Is my automatic audio captioning system so bad? SPIDEr-max: a metric to consider several caption candidates

Etienne Labbé, Thomas Pellegrini, Julien Pinquier

Is my automatic audio captioning system so bad?
SPIDEr-max: a metric to consider several caption candidates

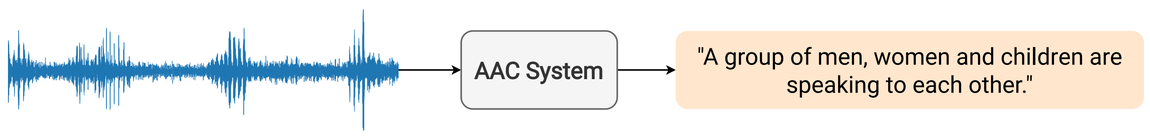
Étienne Labbé, Thomas Pellegrini, Julien Pinquier

IRIT, Université Paul Sabatier, CNRS, Toulouse, France
{etienne.labbe, thomas.pellegrini, julien.pinquier}@irit.fr

Introduction

Automated Audio Captioning task (AAC)

- Describe audio events using natural language
- Cross-modal task between audio event recognition and text generation
- Application for hearing impaired people, information retrieval, human-machine interaction or surveillance



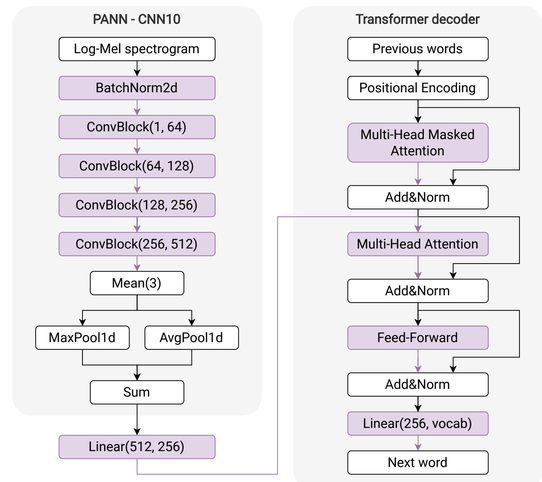
System description

Datasets

- AudioCaps [1]: 46K audio training files (126h) from AudioSet
- Clotho [2]: 3.8K audio training files (24h) from Freesound

Model

- CNN10 from PANN [3] encoder + Transformer decoder [4]
- Beam search decoding with log-probabilities selection



SPIDEr score limitations

- **CIDEr-D metric** [5]: Cosine-similarity of TF-IDF scores for 1-grams to 4-grams
- **SPICE metric** [6]: F-score of semantic propositions extracted from candidate and references
- **SPIDEr metric** [7]: Average of CIDEr-D and SPICE scores
- **Log-probs:** Sum of the log-probabilities given by the model over the sentence size.

Beam search candidates (beam size=5)	CIDEr-D	SPICE	SPIDEr	Log-probs
heavy rain is falling on a roof	0.724	0.400	0.562	-1.018
heavy rain is falling on **a tin roof**	**1.359**	**0.500**	**0.930**	**-0.898**
a heavy rain is falling on a roof	0.787	0.400	0.594	-0.996
a heavy rain is falling on the ground	0.403	0.267	0.335	-1.047
a heavy rain is falling on the roof	0.787	0.400	0.594	-1.079

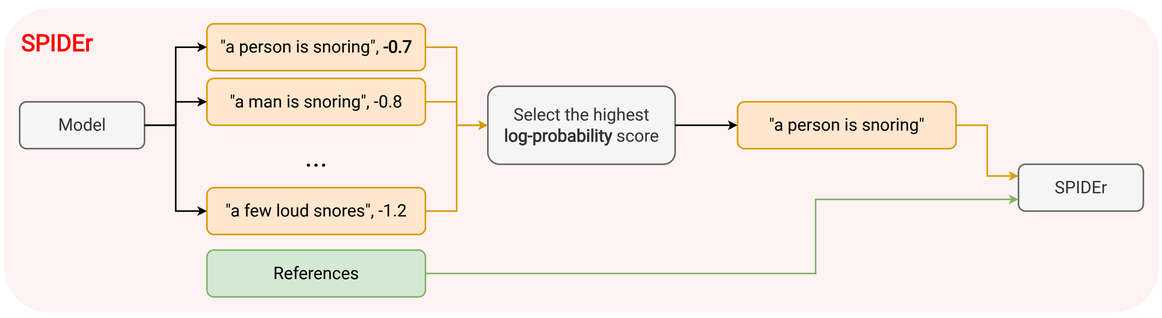
References captions
heavy rain falls loudly onto a structure with a thin roof
heavy rainfall falling onto a thin structure with a thin roof
it is raining hard and the rain hits **a tin roof**
rain that is pouring down very hard outside
the hard rain is noisy as it hits **a tin roof**

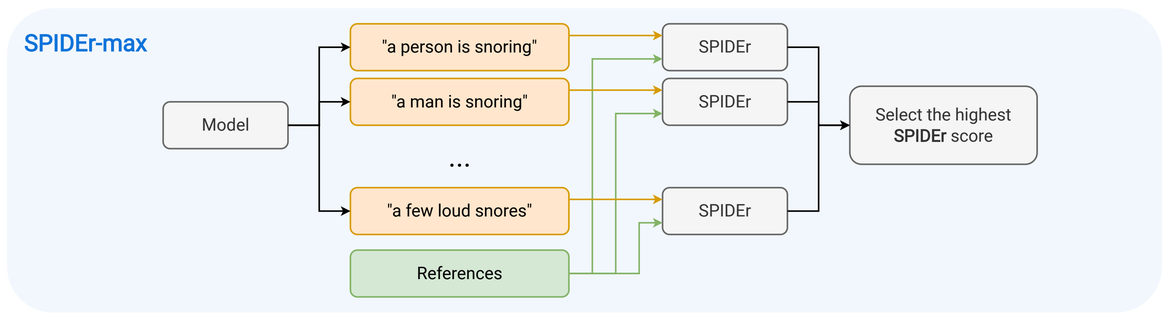
Beam search candidates (beam size=5)	CIDEr-D	SPICE	SPIDEr	Log-probs
a woman speaks and a sheep bleats	0.380	0.000	0.190	-0.745
a woman **speaks and a goat bleats**	**2.308**	**0.211**	**1.259**	-0.767
a man speaks and a sheep bleats	0.477	**0.211**	0.344	-0.768
an adult male speaks and a sheep bleats	0.361	0.100	0.231	-0.799
an adult male is speaking and a sheep bleats	0.282	0.095	0.189	**-0.712**

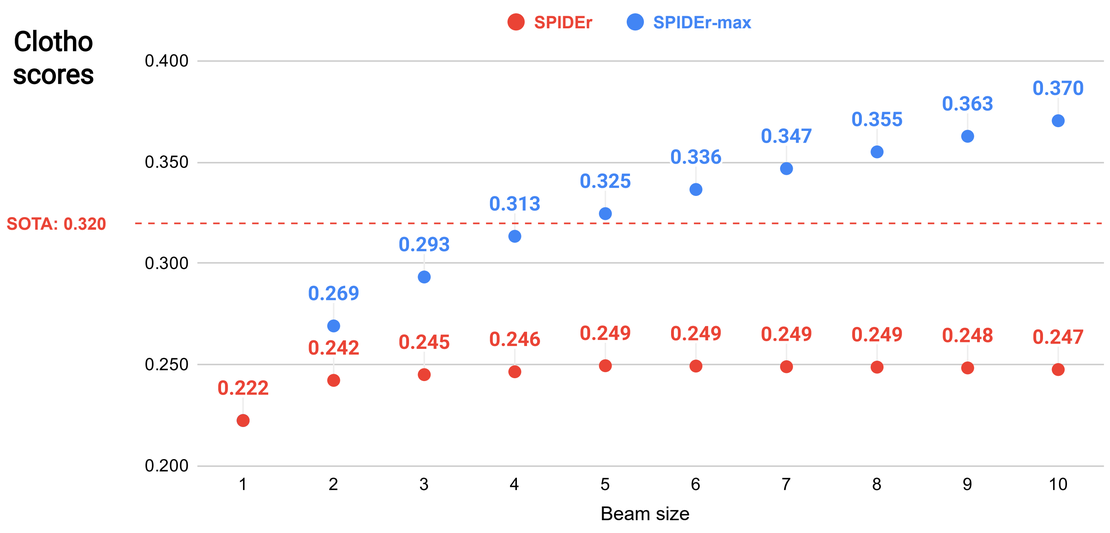
References captions
a man speaking and laughing followed by a goat bleat
a man is speaking in high tone while a goat is bleating one time
a man speaks followed by a goat bleat
a person **speaks and a goat bleats**
a man is talking and snickering followed by a goat bleating

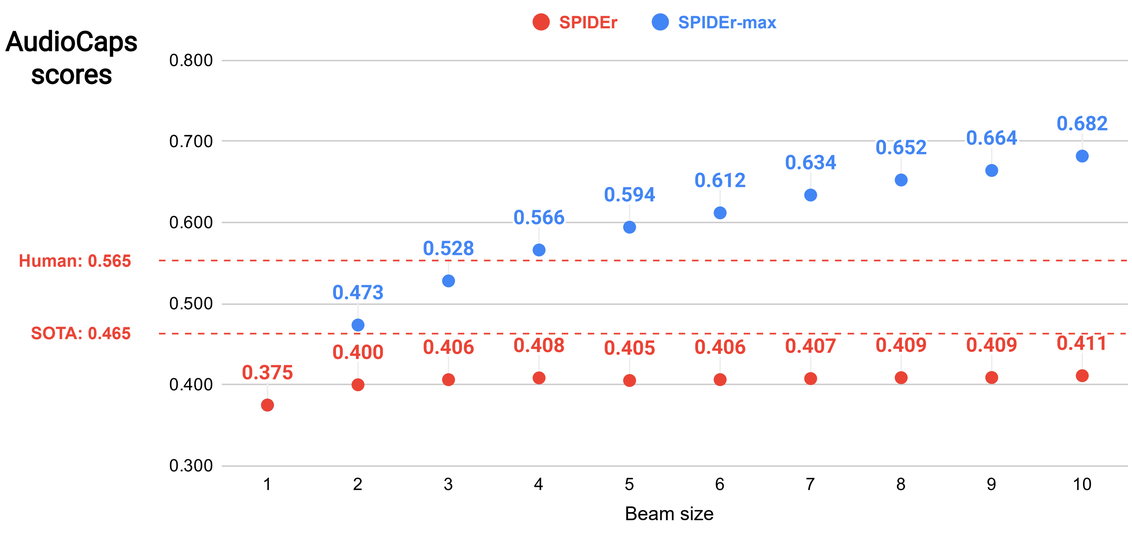
- Beam search caption candidates are similar, but SPIDEr and CIDEr-D scores vary drastically
- Log-probabilities are not strongly correlated with SPIDEr (correlation coefficients of 0.224 on Clotho and 0.259 on AudioCaps)
- Selecting the best candidate automatically is a hard problem because sentences are too similar in most cases

SPIDEr-max takes the maximum SPIDEr score of multiple candidates









Conclusions

- SPIDEr is highly sensitive to the n-grams used and can overestimate the score of similar captions
- SPIDEr-max shows that we can increase SPIDEr score drastically, above SOTA and above Human score on AudioCaps (0.565)
- We need a new metric, more robust to the n-grams used (synonyms, words orders...), closer to the meaning of the sentence

References

[1] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in NAACL-HLT, 2019.

[2] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An Audio Captioning Dataset," arXiv:1910.09387 [cs, eess], Oct. 2019.

[3] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," arXiv:1411.5726 [cs], Jun. 2015.

[4] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.

[5] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762 [cs], Dec. 2017.

[6] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," arXiv:1607.08822 [cs], Jul. 2016.

[7] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient optimization of SPIDEr," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 873–881, Oct. 2017, arXiv: 1612.00370.

DCASE 2022 Workshop | November 3-4 | Nancy, France