Audio Captioning モデルの発達的カリキュラム学習

◎稲垣賢斗，高道慎之介（慶大）

1. はじめに

- オーディオキャプショニングとは，環境音や生活音などの入力音に対して，その内容を説明するキャプションを生成する機械学習タスクである．

→これは言語情報とオーディオ情報の**クロスモーダル**な接続を要求する．

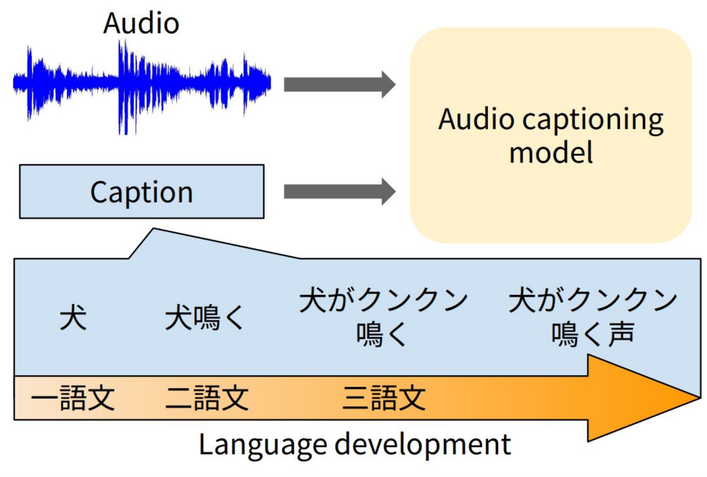
- クロスモーダルな情報処理は，人間が現状の機械学習モデルよりも優れている点の一つ

→**人間の学習メカニズム**を応用できないだろうか？

- 人間の学習過程には，意味のある順序で段階的に情報が提示されることで効率が高まる．

→これに倣った機械学習手法である，**カリキュラム学習**[1]

- 本研究ではオーディオキャプショニングにおいて，**人間の言語発達に倣った段階的なキャプション**の学習データを提示することでモデルの性能や学習過程にどのような変化が生じるかを調査する．



2. 提案手法：発達的カリキュラム学習

- 幼児の言語発達に倣った，**段階的に文長・文法複雑性が上昇するキャプション**を学習ステップに応じて提示する．

→学習初期はキーワードのみ，学習が進んでいくごとに完全な文に近づく．

第一段階　epoch ~20
ex) “大人、猫”
幼児の発話の最初期の一語発話[2]をモデリング．元のキャプションからイベントごとに名詞を一つ抽出する．

第二段階　epoch 20~40
ex) “大人話し、猫鳴く”
二語発話[2]をモデリング．（名詞＋動詞）や（形容詞＋名詞）を抽出．

第三段階　epoch 40~60
ex) “大人の女性が話し、猫がニャーと無く”
文法的な構成が始まる段階をモデリング．第二段階に一単語加えた三語文を構成し，加えて助詞と助動詞を追加．

第四段階　epoch 60~100
ex) “大人の女性が話している声，猫がニャーと鳴く声”
元の完全なキャプション．

3. 実験的評価

3.1 実験条件

- 使用モデル：CNext-trans[3]
  - 事前学習済み畳み込みエンコーダー＋6層のTransformerデコーダー
  - DCASE 2024 Task 6 のベースラインモデル
- データセット：Multi-lingual AudioCaps[4]
  - 10秒程度の環境音と日本語キャプション
  - train: 45,205　validation: 440　test: 883
- 3つの比較手法
  - **カリキュラム無し**：終始同じデータセットで学習
  - **従来手法**：キャプションからストップワード(for, do, theなど)を削除して予測難易度を調整するカリキュラム学習[5]
  - **提案手法**：本研究の発達的カリキュラム学習

3.2 実験結果

- BLUE[6], ROUGE-L[7]，BERTScore[8]において**最も高いスコア**
- CLAIR-A[9]では**最も低いスコア**

最終エポック(epoch 100)でのモデルの評価指標スコア

カリキュラム	BLUE	ROUGE-L	BERTScore	CLAIR-A
無し	0.313	0.501	0.834	0.648
従来手法	0.312	0.502	0.835	**0.650**
提案手法	**0.322**	**0.508**	**0.836**	0.641

単語の一致度　意味の一致度　大規模LLMが評価

3.3 評価・考察

- 下表上の例のように，カリキュラム無しよりも音響イベントについて**正しく記述できている**サンプルが多く見られた．
- 一方で，下表下の例のような**文法エラー**が多く見られた．

→CLAIR-Aでは文法エラーもスコアに含まれる．CLAIR-Aのみで低スコアだったのは文法エラーが原因だと考えられる．

- **提案手法によってオーディオと単語の対応は改善した一方，文法のエラーが増えた．**

最終エポックにおける生成キャプションの例

正解キャプション	カリキュラム無し	提案手法
女性の歌声。その後で、咳。それに続いて、鳥たちがチーチー鳴く声	男性が話している声。その後に、赤ちゃんの泣き声	鳥たちがチューチュー鳴く声
女性の話している声。それに続いて、磁器の皿がガチャンと鳴る音、食べ物と油がシューシューいう音	女性が話している声、同時に食べ物を炒める音	女性が話している声、同時に**食べ物を炒められる**音

3.4 学習過程の観察

- 学習初期において，カリキュラム無しでは教師データに頻出するパターンを一様に出力されている一方，提案手法では正解キャプションに含まれる単語を出力している．

→このような学習過程の違いが最終的なモデルの性能に影響する．

epoch 10 における生成キャプションの例

正解キャプション	カリキュラム無し	提案手法
食べ物が炒められる音、女性の話している声	男性が話している声	女性、水
エンジンがゴロゴロ鳴る大きな音。その後で、エアホーンが鳴る音が三度	男性が話している声	エンジン、車両
子供の泣き声、男性と女性の話している声。それに続いて、車のドアが開く音。その後で、閉まる音	男性が話している声、それに続い始める	子供、女性

Reference

[1] Y. Bengio, ICML, 2009
[2] 小椋 たみ子, 日本言語学会, 2007
[3] E. Labbe, IEEE, 2024
[4] 岡本 悠希, 言語処理学会, 2024
[5] A. Koh, APSIPA ASC, 2022
[6] K. papineni, ACL, 2002
[7] C-Y. Lin, ACL, 2004
[8] T. Zhang, ICLR, 2020
[9] T-H. Wu, arXiv, 2024