History of Medicine Studies

ISSN:1300-669

Volume 14 Issue 1 Jan 2018

AUDIO CAPTIONING REINFORCEMENT LEARNING AT CRNN AND GRU

Dr G Hemalatha[1,] Dr. M Venkata Narayana[2] Dr V Adinarayana Reddy [3], Dr S Zahiruddin[4]

ABSTRACT

The goal of audio captioning is to produce a natural sentence that accurately describes the content of an audio recording. The authors of this study suggest a powerful CRNN encoder coupled with a GRU decoder as a means of approaching this multi-modal challenge. Cross-entropy is not the only method being looked at for producing high-quality captions; reinforcemint learning is also being explored. Our method outperforms the reference model by at least 34% relative improvement on all measures shown. On the Clotho evaluation set1, the Spider score achieved by our proposed CRNNGRU model using reinforcement learning is 0.190. The efficiency was raised to 0.223% once data augmentation was used. We maintained a minimal model size (only 5 million parameters) and placed fourth in the DCASE competition Task 6 based on Spider and second on 5 metrics including BLEU, ROUGE-L, and METEOR without using an ensemble or data augmentation. Reinforcement learning, convoluted recurrent neural networks, and closed-captioning audio are among topics covered in this article.

INTRODUCTION

Joint learning across several modalities is required for the difficult job of automatic captioning. Image captioning, for instance, requires feature extraction and the use of a landgauge model to create descriptive phrases. Similar to audio captioning, video captioning uses feature learning to create captions based on a temporal sequence of visuals and sound. In contrast to the disciplines of images and videos, however, audio captioning receives very less attention [1]. Captioning is an innovative multi-modal activity since it uses text to describe an aural situation in great detail. Audio captioning differs from problems like sound or acoustic event detection, which are solely concerned with single-label estimate of an event, by focusing on the production of rich sentences that accurately and correctly describe an audio. Applications in audio surveillance, automated content description, and machine-to-machine interaction based on the content of a video all stand to benefit greatly from audio captioning. The commercial Propounds Effects [2] audio corpus was used in the pioneering work on audio captioning published in [1]. In this study, a three-layer bidirectional gated recurrent unit (Bigram) encoder and two-layer Bigram decoder were used. The encoder phrase is summed up with the help of attention pooling. Subsequent research in [3] explored audio captioning within the context of Chinose captioning, first proposing a public captioning corpus, centered on talks inside a healthcare facility. Their findings demonstrated that an encoder-decoder GRU network can effectively create audio captions within a constrained region. However, they questioned the efficacy of the standard measures used to assess machine translation's effectiveness. Despite reaching near-human performance on objective measurements, the key argument is that the produced phrases are generally less meaningful in human judgement. Exposure bias is present in audio captioning just as it is in other text generation tasks like machine translation and picture captioning. To better predict the next ground-truth word given the present ground-truth word, neural network-based models are often trained in a "teacher forcing" approach. While ground-truth annotations are only accessible during training, models may use their own predictions of the current word to infer the next word during inference. This causes a buildup of mistakes throughout testing. Misalignment between training goal and assessment measure is another issue in text creation activities. Discrete metrics like BLEU [4], ROUGE-L [5], Cider [6], and METEOR [7] are often used to assess generative models. These non-differentiable metrics, however, cannot be Optimazed in the conventional back-propagation fashion. Evidence from prior research indicates that maximizing both the continuous and discrete assessment measures via Reinforcemint Learning (RL) may help mitigate the effects of exposure bias. In [8], it is suggested that RL be used to teach NLG models how to generate natural language. A generative model is used as the agent, while language and context are considered the external setting

.

[1,2,3] Professor, Department of ECE, K. S. R. M College of Engineering(A), Kadapa

[4] Associate Professor, Department of ECE, K. S. R. M College of Engineering(A), Kadapa

The model's parameters form a policy, and the current produced word is selected in accordance with that policy. The sampled sentence's assessment score (BLEU, METEOR, Cider, etc.) provides the incentive. The gradient of the agent parameters is estimated by utilizing policy-gradient [9]. In order to limit the large variation in rewards, the work in [10] uses the rewards from greedily sampled words as a starting point. The value of the created words is estimated not by randomly selecting from the action space, but rather by using actor-critic approaches [12], as in subsequent work [11]. In this study, we investigate the feasibility of applying the SCST method (first presented in [10]) to the field of audio captioning. We propose our CRNN-based encoder-decoder solution to audio captioning in Section 2 of this work. In Section 3, the front-end characteristics and model parameters used in the experiments are presented. Dissected in our investigation and findings include

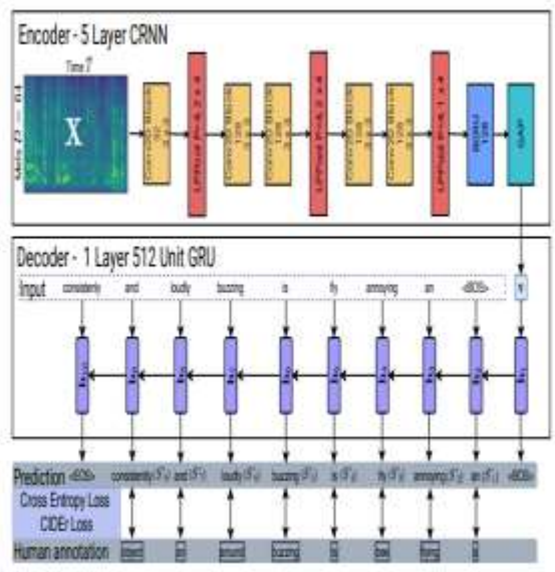

Figure 1: Our proposed encoder-decoder architecture. The encoder is a CRNN model which outputs a fixed-sized 256-dimensional embedding v after a global average pooling layer (GAP).

A convolution block refers to an initial batch normalization, then a convolution, and lastly, a Leakier (slope −0.1) activation. The numbers in each block represent the output channel size and the kernel size. For example, ”32, 3 × 3” means the convolution layer has 32 output channels with a kernel size of 3 × 3. All convolutions use padding in order to preserve the input size. Then a GRU decoder utilizes this audio embedding v or embedding of the word S 0 t at each time-step, to predict the next word S 0 t+1.

APPROACH

Similar to previous audio captioning frameworks [3], our approach follows a standard encoder-decoder model (see Equation (1)).

$$\begin{aligned} \mathbf{v} &= \mathrm{Enc}(\mathbf{X}) \\ [S'_1, \ldots, S'_T] &= \mathrm{Dec}(\mathbf{v}) \end{aligned} \tag{1}$$

The encoder (Enc) is fed an audio-spectrogram (X) and produces a fixed-sized vector representation v, which the decoder (Dec) uses to predict the caption sentence. Specifically, the decoder generates a single word-token S 0 t for each time-step it up until an end of sentence () token is seen (see Figure 1). In audio captioning, decoding differs between training and evalauction stages:

$$\ell_{\mathrm{XE}}(\theta; S, \mathbf{v}) = -\sum_{t=1}^{T} \log p(S_t|\theta; \mathbf{v}) \tag{2}$$

Dec generates word-tokens during training using the embedding v and human-annotated data S under the supervision of a cross-entropy (XE) loss (see Equation (2)), when transcriptions are available. In the absence of transcriptions, word tokens are randomly picked from the decoder based on the audio embedding v during testing and assessment. As can be seen from the above explanation, the quality of v has a direct bearing on the quality of the resulting phrase. Thus, the encoder design and loss function are the primary areas in which our method departs from the prior art. We replace the conventional GRU encoder with a resilient convolutional recurrent neural network (CRNN) since previous encoder models (GRU) may not be enough to provide a robust vector representation. You can see our overall structure in Figure 1. Furthermore, there may be drawbacks to the typical XE training. One problem is that the criteria doesn't take into account context when comparing words. Second, since each word is processed separately, sentences with syntactic errors might be produced. Third, since the model must faithfully mimic a phrase word for word, opitemizing XE always results in repetitive sentences rather than permitting semantically comparable but differently worded seentenses.

For the purpose of audio captioning, we apply reinforcement learning. When using a measure (such BLEU or Cider) as a reward, we may immediately back-propagate it via reenforcement learning. In a nutshell, we formalize our training by reducing the model's error rate in response to a single sampled seentense S 0:

$$\ell_{\mathrm{RL}}(\theta;\mathbf{v}) = -r(S'), S' \sim p(S'|\theta;\mathbf{v}) \qquad (3)$$

where S 0 = [S 0 1, S0 2, . . ., S0 T]. By incorporating the policy gradient method with baseline normalization, the parameter gradients can be estimated as follows:

$$\nabla_\theta \ell(\theta;\mathbf{v}) = -(r(S') - b)\nabla_\theta \log p(S'|\theta;\mathbf{v}), S' \sim p(S'|\theta;\mathbf{v}) \qquad (4)$$

here b is a pre-defined baseline normalization constant to reduce the high variance brought by sampling [12]. We set b as the greedy decoding reward because of its effectiveness in image captioning [10].

Models

Encoder Our proposed encoder is a CRNN model, which has seen success in localizing sound events [13, 14]. The architecttrue consists of a five-layer CNN (utilizing 3 × 3 convolutions), summarized into three blocks, with L4-Norm pooling after each block. The CNN blocks subsample the temporal dimension by factor of 4. A Bigram is attached after the last CNN output, endhanding our model's ability to localize sounds accurately. At last, we use a global average pooling (GAP) layer in order to remove any time-variability to a single, time-independent representation v ∈ R 256. The encoder has 679k parameters, making it comarally light-weight while only using 2.7 MB on disk. Decoder In the context of audio captioning, a decoder takes a fixed-sized embedding and aims to produce a sentence. We use a single-layer GRU with 512 hidden units as our decoder model.

EXPERIMENTS

Dataset

The challenge provides Clotho [2, 15] for the audio captioning task. It contains a total of 4981 audio samples, where the duration is unitfirmly distributed between 15 to 30 seconds. All audio samples are collected from the Freedsound platform. Five native English speakerrs annotate each sample; thus, 24905 captions are available in total. Captions are post-processed to ensure each caption has eight to 20 words, and the caption does not contain unique words, named entiaties or speech transcription. The dataset is officially split into three sets, termed as development, evaluation, and testing, with a ratio of 60%-20%-20%. In the challenge, the development and evaluatetin sets are used for training our audio captioning model while the testing set is for evaluating the model.

Data pre-processing

We extract 64-dimensional log-Mel spectrogram (LMS) as our default input feature. Here a single frame is extracted via a 2048-point Fourier transform every 20 ms with a Hann window size of 40 ms. This results in a X ∈ R T ×D log-Mel spectrogram feature for each input audio, where D = 64 and T is the number of frames. Moreover, the input feature is normalized by the mean and standard deaviation of the development set. For each caption in the dataset, we remove punctuation and convert all letters to lowercase to reduce the vocabulary size. To mark the beginning and the end of sentences, we add special tokens “” and “” to captions. The available training data is split into a model training part, consisting of 90% of available data and a held-out 10% validation set.

Evaluation metrics

A total of eight objective metrics is utilized to evaluate our modelgenerated captions: BLEU@1-4 grams [4], METEOR [7], RougeL [5], Cider [6] and SPICE [16]. A further Spider metric is callcollated as the mean of Cider and SPICE.

Training details

We submit predictions from four models to the challenge:

- CRNN-B (Base). This is our baseline CRNN-GRU encoderdecoder model.

- CRNN-W (Word). Here, the decoder word-embeddings are initialized from Word2Vec word-embeddings trained on the deelopement set captions.

- CRNN-E (Ensemble). Here we fuse CRNN-B and CRNN-W results on output level.

- CRNN-R (Reinforcement). Here we finetune CRNN-W using reinforcement learning. The details for each submission are elaborated in the following. XE training For XE training, teacher forcing is used to accelerate the training process. We evaluate the model on the validation set at each epoch and select the best model according to the highest BLEU4 score. We train the model for 20 epochs and use Adam [17] optimizer with an initial learning rate of 5 × 10−4. The batch size is 32. According to whether Word2Vec is used for word embedding initialization, we get CRNN-B and CRNN-W respectively. Ensemble In order to further enhance performance we merge the outputs of CRNN-B and CRNN-W on word-level. The encoded audio representation v is fed to both CRNN-B and CRNN-W to obtain two-word probabilities p1 and p2.

We ensemble the two models, which means the current word is decoded according to the mean of p1 and p2. Then the current word embedding is fed to CRNN-B and CRNN-W to obtain the next word until is generated. Reinforcement The CRNN-R approach is first initialized by training a CRNN-W model using the standard XE criterion. This model is then finetuned using reinforcement learning, as seen in Section 2, by optimizing the Cider score using policy gradient with baseline normalization. Although [21] optimized Spider by policy gradient in image captioning, we choose Cider as the trainIng objective because Cider optimized model trained by SCST achieved better performance [10]. Cider measures sentence simhilarity through representation by n-gram TF-IDFs while BLEU ofcusseson" hard" n-gram overlaps. Such a" soft" similarity (Cider) may be a better optimization objective compared with BLEU under the condition that one audio corresponds to several semantic similar sentences, possibly composed of different n-grams. The model is trained for 25 epochs using Adam optimizer with a learning rate of 5 × 10−5. Similar to the practice in XE training, we report the best model based on the Cider score on the validation set.

RESULTS

Results

Our results on the Clotho evaluation set are displayed in Table 1 and compared with the DCASE challenge baseline, which consists of a three-layer Bigram encoder and two-layer Bigram decoder. As it can be seen, our initial CRNN-B model largely outperforms the baseline, indicating that a potent encoder is indeed beneficial towards captioning performance. By initializing word embeddings with Word2Vec trained on the development set captions, CRNN-W gets a slight performance improvement in most metrics compared with CRNN-B, except Cider and METEOR. CRNN-E improves performance against both CRNN-B and CRNN-W. Our best performing model is CRNN-R. Interestingly, although CRNN-R is opitemized towards Cider score, the relative improvement in BLEU3 and BLEU4 are more significant than Cider. The improvement in ROUGEL and METEOR is not as significant as other metrics. However, CRNN-R does achieve the best performance in terms of all evaluation metrics, which validates the effectiveness of reinforcemint learning for audio captioning with regards to the official challenge evaluation, our CRNNR achieves the fourth place in DCASE2020 task 6 on the Clotho testing set. However, there is only a slight difference between our submission and the submission ranking the third (0.194 / 0.196).

CONCLUSION

In this paper, we propose a novel audio captioning approach untillazing a CRNN encoder front-end as well as a reinforcement learnIng framework. Audio captioning models are trained on the Clotho dataset. The results on the Clotho evaluation set suggest that the CRNN encoder is crucial to extract useful audio embeddings for captioning while reinforcement learning further improves the perromance significantly in terms of all metrics. Our approach ranked fourth in the DCASE2020 task 6 challenge testing set with a computative result on all metrics except Cider. Notably, our approach is the best performing non-ensemble result without data augmentstation, with the least parameters (5 million). By further utilizing Specie data augmentation, we observe an additional boost in regrads to the Spider score on the evaluation set from 0.190 to 0.223.

REFERENCES

[1] K. Dross's, S. Advance, and T. Virtanen, "Automated audie captioning with recurrent neural networks," in 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct 2017, pp. 374–378.

[2] S. Lipping, K. Dross's, and T. Virtanen, "Crowdsourcing a dataset of audio captions," in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Nov. 2019. [Online]. Available: https: //arxiv.org/abs/1907.09238

[3] M. Wu, H. Dinkel, and K. Yu, "Audio Caption: Listen and Tell," in ICASSP, IEEE International Conference on Acorustics, Speech and Signal Processing - Proceedings, vol. 2019- May. Institute of Electrical and Electronics Engineers Inc., may 2019, pp. 830–834.

[4] K. Papini, S. Roukoops, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, pp. 311–318.

[5] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," Proceedings of the workshop on text summarizetion branches out (WAS 2004), 2004.

[6] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.

[7] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.

[8] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," arXiv preprint arXiv:1511.06732, 2015.

[9] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in Advances in neural information processing systems, 2000, pp. 1057–1063.

[10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.

[11] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio, “An actor-critic algorithm for sequence prediction,” arXiv preprint arXiv:1607.07086, 2016.

[12] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[13] H. Dinkel and K. Yu, “Duration Robust Weakly Supervised Sound Event Detection,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, may 2020, pp. 311–315. [Online]. Available: https://ieeexplore.ieee.org/document/9053459/

[14] H. Dinkel, Y. Chen, M. Wu, and K. Yu, “GPVAD: Towards noise robust voice activity detection via weakly supervised sound event detection,” mar 2020. [Online]. Available: http://arxiv.org/abs/2003.12222

[15] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: An audio captioning dataset,” in 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, May 2020. [Online]. Available: https://arxiv.org/abs/1910.09387

[16] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” CoRR, vol. abs/1607.08822, 2016. [Online]. Available: http://arxiv.org/abs/1607.08822

[17] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[18] Y. Wu, K. Chen, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, “Audio captioning based on transformer and pre-training for 2020 DCASE audio captioning challenge,” DCASE2020 Challenge, Tech. Rep., June 2020.

[19] H. Wang, B. Yang, Y. Zou, and D. Chong, “Automated audio captioning with temporal attention,” DCASE2020 Challenge, Tech. Rep., June 2020.

[20] Y. Koizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, “The NTT DCASE2020 challenge task 6 system: Automated audio captioning with keywords and sentence length estimation,” DCASE2020 Challenge, Tech. Rep., June 2020