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

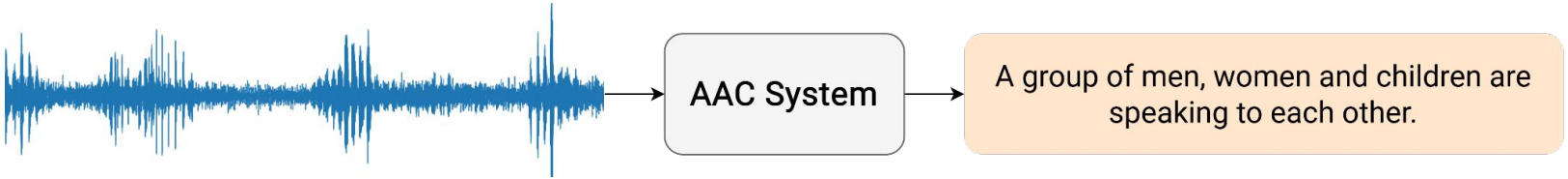
Étienne Labbé, Thomas Pellegrini, Julien Pinquier

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

- PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

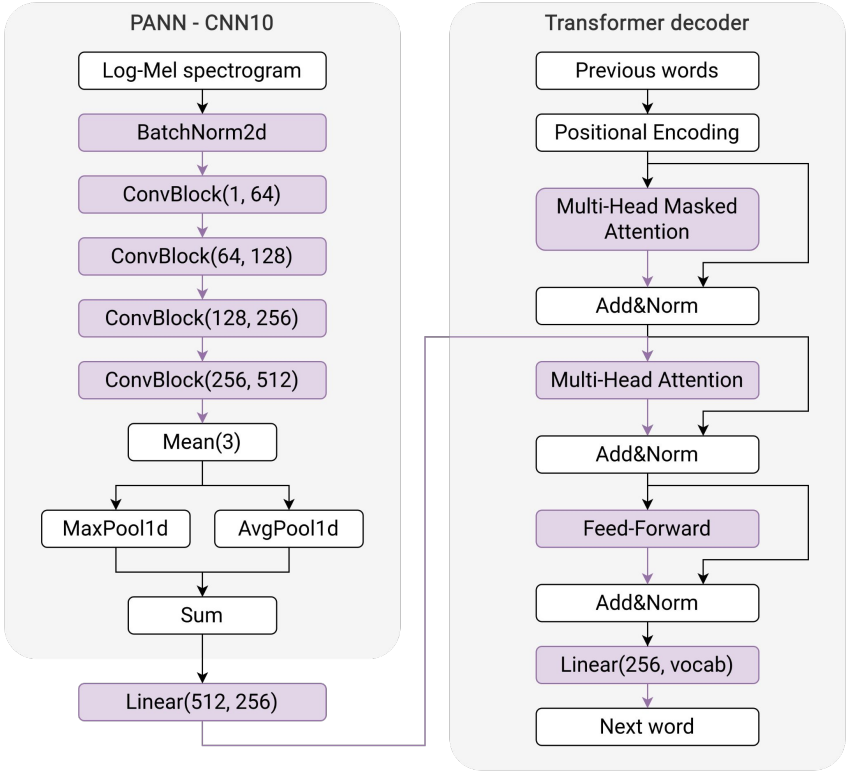
Introduction

- Automated Audio Captioning task (AAC)
- Describe audio events using natural language



System description

- Standard encoder-decoder architecture for AAC
- Pre-trained CNN encoder [1] to recognize events
- Transformer decoder [2] trained from scratch
- Beam search decoding with log-probabilities selection



Captioning Metrics

- Main metrics are inherited from image captioning
- Compare a candidate (prediction) to a list of references (ground truth)
- **CIDEr-D** [3]: Cosine-similarity of TF-IDF scores for 1-grams to 4-grams
- **SPICE** [4]: F-score on a graph of semantic propositions extracted from sentences
- **SPIDEr** [5]: Average of CIDEr-D and SPICE scores

SPIDEr score limitations 1/2

Candidates	SPIDEr	Log-probs
heavy rain is falling on a roof	0.562	-1.018
heavy rain is falling on a tin roof	0.930	-0.898
a heavy rain is falling on a roof	0.594	-0.996
a heavy rain is falling on the ground	0.335	-1.047
a heavy rain is falling on the roof	0.594	-1.079

SPIDEr score limitations 1/2

Candidates	SPIDEr	Log-probs
heavy rain is falling on a roof	0.562	-1.018
heavy rain is falling on a tin roof	**0.930**	**-0.898**
a heavy rain is falling on a roof	0.594	-0.996
a heavy rain is falling on the ground	0.335	-1.047
a heavy rain is falling on the roof	0.594	-1.079

- SPIDEr score varies drastically, even with highly similar captions

SPIDEr score limitations 1/2

Candidates	SPIDEr	Log-probs	References
heavy rain is falling on a roof	0.562	-1.018	heavy rain falls loudly onto a structure with a thin roof
heavy rain is falling on **a tin roof**	**0.930**	**-0.898**	heavy rainfall falling onto a thin structure with a thin roof
a heavy rain is falling on a roof	0.594	-0.996	it is raining hard and the rain hits **a tin roof**
a heavy rain is falling on the ground	0.335	-1.047	rain that is pouring down very hard outside
a heavy rain is falling on the roof	0.594	-1.079	the hard rain is noisy as it hits **a tin roof**

- SPIDEr score varies drastically, even with highly similar captions
- The n-gram “**a tin roof**” gives a much higher SPIDEr score

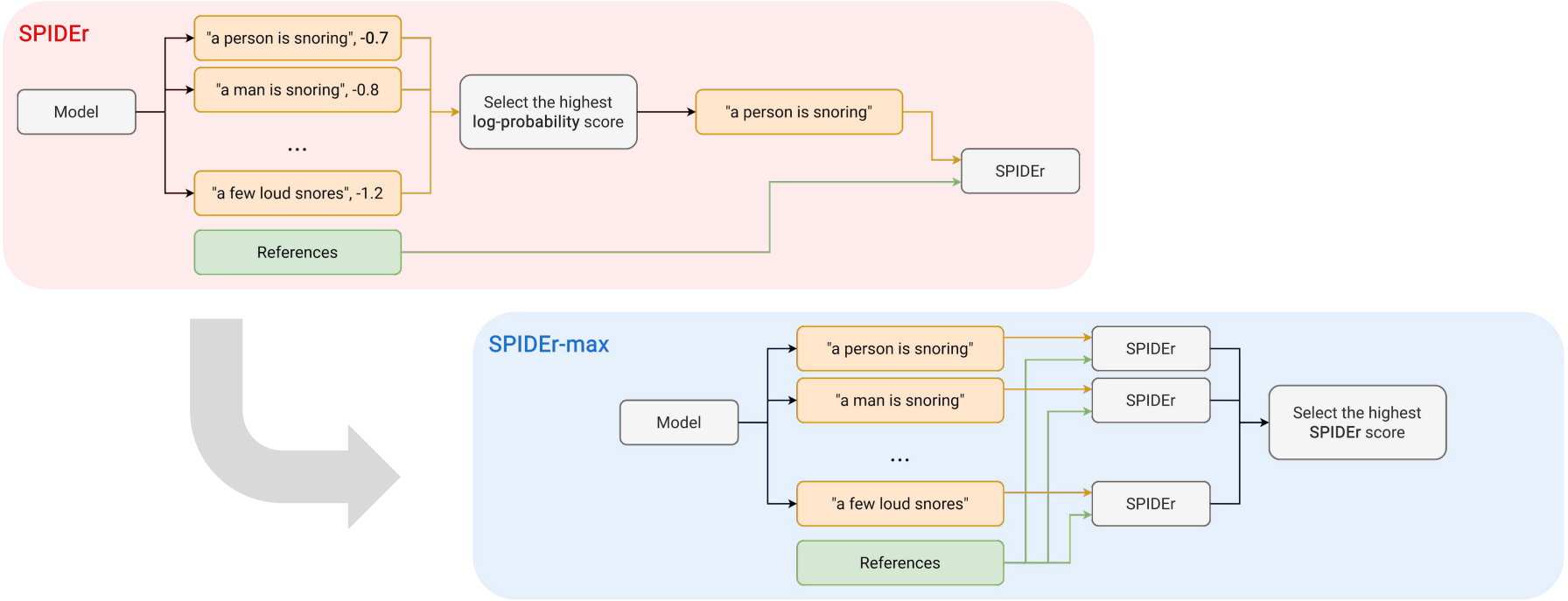
SPIDEr score limitations 2/2

Candidates	SPIDEr	Log-probs	References
a woman speaks and a sheep bleats	0.190	-0.745	a man speaking and laughing followed by a goat bleat
a woman **speaks and a goat bleats**	**1.259**	-0.767	a man is speaking in high tone while a goat is bleating one time
a man speaks and a sheep bleats	0.344	-0.768	a man speaks followed by a goat bleat
an adult male speaks and a sheep bleats	0.231	-0.799	a person **speaks and a goat bleats**
an adult male is speaking and a sheep bleats	0.189	**-0.712**	a man is talking and snickering followed by a goat bleating

- Selection between candidates is a hard problem
- Low correlation between log-probs and SPIDEr (~0.2)

→ Study more these limits: a new metric called **SPIDEr-max**

SPIDEr-max



→ SPIDEr-max shows surprising results, **such as exceeding the "human" SPIDEr score**

Thank you for your attention!

References

[1] 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.

[2] A. Vaswani et al., “Attention Is All You Need,” arXiv:1706.03762 [cs], Dec. 2017.

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

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

[5] 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.