GUIDED CAPTIONING OF AUDIO

Irene Martín-Morató, James Afolaranmi, Annamaria Mesaros

Signal Processing Research Centre, Tampere University, Finland
{irene.martinmorato@tuni.fi, james.afolaranmi@tuni.fi, annamaria.mesaros@tuni.fi

ABSTRACT

This work introduces a guided captioning system that aims to produce captions focused on different audio content, depending on a guiding text. We show that using keywords guidance results in more diverse captions, even though the usual captioning metrics do not reflect this. We design a system that can be trained using keywords automatically extracted from reference annotations, and which is provided with one keyword at test time. When trained with 5 keywords, the produced captions contain the exact guidance keyword 70% of the time, and results in over 3600 unique sentences for Clotho dataset. In contrast, a baseline without any keywords produces 700 unique captions on the same test set.

Index Terms— automatic audio captioning

1. INTRODUCTION

Automatic audio captioning (AAC) is a cross-modal task combining audio signal analysis and natural language processing [1]. Captioning differs from other audio analysis tasks such as detection or classification because it requires not only identification of the sounds, but also a description of the relationships between co-occurring events. Textual descriptions provide more information about the audio content than simple labels, indicating for example which sounds are more prominent and which ones are background, how sounds co-occur or follow each other, or describe attributes, e.g. how loud/quiet or far/near the sound is.

What defines a good caption is subject to the specific situation. Generally speaking, sensory descriptions have as primary function transmitting the main information, which for audio captioning is likely be the main sound event; but the way this information is included in a caption is very subjective [2]. AAC datasets provide captions for training the systems, one or multiple captions per clip [3–5], reflecting to some extent the fact that different descriptions of the same audio clip are correct, even though not identical.

AAC systems are trained in a supervised manner, being fed with the audio file and its corresponding reference captions [6, 7]; evaluation is performed by comparing an automatically predicted caption against the reference captions, to measure how well the predicted caption matches each of the reference captions. Researchers have questioned the use of machine translation or image captioning metrics for evaluating audio captions, because the auditory, temporal and spatial properties of the sound are not the same as objects' properties. As a result, multiple captioning metrics were proposed specifically for AAC, e.g. FENSE [8], SPICE+ [9], CB-score [10], SPIDEr-max [11]. However, the status quo in AAC is still dominated by small training datasets, limited vocabulary, and unclear interpretation of the metrics.

The concept of "guiding text" for captioning has been investigated in [12]; the authors proposed "conceptual captions", where a provided text controls what an image captioning system should focus on. A similar approach was used in [13] for AAC; the authors used a transformer with keyword estimation to generate a caption that contains the estimated keyword. In [14], the authors used keywords estimated from the given audio clip through automatic audio tagging. Furthermore, Xu et.al. [15] focus on improving diversity of the captions without decreasing accuracy. These works focused on improving AAC performance as evaluated with the usual AAC metrics. However, captions containing words from the reference vocabulary will usually have high scores, even if they do not completely describe the audio content; moreover, high semantic similarity does not necessarily reflect the true correspondence with the described sounds, as observed in [10].

The contributions of this paper are as follows: (1) a guided captioning system that can be trained with an arbitrary list of keywords per audio clip; (2) a systematic study of the effect of keywords on the predicted captions. Rather than improving AAC performance in terms of the usual metrics, as done in previous works, we focus on guiding the system towards a specific sound event of interest: a user interested in one particular event will provide a keyword as guidance in order to obtain a description of the specific event. This description may be only partial as to the acoustic content of the clip, but correct and desired by the user. Our experiments show that given the same audio clip as input, it is possible to

This work was supported by Jane and Aatos Erkko Foundation under grant number 200061, "Guided captioning for complex acoustic environments". The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources.

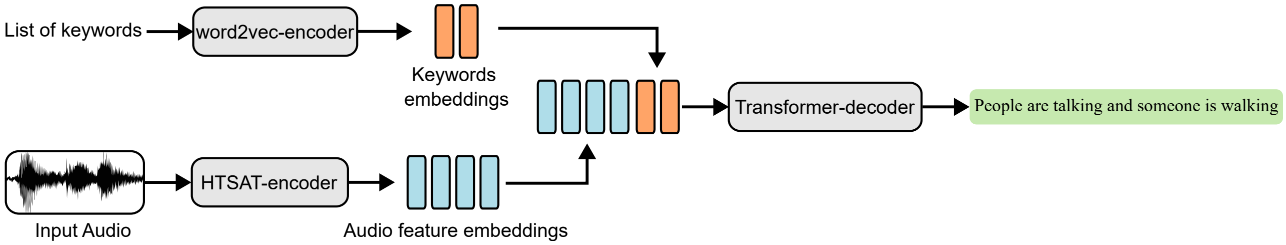

Figure 1: Guided captioning: keywords and audio are provided to the captioning system to produce a caption that is focused on the specific event given as keyword.

produce different captions depending on the provided guidance keyword, resulting in a significantly diverse set of captions compared to a system without keywords.

The paper is organized as follows: Section 2 introduces the concept of automatic audio captioning with keyword guidance, Section 3 presents the datasets and experimental setup; section 4 includes the discussion of the obtained results; section 5 presents conclusions and future work.

2. CAPTIONING WITH KEYWORDS GUIDANCE

The block diagram of the guided audio captioning system is presented in Fig 1. The system consists of two encoders, one for the keywords and another for the audio. The text encoder receives as input a list of keywords and will provide textual guidance to the model in the form of text embeddings; the audio encoder receives as input the raw audio signal to be transformed into audio feature embeddings. As a text encoder we trained a Word2Vec [16] model on the vocabulary of the dataset used in each experiment. To obtain the feature embeddings from the raw audio, we use the HTSAT transformer model [17] which is pre-trained on AudioSet.

The output of the text encoder, representing the keyword embeddings, is concatenated at the end of the audio embeddings obtained at the output of the audio encoder, forming the input for the transformer-decoder. The transformer-decoder has a standard architecture, as in [14], and is followed by a fully connected linear layer that outputs word probability. It has two hidden layers with a dimension of 768 and uses GELU activation functions in the feed-forward process between the hidden layers. The output of the transformer generates the captions based on the combined information from text and audio. The vocabulary used for training the model was collected from the reference captions for each dataset separately. KeyBERT [18] was used to extract N keywords for each clip, representing the words that best describe the reference captions.

The model was trained from scratch as opposed to using a pre-trained model, to accommodate for the concatenation of text embeddings and audio embeddings before decoding. For testing the model using textual guidance, we used two different setups: (1) using N keywords at once for guidance, same as in the training; and (2) using one keyword at a time. For the second setup, each clip is tested multiple times, each time with a different keyword. The keywords used in testing are obtained from the reference captions using the same procedure as for training, therefore they represent correct acoustic content for each clip.

3. EXPERIMENTAL SETUP

We use three datasets for our experiments, Clotho [3], collected based on Freesound [19] content, MACS [5], which contains audio clips of everyday acoustic environments, and AudioCaps [4], a subset of AudioSet [20].

Clotho contains 5929 recordings of 15 to 30 seconds long, each audio clip having five reference captions. We extract five keywords from the captions using KeyBERT ($N = 5$). The experiments are run on the development set of Clotho using the provided training/validation/test split. MACS contains 3930 recordings from TAU Urban Acoustic Scenes 2019 development dataset, from three acoustic scenes, each file being 10-seconds long. Captions and tags were collected at the same time for the data. A list of tags was provided to annotators to indicate what sounds they hear in the clip, after which they were asked to provide a one-sentence description. Here, we can use the tags provided by annotators as keywords (so N varies from 1 to 7 per clip). The experimental split is created based on the TAU Urban Acoustic Scenes Development set, with the included clips. AudioCaps contains 51308 clips, of which only 46721 are available now[1]. From these, 886 clips are used for testing. Because AudioCaps annotators had access to the AudioSet tags, we have tags available for the clips and can use them to guide the AAC system.

When using tags as keywords, it is important to note that: (1) the tags are not necessarily keywords that are extracted from the captions; (2) the tags can be single words (music) or compound terms (dog barking); (3) the number of tags per clip varies, so in this case N words provided as guiding text will be the number of tags for each clip. For Clotho we use $N = 5$ for all clips.

[1] Audio clips downloaded June 2024.

Training keywords	Guidance keywords	BLEU_1	BLEU_4	CIDEr	SPIDEr	% exact	% synonym	unique captions
None	None	56.24	15.19	39.35	26.21	-	-	737
kBERT 1	kBERT 1	58.99	17.14	47.02	30.22	47.08	11.77	774
kBERT 5	kBERT 5	66.13	20.65	63.64	**40.16**	44.30	11.40	944
kBERT 1	1 (all)*	57.73	16.38	41.31	27.14	40.52	10.37	2463
kBERT 5	1 (all)*	56.85	14.30	39.35	25.98	**72.94**	3.50	**3673**

* Five keywords extracted with kBERT for a clip are provided as guidance one at a time.

Table 1: Guided captioning results on CLOTHO dataset for different training and test setups: baseline (no keywords) and using 5 keywords extracted with keyBERT (kBERT). The main setup of the guided captioning system is highlighted with light gray.

The datasets differ on the number of unique captions (Clotho: 29614, MACS: 10594, AudioCaps: 47737), and the lexical diversity of the datasets also varies. The moving average type-to-token ratio (MATTR) [21] using a window of 500 tokens is 0.385 for Clotho, 0.302 for MACS and 0.415 for AudioCaps, indicating a richer vocabulary for the latter.

The main setup of the proposed system is to train it with the available N keywords per clip, and test it with one keyword as guidance. Each test audio clip is repeatedly tested with different keywords, and the produced captions are evaluated independently. This is marked in the tables in gray. As an ablation study, we compare the results with different setups. We first construct a baseline system as a plain AAC system using the same architecture but trained and tested without any keywords or guidance. We also train and test the system with only one keyword per clip, and train and test with N keywords at once. For MACS and AudioCaps, the ablation experiment also includes using for guidance all available tags per clip (variable N) in addition to the experiment with $N = 5$ keywords extracted with KeyBERT.

4. RESULTS AND DISCUSSION

The results of the system on Clotho are presented in Table 1. The performance of the baseline (None/None combination, on row 1), are aligned with the performance presented in the DCASE Challenge, placing the system around 6th place in the 2023 challenge. Training and testing with one keyword results in a significantly higher CIDEr and SPIDEr than of the baseline, which is further markedly improved when the system is trained and guided with 5 keywords at the same time. When the guidance goes through all keywords one at a time (lower half in Table 1), the system performs comparable with the baseline which does not use any guidance. However, in this experimental setup there are 5 times more test cases, because each clip is tested 5 times (once with each keyword). The advantage brought by using the most representative keyword per clip is lost when the averaging is done over all keywords, since there is more variety in the n-grams content of the predictions. Similarly, there is much less overlap in n-grams between captions containing one keyword compared to (potentially) five.

However, if we look at the generated captions, we observe that with different keywords the system produces a much higher number of unique sentences. To quantify the effect of the keywords guidance, we include to Table 1 the % of the times the generated caption contains the exact match of the guidance keyword or a synonym of it, respectively. When guided with 5 keywords, the % exact is calculated as the proportion of keywords present in the caption (so 1 of 5 counts as 20%). The keywords are most often present in the predicted caption exactly as such, rather than a synonym, due to the limited vocabulary of the system.

Results for AudioCaps and MACS are presented in Table 2. Ablations include the use of tags and KeyBERT produced keywords as guidance (none and five). When using tags, the number of keywords is equal to the number of tags available for each clip. While the numbers differ, the behavior is similar to what we observed on Clotho: guidance with five keywords at test time gives the best AAC metrics performance, while training with five and guiding with one keyword has similar AAC performance as the baseline (no guidance) but a much higher number of unique sentences. Particularly, for the case of AudioCaps, we achieve a SPIDEr score of 62.43% with 875 unique captions for 886 test audio files. Guiding the captioning process with a single keyword results in better scores when using tags rather than the KeyBERT keywords, but produces more repetitive captions, as shown by the smaller number of unique sentences. For MACS, the difference is not significant in CiDEr and SPIDEr score, likely due to the reduced lexical diversity and smaller vocabulary than the other datasets.

Table 3 provides a few examples of captions generated by the different setups for a clip in Clotho. It is evident that the use of keywords results in sentences containing the provided keywords. While the baseline produces a caption containing as much information as possible, the guided captions refer to different aspects of the environment through the keywords:

Dataset	Training keywords	Guidance keywords	BLEU$_1$	BLEU$_4$	CIDEr	SPIDEr	% exact	% synonym	unique captions
AudioCaps	None	None	69.92	27.74	72.50	45.39	-	-	608
	Tags	Tags	71.59	28.47	77.48	48.17	47.50	15.10	612
	kBERT	kBERT	86.82	33.17	102.4	**62.43**	75.28	10.34	**875**
	Tags	1 (all)*	70.80	26.90	69.04	43.65	46.82	13.26	1086
	kBERT	1 (all)*	51.08	14.50	36.66	23.62	96.27	0.11	1325
MACS	None	None	73.38	22.27	29.78	22.87	52.61	2.67	235
	Tags	Tags	75.61	24.83	32.43	24.44	53.75	2.77	173
	kBERT	kBERT	73.36	24.66	40.49	**28.77**	43.02	7.50	**523**
	Tags	1 (all)*	75.10	24.06	28.71	22.16	51.30	3.58	441
	kBERT	1 (all)*	69.51	20.32	29.86	22.41	48.60	4.80	1301

* All keywords for a clip are provided as guidance one at a time.

Table 2: Guided captioning results on AudioCaps and MACS datasets for different training and test setups: baseline (no keywords), using metadata labels (Tags) and using 5 keyBERT extracted labels as keywords (kBERT).

Keyword	Generated Caption
-	an announcement is made over a loudspeaker while people are talking in the background
crowded	people are talking in a **crowded** area and walking in a crowded area
restaurant	people are talking and moving in a **restaurant**
crowd	a **crowd** of people are talking in an enclosed space
busy	people are talking in a **busy** area with each other in the background
eating	a person is **eating** something and people are talking in the background

Table 3: Example captions generated for *je_PittsPhipps.wav* file in the CLOTHO dataset: baseline (no keywords) and guidance with 5 different keywords. Two of five reference captions for this clip contain the word "restaurant".

scene (restaurant), attributes (busy, crowded), sound sources (crowd, eating). When evaluated with the captioning metrics, the caption produced by the baseline has the potential to being scored higher than the others due to containing more n-grams. On the other hand, there are specific terms, in this example "restaurant", not picked up by the baseline. This is a good example of guidance, the focus on specific content instead of producing a generally good description. However, the guided captions quite often contain repeated keywords; the system likely requires a more careful optimization of the training process w.r.t. the length of the generated sentences. In this work, we kept the training process the same for all the scenarios, not optimizing them separately.

To verify the effect of random keywords on the guided captioning system, we feed as guidance five keywords that are not related to the content of the clip. The SPIDEr scores for the three datasets with this setup are 18.7 for Clotho, 10.7 for AudioSet and 20.5 for MACS, all smaller than the equivalents that are guided with correct keywords (the kBERT/kBERT line in the tables). Furthermore, if the system is not provided any keyword at test time, its SPIDEr scores are 18.8, 21.3 and 20.5, respectively, showing that the guidance does have a quantifiable effect on the system output.

5. CONCLUSIONS

This paper presented a guided captioning system to enhance the relevance of certain audio events in the generated caption. As a design choice, the system is not optimized for typical AAC metrics, and instead it focuses on user-provided keywords, which the AAC metrics fail to adequately evaluate. Because describing audio content is subjective to the annotator perception of the acoustic environment, there may be multiple correct ways to describe the same content; AAC metrics evaluate the largest overlap and penalize automatic captions with lesser content. Our focus on directing the system towards user-requested events intends to reduce this requirement. We demonstrated the system's capability to produce a diverse set of descriptions aligned with the provided keyword. Future work will focus on better evaluating the captioning outputs based on the guidance keyword, since finding matching n-grams is not sufficient, nor necessary.

6. REFERENCES

[1] X. Mei, X. Liu, M. D. Plumbley, and W. 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.

[2] B. Winter, *Sensory Linguistics: Language, perception and metaphor*, ser. Converging Evidence in Language and Communication Research. John Benjamins, Apr. 2019.

[3] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an Audio Captioning Dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[4] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. of the 2019 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 119–132.

[5] I. Martín-Morató and A. Mesaros, "Diversity and bias in audio captioning datasets," in *Proceedings of the 6th Workshop on DCASE*, Nov. 2021, pp. 90–94.

[6] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A chatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–15, 2024.

[7] F. Gontier, R. Serizel, and C. Cerisara, "Automated audio captioning by fine-tuning BART with AudioSet tags," in *Proceedings of the 6th Workshop on DCASE*, Barcelona, Spain, November 2021, pp. 170–174.

[8] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?" in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 981–985.

[9] F. Gontier, R. Serizel, and C. Cerisara, "Spice+: Evaluation of automatic audio captioning systems with pre-trained language models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] I. Martín-Morató, M. Harju, and A. Mesaros, "A summarization approach to evaluating audio captioning," in *Proceedings of the 7th Workshop on DCASE*, Nov. 2022, pp. 116–120.

[11] E. Labbé, T. Pellegrini, and J. Pinquier, "Is my automatic audio captioning system so bad? SPIDEr-max: A metric to consider several caption candidates," in *Proceedings of the 7th Workshop on DCASE*, Nancy, France, November 2022.

[12] E. G. Ng, B. Pang, P. Sharma, and R. Soricut, "Understanding guided image captioning performance across domains," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend, Eds. Association for Computational Linguistics, Nov. 2021, pp. 183–193.

[13] Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito, "A transformer-based audio captioning model with keyword estimation," in *Proc. Interspeech 2020*, 10 2020, pp. 1977–1981.

[14] X. Mei, X. Liu, H. Liu, J. Sun, M. Plumbley, and W. Wang, "Automated audio captioning with keywords guidance," DCASE2022 Challenge, Tech. Rep., May 2022.

[15] X. Xu, M. Wu, and K. Yu, "Diversity-controllable and accurate audio captioning based on neural condition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 971–975.

[16] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*, 2013.

[17] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 646–650.

[18] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT." 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265

[19] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM Int. Conf. on Multimedia (MM'13)*. Barcelona, Spain: ACM, Oct. 2013, pp. 411–412.

[20] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[21] M. Covington and J. McFall, "Cutting the gordian knot: The moving-average type-token ratio (MATTR)," *Journal of Quantitative Linguistics*, vol. 17, pp. 94–100, 05 2010.