xubo Liu, Mark D Plumbley, and Wenwu Wang. Automated audio captioning: An overview of recent progress and new challenges. **EURASIP journal on audio, speech, and music processing**, Vol. 2022, No. 1, p. 26, 2022.

[5] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. In **ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 6735–6739. IEEE, 2024.

[6] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. **arXiv preprint arXiv:2301.12503**, 2023.

[7] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In **International Conference on Machine Learning**, pp. 13916–13932. PMLR, 2023.

[8] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 736–740. IEEE, 2020.

[9] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 119–132, 2019.

[10] Irene Martín-Morató, Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, Maximo Cobos, and Francesc J Ferri. Sound event envelope estimation in polyphonic mixtures. In **ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 935–939. IEEE, 2019.

[11] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[12] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, Vol. 1, , 2020.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, Vol. 35, pp. 27730–27744, 2022.

[15] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, 2024.

[16] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. **arXiv preprint arXiv:2307.16372**, 2023.

[17] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-acd: A large-scale dataset for audio-language representation learning. In **Proceedings of the 32nd ACM International Conference on Multimedia**, pp. 5025–5034, 2024.

[18] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In **ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 1–5. IEEE, 2023.

[19] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. **arXiv preprint arXiv:2307.06942**, 2023.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In **International conference on machine learning**, pp. 19730–19742. PMLR, 2023.

[21] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. **arXiv preprint arXiv:2110.05069**, 2021.

[22] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, Vol. 364, , 2019.

[23] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. **arXiv preprint arXiv:1907.06987**, 2019.