6T-03

Listen and Tell: 深層学習を用いた音響シーンのキャプション生成

岩月道生[1], 周藤唯[1], 糸山克寿[1], 西田健次[1], 中臺一博[1,2]
1 東京工業大学工学院システム制御系
2 ホンダ・リサーチ・インスティチュート・ジャパン

1 はじめに

本稿では，深層学習を利用して音響信号からその内容を説明するキャプションを生成する手法について報告する．音響信号に対する認識については，音の種別を認識する音源同定や，音声を対象とした音声認識といった研究が行われている．しかし音響信号はこれら以外にもタイミングや位置情報など多様な情報を持っている．こうした情報も含めた総合的な音響シーンの理解ができれば，音響信号の検索などに役立つ．本稿では，音響シーンの表現形式として，音源の種類と順序を自然言語で説明するキャプションを生成する．

2 提案法

画像に対するキャプション生成については深層学習を使った Show and Tell [1] をはじめとした報告がある．この画像に対するキャプション生成手法を参考に，音響信号に対応したキャプション生成モデルを提案する．画像と音響信号は 1) 入力形式が時系列信号かどうか，2) 入力信号長が固定か可変かという主に 2 つの性質の違いがあるので，画像からキャプションを生成する手法をそのまま音響信号に適用することはできない．このため，前者の問題に対しては音響信号のスペクトログラム表現を用いて入力信号の形式を画像に合わせる，後者の問題に対しては，複数のスペクトログラムを用いることにより可変長音響信号を疑似的に表現することで解決を図った．具体的な手法を以下に示す．

2.1 音響信号を対象としたキャプション生成手法

Show and Tell [1] をはじめとする深層学習を用いた画像からのキャプション生成モデルは，画像を CNN (Convolutional Neural Network) を用いて固定長の中間ベクトル表現に変換し [2]，その中間ベクトル表現を RNN (Recurrent Neural Network) を用いてキャプションに変換する [3] という構造をしている．

このモデルを基に音響信号のキャプション生成モデルを考えるとき，前述の 2 つの問題が発生する．1 つ目の入力信号形式の違いについては，画像は色情報を含めた 3 次元のデータであるが，音響信号は 1 次元である．よって音響信号をメル周波数スペクトログラムに変換することによって 1 チャンネルのグレースケール画像として扱えるようにし，CNN に入力できるようにした．2 つ目の問題は，画像の場合はリサイズによってデータの大きさを統一して用いているのに対し，音響信号は時系列の可変長データであるという点である．そこで波形データを一定間隔ごとに固定長で切り取り，切り出した複数の波形データそれぞれについてスペクトログラムを計算，最終的にそれらを RNN を用いて一括して入力することにより，固定サイズの画像のシーケンスとして，可変長音響信号を表現する手法を導入した (Fig. 1).

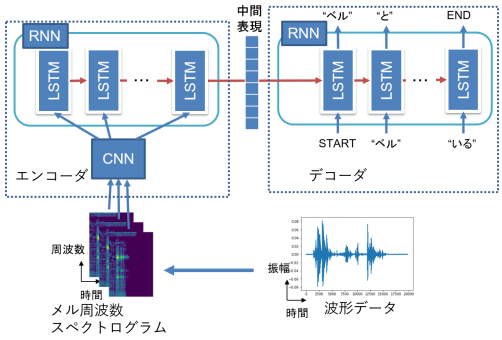

Fig. 1 音響信号のキャプション生成モデル (提案法)

3 評価実験

提案モデルの有効性を示すためにデータセットを作成し，評価実験を行った．

3.1 データセット

提案モデルの学習と評価にあたって音響信号とキャプションのデータセットが必要になる．よって，単一クラスのみが含まれた音源をランダムに合成した混合音と，それに対応するキャプションのデータセットを作成した (Fig. 2). 音源には FSDKaggle2018 Dataset [4] の 41 クラスを使用し，混合音は 3 つ程度の音源をオバーラップがないように接続して作成した．ただし，音源信号間には 0〜0.6 秒程度のランダムな長さの無音信号を挿入した．今回のタスクでは，モノラル入力音響信

Michio Iwatsuki[1], Yui Sudou[1], Katsutoshi Itoyoma[1], Kenji Nishida[1], Kazuhiro Nakadai[1,2]
1 Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
2 Honda Research Institute Japan Co., Ltd.

号に対して種類と順序を説明するキャプションを生成するものとした．混合音とキャプションのペア 18000 個分を学習データセット，2000 個分を評価データセットとした．

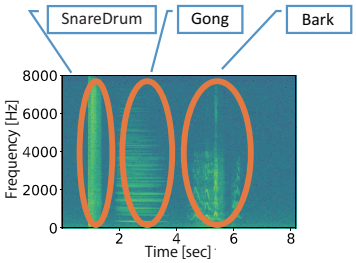

Fig. 2　混合音とそのキャプションデータ

3.2　学習と実験条件

モデルの学習は 2 段階に分けて行う．最初に音源識別用の CNN を用意する．これには，既存のトレーニング済みのモデル [5] を用いた．次にこの CNN に対して，3.1 節で作成した学習データセットを用いてモデル全体の転移学習を行った．なお波形データは 1.28 秒間ずつ，50%オーバーラップするように切り取った．

3.3　実験結果

まず音源の順序が正しく表現できているかをみる．正解と完全に一致したキャプションが生成された割合は 71.6%となった．一方，音源の順序を問わず，キャプションに含まれる音源の種類が一致したものは 73.5%となった．以上より，音源の種類は当たっているが，順序が誤っているものは 1.9%のみであり，音源の順序情報を表現できていることが分かる．

次に含まれる音源の種類が正しく表現できているかをみる．正解と一致しなかった出力キャプションの例を Table 1 で示す．このようにエラーの原因は音源識別間違いと音源数の間違いに分けられる．そこでキャプションに含まれるラベル単語のみを抜き出してラベル列を作成し，そのラベル列に対して挿入誤りの数，削除誤りの数，置換誤りの数を計算した．全評価データ 5969 個のうち，挿入誤りの数: 40，削除誤りの数: 20，置換誤りの数: 598 となった．置換誤りが多いことから音源の有無よりも音源識別での間違いが多いことが分かる．

4　まとめ

本研究では，深層学習を用いて音響信号からその内容を説明するキャプションを生成する手法を提案した．画像でのキャプション生成モデルを音響信号でのモデルに適用するために，音響信号に対するスペクトログラム表現と，複数のスペクトログラムを用いた可変長音響信号に対する固定長ベクトル表現を導入することによってモデルを拡張した．提案モデルを混合音とキャプションのデータセットで学習させた結果，完全に一致したキャプションが生成される割合は 71.6%となり，また，音源の順序が表現できていることが分かった．

今後は音響信号に含まれる音源の種類・順序だけでなく，タイミングや音源定位と組み合わせた音源位置の情報の表現にも取り組む予定である．

Table 1　出力例

正解キャプション	出力キャプション
Writing の後に MicrowaveOven, その後に Telephone が鳴っている	Writing の後に MicrowaveOven, その後に Saxophone が鳴っている
Fireworks の後に Shatter が鳴っている	Fireworks の後に Fireworks, その後に Shatter が鳴っている

謝辞　本研究は，JSPS 科研費 16H02884，16K00294，17K00365 および，JST ImPACT タフロボティクスチャレンジの助成をうけた．

参考文献

[1] Vinyals, Oriol, et al. ”Show and tell: A neural image caption generator.” Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164). 2015.

[2] Sermanet, Pierre, et al. ”Overfeat: Integrated recognition, localization and detection using convolutional networks.” arXiv preprint arXiv:1312.6229. 2013. arXiv:1312.6229. 2013.

[3] Bahdanau, Dzmitry, et al. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473. 2014.

[4] Fonseca, Eduardo, et al. ”General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline.” arXiv preprint arXiv:1807.09902. 2018.

[5] Il-Young, Jeong, et al. ”Audio tagging system for DCASE 2018: Focusing on label noise, data augmentation and its efficient learning.” DCASE2018 Challenge. 2018.