AUTOMATED AUDIO CAPTIONING

Technical Report

Arunodhayan Sampathkumar, Danny Kowerko

Technische Universität Chemnitz, Juniorprofessur Media Computing, Chemnitz, Germany
{Arunodhayan Sampathkumar, Danny Kowerko}@informatik.tu-chemnitz.de

ABSTRACT

The audio captioning is a novel approach to describe an audio scene based on human like perception. The human like perception of audio events not only perform detection and localization, but also tries to summarize the relationship between different audio events. The DCASE2020 has developed a strongly labelled caption dataset to perform automated audio captioning. In this research, mel spectrogram is used to extract the audio features. A Recurrent Neural Network (RNN) encoder-decoder is employed to train the dataset. Finally the network is evaluated using the MS COCO metrics where *BLEU3&BLEU1* scores were strong and is discussed in detail in section 5.

Index Terms— Audio captioning, RNN, Natural Language

1. INTRODUCTION

In recent years there has been a steady growth in advancement in audio classification which mainly focuses on speech and music processing [1]. The current state of art audio focuses on classification, detection and localization of acoustic events [2]. Acoustic scenes in real time has multiple overlapping sound events [2]. Humans not only detect and classify sound events, we inspect the inner relationship between the individual sounds and describe in natural language. This super ability is more challenging for machine perception.

The classification of human-made acoustic events is important for the monitoring and recognition of human activities or critical behavior. In our experiments on acoustic event classification for the utilization in the sector of health care, we defined different acoustic events which represent critical events for elderly or people with disabilities in ambient assisted living environments or patients in hospitals. This contribution presents our work for acoustic event classification using deep learning techniques. We implemented and trained various convolutional neural networks for the extraction of deep feature vectors making use of current best practices in neural network design to establish a baseline for acoustic event classification. We convert chunks of audio signals into magnitude spectrograms and treat acoustic events as images. Our data set contains 20 different acoustic events which were collected in two different recording sessions combining human and environmental sounds. Our results demonstrate how efficient convolutional neural networks perform in the domain of acoustic event classification [1], [3], [4].

Automated audio captioning is an exciting area of research which focuses on generating automatic textual description for an audio signal. The automated audio captioning method does not predict the sound event rather generates textual description for a given audio signal. For example, if an audio signal contains a laughing event or drinking event, it generate a textual description like "the people are laughing" ,"A person opens a canteen, quickly gulps the water and then closes the canteen." [5]

The automated audio captioning is inspired from different multimedia captioning methods. Image captioning was the first attempt to create a caption based on human like perception, using natural language processing. As a follow up of this research, automated video captioning was developed [6]. The image and video captioning was developed based on the existing works in machine translation [7]. The automated image and video captioning was divided into two methods namely encoder and decoder. The encoder processes the input example data, images or frames to create a feature representation using Convolutional Neural Network(CNN) [5]. Later, decoder take the input data and helps in generating sequences using RNN namely Long Short Term Memory(LSTM) or Gated Recurrent Unit(GRU). In some research implementations, multi-layer RNN was used for encoder. In the cited works, decoder uses multi-layer or single-layer RNN follwed by Fully Connected Layer(FC) [7].

In this research paper and automatic audio captioning algorithm, based on a neural network, is developed using the clotho dataset [5]. The dataset consist of an audio file with a generated textual description which is as close as human assigned captions. The trained network is evaluated using the Microsoft COCO caption evaluation namely *BLEU*, $ROUGE_L$, *METEOR* and *CIDEr* [8]. The description of the evaluation metric is discussed in this section 5. The research paper is organized as follows: experimental setup in section 2, data preprocessing in section 3, network architecture in section 4, evaluation metric and results in section 5 & 6

2. EXPERIMENTAL SETUP

Our hardware and software configuration is described in the following bullet points.

- Hardware Configuration:
  - Intel Core i7
  - 8 GB RAM
  - Operating system: Ubuntu 18.04 LTS
  - CPU (4 cores @ 2.40 GHz)
  - NVIDIA Titan X with 12 GB graphics memory
- Software:
  1. *Programming Language:*

- Python 3.7

2. *Libraries:*
   - CUDA 10.0 tool kit for GPU acceleration
   - CUDnn 7.5.4
   - Pytorch=1.4.0
   - Torchvision=0.4.1
3. *Dependencies:*
   - numpy 1.14.2
   - scipy 1.0.0
   - python speech features 0.6
   - pydub 0.21.0
   - python-openCV=4.2.0
   - librosa=0.7.2
   - matplotlib 2.2.0
   - cmake 3.5.1
   - cython 0.29.2

3. DATA PRE-PROCESSING

Caption preprocessing: The Clotho dataset has three splits namely the development set consisting of 2893 audio files, evaluation and test set consisting of 1045 audio files, respectively [5]. Each audio file has been associated with a minimum of 5 to 6 captions. In most cases captions are not proper sentences but a set of keywords. In a caption, some keywords repeat more than once and in different form (i.e singular or plural). Repeated keywords of the same form are removed, i.e. "Rain Rain" one word is removed resulting in "Rain". The punctuations are removed in order to reduce the unique words or tokens which are difficult to predict. Most of the words in captions have typographic errors which may lead to a high number of unique words. This error is resolved by introducing a python package named *Enchant* which removes all the words which do not belong to US or UK dictionaries.
Audio Feature Extraction: The feature extraction for the audio file is performed using a *Librosa* python package [9]. Each recording is 30 seconds long on average with a sampling rate of 44.1kHz, sample width of 16 bits and single channel. We used log-mel spectrogram with FFT length 512, melband 64, hanning window and hop size 256 into a spectrogram of size (melbands x frames) as given in the Figure 1.

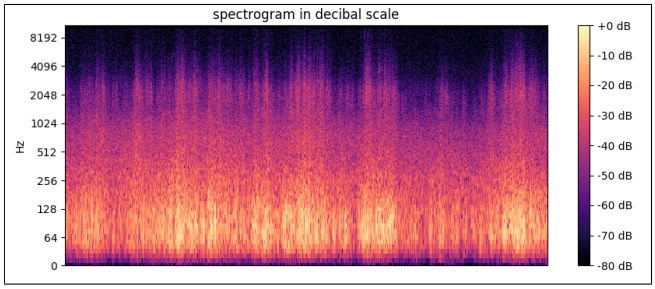

Figure 1: Extracted spectrogram of 25 seconds length consisting of a person drinking sound. For details, see text.

4. NETWORK ARCHITECTURES

From the state of art, combinations of CNN and RNN performed well for image and video captioning. The neural network consists of an encoder and a decoder as presented in the Figure 2. The encoder is a two layered or one layered bidirectional GRU with Relu as activation function and a dropout at the end of the encoder. For one layered bidirectional GRU the input size is 64 cells, 256 hidden cells and 512 output cells. For two layered bidirectional GRU the input size is 512 cells, 256 hidden cells and 512 output cells. The network is inspired from the baseline provided [10].

The decoder is of two layered with a dropout, single directional GRU and a fully connected classifier layers. For the single directional GRU, the input size is the output size from the encoder that is 512 cells and 256 output cells with relu as activation function. The FC has 256 input cells and 4367 output cells with a linear activation function. The dropout size of encoder and decoder is 0.1 and 0.2, respectively. The network is trained for 200 epochs with a batch size of 24.

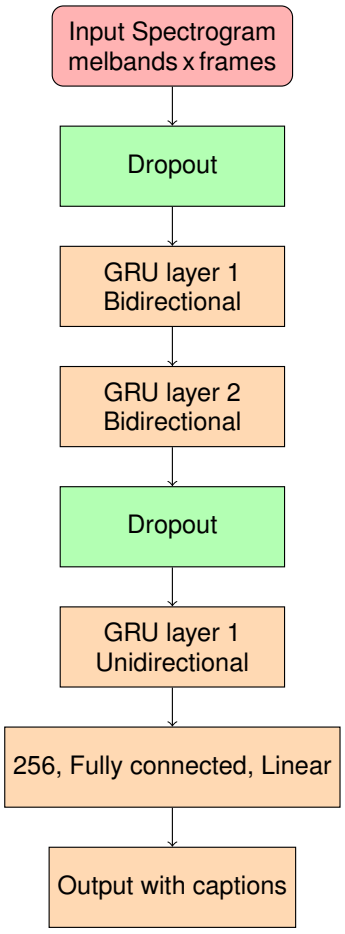

Figure 2: Network architecture for audio captioning.

5. EVALUATION METRICS

To evaluate the quality of captions the following metrics are followed to evaluate the model. BLEU, $ROUGE_L$, METEOR, CIDEr

are the evaluation metrics used in this research.

BLEU: BLEU is a machine translation metrics that analyses the co-occurrences between the candidate and the reference statements. BLEU is computed using weighted geometric mean. BLEU has performed good for corpus level comparisons with a high level of matches.

$\boldsymbol{ROUGE_L}$: $ROUGE_L$ is designed to evaluate summarization texts. It uses a measure based on Longest Common Subsequence (LCS). The LCS measure the F-score between the pair of sentences. The F-score is measured by calculating the precision and recall.

METEOR: METEOR is calculated in the candidate and reference sentences by generating an alignment between the words. This alignment is based on matching the exact token from the WordNet synonyms, stemmed tokens and from the paraphrases.

CIDEr: This metric was designed to evaluate the consensus in image captioning by performing a Term Frequency Inverse Document Frequency (TF-IDF) by weighting for each n-grams of words.

The cited research work explain in detail regarding the evaluation metrics [8].

6. RESULTS

The dataset consisting of English captions and the results are evaluated based on the test dataset provided. The metrics used for evaluating the dataset was discussed in section 5. The table presents the results of evaluation and test dataset comparing it with the baseline. The BLEU3 & BLEU1 has achieved better scores when compared with the baseline.

Evaluation Metrics	Evaluation dataset	Baseline
$BLEU_1$	0.4417	0.389
$BLEU_2$	0.1397	0.136
$BLEU_3$	0.365	0.055
$BLEU_4$	0.0094	0.015
$METEOR$	0.0876	0.084
$ROUGE_L$	0.2543	0.262
$CIDER$	0.070	0.074
$SPICE$	0.0251	0.033
$SPIDER$	0.061	0.054

Table 1: Audio captioning results compared with baseline.

7. CONCLUSION AND FUTURE WORK

In this paper we present an automated audio captioning method to provide captions for audio signals. We present an approach where audio signals were converted to log-mel spectrograms which serve as feature used in an RNN and predict the caption for each audio file. The results obtained using this method is better when compared with the baseline results. We were able to improve the baseline results for most of the evaluation metrics given in this challenge to a minor extend. The highest improvement was achieved for Bleu3.

8. REFERENCES

[1] A. Sampath-Kumar, R. Erler, and D. Kowerko, "CNN-based Audio Classification for Environmental Sounds, Ambient Assisted Living and Public Transport Environments using an Extensive Combined Dataset," in *Chemnitzer Informatik Berichte 2020*, vol. CSR-20-01. Chemnitz: Universitätsbibliothek/Universitätsverlag, Jan. 2020, pp. 29–66.

[2] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," *CoRR*, vol. abs/1902.09254, 2019. [Online]. Available: http://arxiv.org/abs/1902.09254

[3] "Acoustic Event Classification Using Convolutional Neural Networks," Chemnitz. [Online]. Available: https://dl.gi.de/handle/20.500.12116/3989

[4] A. Sampath Kumar, R. Erler, and D. Kowerko, "A real-time demo for acoustic event classification in ambient assisted living contexts," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2205–2207. [Online]. Available: https://doi.org/10.1145/3343031.3350600

[5] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," *CoRR*, vol. abs/1910.09387, 2019. [Online]. Available: http://arxiv.org/abs/1910.09387

[6] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *CoRR*, vol. abs/1412.4729, 2014. [Online]. Available: http://arxiv.org/abs/1412.4729

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CoRR*, vol. abs/1411.4555, 2014. [Online]. Available: http://arxiv.org/abs/1411.4555

[8] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

[9] B. McFee, M. McVicar, O. Nieto, S. Balke, C. Thome, D. Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, D. Ellis, F.-R. Stoter, D. Repetto, S. Waloschek, C. Carr, S. Kranzler, K. Choi, P. Viktorin, J. F. Santos, A. Holovaty, W. Pimenta, and H. Lee, "librosa 0.5.0," Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.293021

[10] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," *CoRR*, vol. abs/1706.02291, 2017. [Online]. Available: http://arxiv.org/abs/1706.02291