ACES: Evaluating Automated Audio Captioning Models on the Semantics of Sounds

Gijs Wijngaard
Maastricht University

Elia Formisano
Maastricht University

Bruno L. Giordano
CNRS and Université Aix-Marseille

Michel Dumontier
Maastricht University

***Abstract*—Automated Audio Captioning is a multimodal task that aims to convert audio content into natural language. The performance of audio captioning systems is evaluated on quantitative metrics applied to the text representations. Previously, researchers have applied metrics from machine translation and image captioning to evaluate a generated audio caption. Inspired by cognitive neuroscience research on auditory cognition, in this paper we present a novel metric approach that evaluates captions taking into account how human listeners derive semantic information from sounds: Audio Captioning Evaluation on Semantics of Sound (ACES).**

***Index Terms*—automated audio captioning, evaluation metric, semantics**

I. Introduction

Automated audio captioning (AAC) is an emergent field of audio processing. As introduced in 2017 by Drossos et al. [1], the goal of AAC is to describe the content of an audio clip using natural language, i.e. using structured text that contains a description of the sound. The performance of these AAC models is measured by metrics that compare model-predicted captions to corresponding human annotations.

Standard AAC models are based on the encoder-decoder architecture [2]. In this architecture, an audio encoder converts the input audio into an embedding, which gets fed into the decoder. A encoder-decoder learns the structure of the caption by minimizing cross-entropy loss on the probabilities of the decoder outputs. During inference, the decoder calculates the most probable sentence given the embedded audio.

These AAC models are benchmarked using specifically designed metrics. As metrics are based on different criteria (see below), it is also possible to benchmark metrics, for example, by measuring how well they align with human judgement. For this, Zhou et al. [3] introduced the FENSE benchmark, which assesses how the AAC metrics score compares with human evaluation based on four caption categories.

In this paper, we propose a metric to evaluate audio captioning algorithms on annotated datasets. This metric is inspired by cognitive neuroscience research on how human listeners derive and describe semantic information from sounds. It combines measures of overall semantic similarity with specific measures that evaluate the correspondence of generated and reference captions with respect to semantic categories particularly relevant for sound descriptions. We name this metric Audio Captioning Evaluation on Semantics of Sounds (ACES) [1].

II. Related work

A standard toolset of audio captioning evaluation algorithms are the evaluation metrics from the Microsoft COCO Caption Dataset [4]. This set of evaluation metrics was originally intended to measure performance of models on image captioning tasks, but has since been adopted in other domains, including the field of audio captioning. The toolset includes the metrics BLEU [5], ROUGE [6], METEOR [7], CIDEr [8]. Later versions of COCO Caption Evaluation also contain SPICE [9]. The current standard in AAC metrics is SPIDEr, a combination of CIDER and SPICE, which outperforms both metrics based on human evaluation of a randomly sampled COCO test set (10.56% increase compared to a baseline MLE model) [10].

A. Drawbacks of current metrics

There are several drawbacks involved in these metrics. One being that BLEU, ROUGE, CIDEr and METEOR are sensitive to n-gram overlap, which is neither necessary nor sufficient for two sentences to convey the same meaning [9]. For example, the captions “Rain coming from a big cloud”, and “Music coming from a big band”, are semantically dissimilar but have a high n-gram score due to the similarity of the wording. SPICE was created to mitigate this problem but assumes that both the candidate and reference captions have exactly the same wordings in its entities, attributes, and relations. This is not always the case. The candidate sentence “Young woman talking with crunching noise” and the reference sentence “Paper crackling with female speaking lightly in the background.” result in a SPICE score of 0, but one can clearly see its semantic affinity [3]. The captions could also possibly contain no entities (e.g. “Metallic scraping that stops and then starts again”) or no relations (e.g. “Very loud static sound without any other noise”). In this case, no scene graph can be composed or calculated, and the SPICE score returns a low value regardless of semantic relatedness.

Here, we propose a novel metric based on semantic role structure. Semantic-role metrics have been previously investigated by Lo et al. [11], who proposed the MEANT

[1] Code, data and models available at: https://github.com/GlJS/ACES

metric in the context of Machine Translation. A drawback of this metric is the use of syntactic parsing of the sentence, which results in a low score when synonyms are used. Pretrained models such as BLEURT [12] and BERTScore [13] have also been used as a metric for various machine learning tasks. One of the benefits of these models is that the cosine similarity of two words that are semantically similar still results in a high score.

This idea of describing sentences based on semantics was also used in other recent metrics in Audio Captioning. The FENSE metric [3] uses a language model to calculate similarity and a fluency penalty added to capture coherent structures in audio captions. The SPICE+ metric [14] tries to solve issues that arise with its counterpart SPICE, by taking the evaluation into a language model framework. These models are capable of measuring the quality of the generated caption well, but do not take into account the meaning of sentences based on semantic categories.

Our model combines both the strength from semantic entity recognition similar to the MEANT metric, as well as using the latest pretrained models to capture the semantic similarity of captions. We propose using a combination of cosine similarity between the candidate and the reference caption, together with an F1 score from the syntactic parsing of the candidate and reference.

B. Research in other fields

Analogies can be drawn to active research areas in the field of natural language processing (NLP): (1) semantic role labelling (SRL). In this task, predicate-argument structures are modeled from sentences. (2) Part-of-speech tagging (POS), words are tagged based on which part-of-speech they provide, with tags such as subject and adverb. In our work, the dataset is directly labeled by us, but there is a certain correlation between our labels and the way an automated SRL or POS model would label them. For example, the *ARG-0* and *V* labels in a SRL model would correlate to the WHO and HOW property in our model respectively.

Research in the field of psychology and neuroscience has shown that humans listen to sounds to derive information on sources, events and changes in the environment and that this information is reflected in listeners' verbal descriptions of everyday sounds [15]. When asked to describe sounds, listeners refer to the presence of animate (*who*) or inanimate (*what*) sources, identify mechanisms or actions of sound generation (*how*) and, eventually, to a spatial (*where*) or temporal (*when*) context (see Table I).

III. Method

The backbone of the proposed metric is a NLP model that is capable of classifying words from human (or model) generated captions into a set of semantic categories. These categories reflect different dimensions of the semantic information that listeners derive from sounds and have been derived from a recent survey on the semantics of everyday sounds [15] (see Table I). This word model is then applied to both the candidate and the reference captions and a score is calculated that reflects the similarity of their semantic categories.

TABLE I
Sound descriptor categories on the captions of sounds. Items marked with * exist only in the dataset of 13 labels.

Label	Description
WHO	sound-generating agent
WHO/WHAT PROPERTY*	describes object or person
WHAT	vibrating object and substance
HOW	actions or mechanisms
HOW PROPERTY*	specifies action
WHEN	temporal context
WHERE	spatial context
WHAT/WHERE*	surfaces that contribute to acoustics
SOUND TYPE	sounds at signal level
SOUND PROPERTY	attributes of the auditory sensation
NON-AUDITORY SENSATION	non-auditory attributes of sound
OTHER	labels that do not describe sound
O	omitted labels

A. Model training

To obtain this word-classification model, we annotated a random subset of captions from the Clotho dataset and finetuned various pre-trained NLP models (see below). Specifically, we generated two word-labeled caption datasets using the Prodigy web annotation tool [16]. For each caption, we labeled each word using one label from a set of 10 (dataset 1) or of 13 labels (dataset 2), respectively (see Table I).

Dataset 1 consists of 2300 captions labeled by 3 annotators. The annotators labeled 694, 1387 and 219 labels captions respectively. The average inter-rater agreements (Cohen's kappa coefficient) for the first dataset were 0.808, 0.836 and 0.839. In these 2300 captions, some captions were duplicates to calculate inter-agreement statistics. However, to train our models, duplicates were filtered out to ensure that they did not lead to biased results from training items multiple times per epoch or data leakage in the test set. The final dataset consisted of 1158 unique captions. In these 1158 captions, the annotators labeled 285, 727 and 146 respectively.

We applied a similar procedure to dataset 2, which was labeled by 2 annotators, with an average Cohen's kappa coefficient of 0.794. The dataset initially contained 989 captions, where the annotators labeled 494 and 495 labels, respectively. After removing duplicates, the final dataset contained 500 unique captions.

Using the labeled datasets, we finetuned various pre-trained Transformer encoder models with an classification head (see Table II) as implemented in the HuggingFace Transformers library [17]. During tokenization, each caption was split into tokens, where each token corresponds to a label. Tokens from words that did not correspond to a label in our dataset, were assigned the label O. For training, AdamW optimizer [18] was used, with a learning rate of 2e-5 and weight decay. The labeled dataset was

TABLE II
COMPARISON OF BERT-BASED MODELS ON TOKEN CLASSIFICATION OF OUR DATASET. THE HIGHEST SCORES ARE IN BOLD. THE GROUPS OF MODELS ARE 10 LABELS (TOP) AND 13 LABELS (BOTTOM).

Name	F1	F1 How	F1 What	F1 Where	F1 Who
BERT	**0.816**	0.883	0.806	0.785	**0.853**
RoBERTa	0.812	**0.895**	0.793	**0.792**	0.800
XLM RoBERTa	0.797	0.858	0.805	0.780	0.810
ALBERT	0.806	0.870	**0.812**	0.768	0.837
DeBERTa	0.797	0.856	0.801	0.756	0.839
BERT	0.803	0.908	0.806	0.860	0.912
RoBERTa	**0.842**	0.918	0.847	**0.923**	0.909
XLM RoBERTa	**0.842**	**0.937**	0.856	0.839	0.929
ALBERT	0.830	0.914	**0.875**	0.845	**0.955**
DeBERTa	0.809	0.887	0.871	0.847	0.879

divided into 80% train and 20% test sets. The model was trained for 5 epochs, since initial tests showed that after 5 epochs overfitting might occur. For each trained NLP model, we report averaged and category-specific (Who-What-How-Where) F1-scores for the two distinct variants obtained using 10 (top) or 13 (bottom) labels (Table II).

B. Metric definition

This finetuned model is then applied to both the candidate and the reference captions. Both the (word) embeddings from the penultimate layer, as well as the predicted labels after the final linear layer with dropout are taken into account. On the basis of these output labels, (word) embeddings are categorized and combined. For example, for the sentence "a person is walking on a hard surface" the token embeddings of *person*, *walking on* and *hard surface* are categorized into *WHO*, *HOW* and *WHAT* respectively and are compared with the corresponding embeddings of the reference caption. Specifically, for each candidate and reference caption C and R, for each candidate and reference token $c \in C$ and $r \in R$, given the label of $c, r \in \mathbb{C}_l \cap \mathbb{R}_l$, the cosine similarity is calculated as follows:

$$\mathrm{CosSim}(c,r) = \frac{1}{|r|}\sum_{i=1}^{|r|}\frac{1}{|c|}\sum_{j=1}^{|c|}\frac{c_i \cdot r_j}{\|c_i\|\|r_j\|}$$

This cosine similarity predicts the relation of each pair of tokens so that similar tokens in sound descriptors also get a high score, contrary to using the n-gram.

In addition to cosine similarity, a metric based on candidate and reference label overlap helps to penalize predicted captions that do not have the correct reference labels.

$$\mathrm{P}(\mathbb{C},\mathbb{L}) = \frac{\mathbb{C}_l \cap \mathbb{R}_l}{\mathbb{C}_l} \quad \mathrm{R}(\mathbb{C},\mathbb{L}) = \frac{\mathbb{C}_l \cap \mathbb{R}_l}{\mathbb{R}_l} \quad \mathrm{F1} = \frac{2 \cdot P \cdot R}{P + R}$$

The ACES metric is defined as follows:

$$\mathrm{ACES} = \frac{\mathrm{CosSim}(c,r) + \mathrm{F1}(\mathbb{C}_l, \mathbb{C}_r)}{2}$$

ACES returns a single value for each candidate and reference caption C and R. Getting a single value as a result for all candidates and references can be calculated by taking the average for each individual ACES score. This score can be used to get an overall impression of how good the model performs on a set of captions.

TABLE III
SEVERAL MODELS WERE MEASURED AGAINST METRICS, INCLUDING THE ACES METRIC.

	BERTScore	FENSE	SPIDEr	ACES$_{10}$	ACES$_{13}$
Baseline 2021 [2]	0.823	0.251	0.064	0.512	0.516
Baseline 2022 [3]	0.905	0.434	0.231	0.730	0.724
PANNs + BART	0.907	0.454	0.252	0.736	0.725
PASST + BART	0.908	0.463	0.260	0.734	0.723

IV. EXPERIMENTS

Our models were evaluated by comparing them to similar metrics when applied to audio captioning model results. Next to that, our models are tested against the FENSE benchmark to allow to measure its predictions to human evaluation.

A. Model evaluation

We run and evaluated several AAC models using the BERTScore, FENSE, SPIDEr and the proposed ACES metrics. The DCASE baselines of this year and last year were used as benchmarks for the metrics. The 2021 baseline model was an encoder-decoder architecture based on GRU's [19], whereas the 2022 model had a VGGish encoder and a BART-based decoder. In addition to the 2022 model baseline, two other versions were added. In these two versions, the VGGish encoder network was replaced by a PANNs encoder [20] and PaSST [21] encoder, respectively.

In Table III the results of the evaluation of AAC models are shown. It can be noted that BERTScore returns a relatively high value with less discrepancy, whereas SPIDEr metrics start with relatively low values. Interestingly, the ACES scores are not perfectly aligned with the other scores: PANNs + BART scores the best of all models on the ACES metric, whereas PASST + BART scores the best on all other metrics. Also, ACES$_{13}$ is even less aligned than ACES$_{10}$ with an average Kendall's τ of $\frac{1}{3}$ in ACES$_{13}$ to $\frac{2}{3}$ in ACES$_{10}$. The differences between scores for ACES$_{13}$ are smaller in comparison with ACES$_{10}$.

B. Human evaluation

We utilised the FENSE benchmark [3] to evaluate our score against human evaluation. The FENSE benchmark consists of four components, each to calculate the quality of a candidate metric against its reference: (1) human-correct (HC), where both captions are correct references. (2) human-incorrect (HI), where one caption is a reference from a different sentence. (3) human-machine (HM) and (4) machine-machine (MM) where one and both captions

[2] https://github.com/audio-captioning/dcase-2021-baseline
[3] https://github.com/felixgontier/dcase-2022-baseline

TABLE IV
PERFORMANCE OF VARIOUS METRICS ON AUDIOCAPS AND CLOTHO, HIGHEST VALUES IN BOLD.

Metrics	AudioCaps					Clotho-Eval				
	HC	HI	HM	MM	Total	HC	HI	HM	MM	Total
SPIDEr	53.2	89.9	84.1	55.2	65.4	47.9	88.1	67.9	52.5	59.8
BERTScore	60.6	**97.6**	**92.9**	65	74.3	57.1	**95.5**	70.3	61.3	67.5
BLEURT	**77.3**	93.9	88.7	72.4	79.3	59	93.9	75.4	67.4	71.6
Sentence-BERT	64	99.2	92.5	73.6	79.6	60	**95.5**	75.9	66.9	71.8
FENSE	64.5	98.4	91.6	**84.6**	**85.3**	60.5	94.7	**80.2**	**72.8**	**75.7**
SPICE+	59.1	85.4	83.7	49	62	46.7	88.1	70.3	48.7	57.8
SPICE+emb	63.5	96.4	91.6	70	77	**61**	94.7	76.3	61.6	68.9
ACES_{10}	52.2	74.5	65.7	55.7	59.9	55.7	84	53.4	57	60.5
ACES_{13}	55.7	77.7	69	53.9	60.6	56.2	83.6	58.6	58.2	62

are from a captioning model respectively. Our metrics are compared to several other metrics in the FENSE benchmark (see Table IV).

Overall, the metric does not perform on par with other metrics in the FENSE benchmark, which could indicate that the metric is not aligned with human evaluation. Future work and more evaluation will indicate how to improve the metric on this regard in performance on this benchmark. was

V. CONCLUSION

In this paper, Audio Captioning Evaluation on Semantics (ACES) is introduced, a metric based on audio semantic research. This paper showed the possibility of an AAC evaluation metric that combined both semantic similarities and semantic entity labeling.

Although the model trained with 13 labels (ACES_{13}) was better compared to ACES_{10} on the FENSE benchmark for human evaluation, it did not align well with other metrics when used for the evaluation of AAC models (Table III). Further work will need to highlight why model evaluation is not consistent with human evaluation.

Some improvements could be made regarding training our language model. Currently, a 1120 and 500 captions dataset was utilised that resulted in a 84.2 and 81.2% F1 score for both our models on correct classification. More labelled data would result in better classification. There was also a limitation of only using one dataset, Clotho. Adding AudioCaps or other datasets could diversify the training data.

REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated Audio Captioning with Recurrent Neural Networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, no. arXiv:1706.10006, Oct. 2017.

[2] 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, Oct. 2022.

[3] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can Audio Captions Be Evaluated with Image Caption Metrics?" in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. arXiv, Jan. 2022.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO Captions: Data Collection and Evaluation Server," Apr. 2015.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.

[6] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[7] A. Agarwal and A. Lavie, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," *Proceedings of WMT-08*, 2007.

[8] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-Based Image Description Evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[9] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *ECCV 2016*. Springer International Publishing, Jul. 2016.

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

[11] C.-k. Lo, A. K. Tumuluru, and D. Wu, "Fully Automatic Semantic MT Evaluation," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, Jun. 2012, pp. 243–252.

[12] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning

Robust Metrics for Text Generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations*, Feb. 2020.

[14] F. Gontier, R. Serizel, and C. Cerisara, "SPICE+: Evaluation of Automatic Audio Captioning Systems With Pre-Trained Language Models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2022.

[15] B. L. Giordano, R. de Miranda Azevedo, Y. Plasencia-Calaña, E. Formisano, and M. Dumontier, "What do we mean with sound semantics, exactly? A survey of taxonomies and ontologies of everyday sounds," *Frontiers in Psychology*, vol. 13, 2022.

[16] "Prodigy · Prodigy · An annotation tool for AI, Machine Learning & NLP."

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-Art Natural Language Processing," Association for Computational Linguistics, pp. 38–45, Oct. 2020.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *ICLR 2019*, Jan. 2019.

[19] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Sep. 2014.

[20] 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, Aug. 2020.

[21] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Interspeech 2022*, Mar. 2022.