A SUMMARIZATION APPROACH TO EVALUATING AUDIO CAPTIONING

Irene Martín-Morató, Manu Harju, Annamaria Mesaros

Computing Sciences Tampere University, Finland
{irene.martinmorato, manu.harju, annamaria.mesaros}@tuni.fi

ABSTRACT

Audio captioning is currently evaluated with metrics originating from machine translation and image captioning, but their suitability for audio has recently been questioned. This work proposes content-based scoring of audio captions, an approach that considers the specific sound events content of the captions. Inspired from text summarization, the proposed measure gives relevance scores to the sound events present in the reference, and scores candidates based on the relevance of the retrieved sounds. In this work we use a simple, consensus-based definition of relevance, but different weighing schemes can be easily incorporated to change the importance of terms accordingly. Our experiments use two datasets and three different audio captioning systems and show that the proposed measure behaves consistently with the data: captions that correctly capture the most relevant sounds obtain a score of 1, while the ones containing less relevant sounds score lower. While the proposed content-based score is not concerned with the fluency or semantic content of the captions, it can be incorporated into a compound metric, similar to SPIDEr being a linear combination of a semantic and a syntactic fluency score.

Index Terms— audio captioning, evaluation, content-based retrieval

1. INTRODUCTION

Automated captioning, the description of images, audio, or video content using unrestrained natural language, is an active research topic in all these fields. The first works in image captioning defined it as a machine translation task, and evaluated it using metrics from machine translation such as BLEU [1], METEOR [2] and ROUGE [3], which are primarily based on n-gram overlap between the reference and candidate caption. Subsequently, it was observed that these metrics do not correlate well with human opinion [4], resulting in development of new metrics optimized for image captioning such as SPICE [5] and CIDEr [6]. SPICE measures performance using a graph-based semantic representation that explicitly encodes the objects, attributes and relationships found in image captions, while CIDEr measures how well a candidate sentence matches the consensus of a set of image descriptions, using n-grams weighted using Term Frequency Inverse Document Frequency (TF-IDF) weighting, combining n-grams of varying lengths (typically up to 4-grams). Further optimization in image captioning resulted in SPIDEr [7], a linear combination of SPICE and CIDEr.

Audio captioning was defined and evaluated the same way as image captioning [8]. In the last few years, the DCASE Challenge has accelerated development of audio captioning methods, seen clearly in the significant improvement of the evaluation results. Approaches are based on encoder-decoder systems, with the decoder usually a recurrent network with sequence-to-sequence modeling. Recently, the use of transformers has become very popular [9], with pretrained models such as BERT [10] and BART [11] consistently ranked as state of the art. The evaluation scheme has not changed, and consists of the same set of metrics, from BLEU and ROUGE to CIDEr, SPICE and SPIDEr.

The recent development of large language processing models, notably BERT [10], brought a new approach to measuring similarity of text, initiating research into metrics more suitable for measuring audio captioning outputs. For example the work in [12] proposed FENSE, a new metric which combines Sentence-BERT for semantic similarity [13] with an error detector to penalize erroneously formed sentences. FENSE was shown to correlate with human judgments, in experiments that evaluated the output of four different captioning systems on the Clotho dataset [14], and showed that FENSE ranked the best systems the same way as humans did.

We introduce a novel perspective to audio captioning, namely summarization. Instead of cross-modal machine translation, we regard audio captioning as *cross-modal summarization*. A detailed description of a complex acoustic scene using natural language would include information on all the different sounds present at the scene. However, humans in fact do not care or pay attention to everything, and may consider some content irrelevant. We assume that annotators required to describe audio content would include the most relevant content, subject to their own judgement. In this respect, the textual description can be viewed as a summary.

In this paper we propose a novel measurement for audio captioning, which considers the captions content in terms of sound events. The different sounds mentioned in the captions are given relevance scores based on the annotators' consensus in producing the reference captions. A candidate caption is then evaluated based on the relevance of its content with respect to the reference information, given that the optimal caption for an audio will contain the topmost relevant sounds. In effect, this is a content-based scoring scheme that rewards the captions which retrieve the most relevant sounds in the captioned audio. The contributions of this work are the following: (i) we formulate audio captioning as audio-to-text summarization; (ii) we propose a method to estimate relevance of sound events based on multiple reference captions; (iii) we propose a relevance-based score for evaluation of audio captions content.

The paper is organized as follows: Section 2 introduces the proposed sound relevance estimation scheme and the proposed metric for evaluating candidate captions. Section 3 presents evaluation results using different captioning systems and datasets and introduces ideas for possible further development. Finally, Section 4 presents the conclusions.

This paper has received funding from Jane and Aatos Erkko Foundation, Finland.

2. CONTENT-BASED EVALUATION OF CAPTIONS

This work is inspired by the Pyramid method for evaluating content selection of textual summaries [15]. It was observed that among textual summaries produced by humans, many seem equally good without having identical content. This is valid for human-produced audio captions too, in particular for complex scenes. The Pyramid method starts by annotating "summarization content units" (SCU), then uses these SCUs to produce an optimal summary. The Pyramid score is defined as the ratio of a candidate summary to the ideal summary having the same number of SCUs.

Considering the caption as a summary, we define our content units as sound events. In consequence, we will evaluate a caption based on the sound events that are mentioned in this caption. The sound events are extracted from the captions using the AudioSet ontology [16], on the grounds that while it may be an incomplete list of possible sounds, the ontology provides a large set of the most common sounds. The human-produced reference captions are processed to extract the sound events from each caption, which are then assigned a relevance as explained in the following.

2.1. Estimating relevance of sound events

In its simplest way, the relevance of a sound event to an audio clip can be defined based on how many annotators have referred to it in the caption they provided. Consider the example from Table 1: given ten captions for one audio clip, four of them mention birds singing, two mention car passing by, all ten mention children laughing, and nine mention children talking.

We define relevance of sound event i as its consensus-based weight, calculated as:

$$rel_i = \frac{N_i}{\sum_{j=1}^{M}(N_j)} \tag{1}$$

where N_i is the number of times sound event i is mentioned in the captions assigned to a clip, with M being the total number of sound events mentioned in all captions. The effect of this definition is that a relevance of 1 is distributed among all the mentioned sound event classes M based on how frequently they appear in the captions:

$$\sum_{i=1}^{M} rel_i = 1. \tag{2}$$

For the example in Table 1, bird singing has a relevance of 0.16, while children laughing has a relevance of 0.40.

One may argue that in certain cases the rare sound events may be more relevant to a clip, instead of the most commonly mentioned ones. Depending on the application and the desired output, the relevance of individual sound events to a clip can be estimated based on direct frequency as in the example above, or using TF-IDF weighing to give more weight to sounds that are very specific to a clip. For simplicity, we only use the former approach in this work, and leave other weighing schemes for future development.

2.2. Evaluation of candidate captions

The content-based score (CB-score) of a candidate caption C containing K sound events is defined as the ratio between the relevance of its content and the relevance of the optimal caption C_K containing the same number K of sound events:

Label	Freq	Relevance
Bird singing	4	0.16
Car passing by	2	0.08
Children laughing	10	0.40
Children talking	9	0.36

Table 1: Consensus-based relevance of sound events to a clip

$$\text{CB-score} = \frac{\sum_{j=1}^{K} rel_j}{\sum_{k=1}^{K} rel_k} \tag{3}$$

where sound events j belong to the candidate caption C, and events k belong to the optimal caption C_K. The optimal caption C_K is defined as containing the most relevant K of the M sound events mentioned in the reference captions. This means that for the example in Table 1 the optimal caption containing only one sound event would contain "children laughing", while the optimal caption containing two sound events would contain "children laughing" and "children talking". Table 2 gives examples of CB-score calculation for different captions, using the relevance scores from Table 1.

3. EXPERIMENTS

For each audio clip, the reference captions are processed to extract sound events and estimate their consensus-based relevance. Then, the candidate caption is evaluated against the optimal caption as explained in the previous section. Experiments were performed using three different systems and multiple datasets, in order to verify the behavior of the metric. For comparison, SPIDEr scores (as used in DCASE Challenge 2022) and FENSE scores were calculated.

3.1. Captioning datasets and systems

The datasets used for evaluating the behaviour of the proposed metric are Clotho [14] which has five captions per clip, for audio clips 15 to 30 seconds long that were collected from Freesound [17] and AudioCaps [18], consisting of 10-second long audio clips from AudioSet [16], for which only the test split has five captions per clip, the rest has only one caption.

We use three different systems to generate automatic captions. The first system is the DCASE task6 subtask A baseline system[1], which is a sequence-to-sequence transformer based on BART model[2]. The second system (ED-RNN) consists of an Encoder-Decoder architecture with an attention layer in the decoder and bi-directional RNN in the encoder. The third model (AACTrans) is a sequence-to-sequence transformer similar to the baseline system, but having smaller number of parameters and using greedy token generation. All three models use VGGish [19] features as inputs. For brevity, we do not present more details about the systems, since they are irrelevant for the objective of this study. The systems are separately trained and tested using Clotho and AudioCaps, respectively, using the training/test splits provided with the datasets.

[1] https://github.com/felixgontier/dcase-2022-baseline

[2] https://huggingface.co/docs/transformers/model_doc/bart

Candidate captions:
C1. Children are talking outside.
C2. A dog is barking at a car passing by.
C3. A car is passing by a group of children that are playing and laughing.

	Sound events	Relevance	Ideal caption content	CB-score
C1	Children talking	0.36	Children laughing	0.36/0.40 =0.90
C2	Dog barking, Car passing by	0.08	Children laughing, Children talking	0.08/(0.40+0.36)=0.11
C3	Car passing by, Children laughing	0.08+0.40	Children laughing, Children talking	0.48/0.76=0.63

Table 2: Example of CB-scores for candidate captions, based on relevance scores from Table 1.

	CLOTHO			AudioCaps		
System	SPIDEr	FENSE	CB-score	SPIDEr	FENSE	CB-score
Baseline	0.22 (0.21, 0.24)	0.46 (0.45, 0.47)	0.49 (0.46, 0.51)	0.34 (0.32, 0.37)	0.57 (0.56, 0.58)	0.63 (0.60, 0.65)
ED-RNN	0.15 (0.14, 0.16)	0.41 (0.40, 0.43)	0.40 (0.38, 0.43)	0.30 (0.28, 0.33)	0.54 (0.53, 0.55)	0.62 (0.59, 0.64)
AACTransformer	0.19 (0.18, 0.21)	0.40 (0.39, 0.41)	0.47 (0.45, 0.50)	0.35 (0.32, 0.37)	0.54 (0.53, 0.55)	0.65 (0.63, 0.68)
Reference caption	0.58 (0.55, 0.61)	0.58 (0.57, 0.59)	0.64 (0.62, 0.66)	0.56 (0.52, 0.59)	0.68 (0.68, 0.69)	0.76 (0.75, 0.78)

Table 3: Performance of different systems on Clotho and AudioCaps datasets.

3.2. Practical implementation details

The extraction of sound events from the textual captions is not possible using simple word matching, due to the unconstrained nature of the annotation process. We therefore match individual terms and their synonyms in a controlled manner, to extract what we call sound event tokens as well as possible. The processing steps are the following: First, AudioSet vocabulary and the reference captions are tokenized and lemmatized to obtain the root form for each word. Then, for each token in the caption, its first order synonym and first order hypernym are extracted from WordNet [20], based on its POS (part of speech) obtained using spaCy [21] and nltk [22]. This allows identifying words that are related to sounds, when they do not match the exact vocabulary of AudioSet. AudioSet vocabulary is used as a two-level hierarchy to standardize the depth of the vocabulary. If the child node in AudioSet does not match the extracted sound event from the caption but its synonym does, the extracted token is matched to the parent node. For example "people *talking*" is processed to "group, people, citizenry, *speak*, talk, communicate", where *speak* is a child of *speech* in AudioSet, therefore *talking* is matched to *speech*.

We expect that this matching process will result in some amount of errors, in some cases matching wrong terms. However, we consider that the vocabulary used in the captions and in general for describing sounds are a small subset of WordNet, and words are mostly used with their most common meaning, which means that wrongly matched synonyms should not affect the scores very much. There are also cases where a correspondence to the AudioSet vocabulary is not found, and therefore the process fails to identify sounds.

3.3. Numerical results and analysis

The evaluation results are presented in Table 3 for the three systems, comparing SPIDEr, FENSE and CB-score. Confidence intervals for the metrics were calculated using the jackknife resampling procedure on the system output. To understand the dynamics of the metric values and to have better insight on the expected upper bound for each metric, we selected randomly one of the five reference captions and compared it to the other four using the three metrics. The results of this evaluation are presented in the last row of Table 3.

Among the captioning systems, we observe that the baseline has the best performance on Clotho, a result which is consistent for all three metrics. On AudioCaps, the baseline and AACTransformer have similar performance, with confidence intervals of the metrics overlapping. In comparison, the evaluation of one reference caption against the others produces similar SPIDEr score on both datasets, while FENSE and CB-score are significantly higher on AudioCaps. It may be noteworthy that the datasets are of different size, with 46k and 29k captions available for training in AudioCaps and Clotho, respectively. Since usually larger datasets lead to more robust models, the size difference may explain the AudioCaps better scores.

To understand the meaning of these average values, the distribution of the scores calculated for the baseline system are illustrated in Fig. 1. We observe that SPIDEr is very concentrated close to 0, while FENSE looks normally distributed between 0 and 1. CB-score has many values at the extremes, with 0 corresponding to the case when no sounds were matched between the evaluated caption and the reference ones, and 1 for the case when the evaluated caption contained the most relevant sound(s). For Clotho 33% of the captions produced for the test split have CB-score 0, while 27% have a score of 1. On the other hand the FENSE score is more moderate, with most captions scoring between 0.4–0.6, only 13 files having a FENSE score of 0. SPIDEr scores are under 0.25 for 66% of the captions. AudioCaps data obtains more balanced scores, with only around 13% of the captions in the test split getting a CB-score of 0, while the maximal score of 1 is obtained by around 30% of the files, similar to Clotho.

For comparison, Fig. 2 shows the CB-score and FENSE distributions for evaluating one reference caption (selected randomly) against the others. Interestingly, 17% of the Clotho test split gets CB-score 0, which means that the captions produced by different people do not contain the same sound events. This information is not indicated by FENSE, which measures the general semantic similarity of the sentences. For both Clotho and AudioCaps, a CB-score of 1 is obtained by around 40% of the human-produced captions.

To analyze the content evaluated by the CB-score, we investigate the number of sound tokens extracted from the caption. The statistics are shown in Fig. 3; the prediction uses the output of the

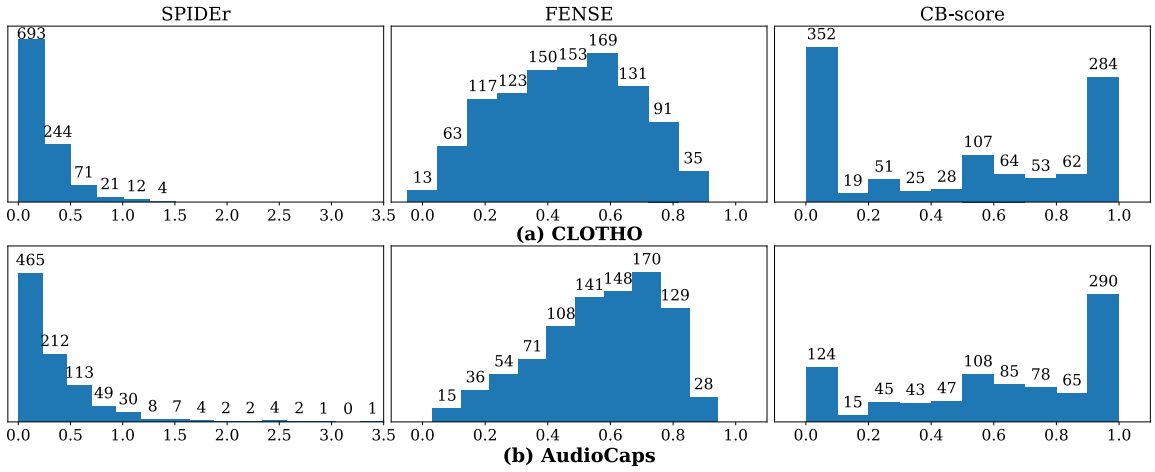

Figure 1: Distribution of SPIDEr, FENSE and CB-score on the test splits of Clotho and AudioCaps for the baseline system.

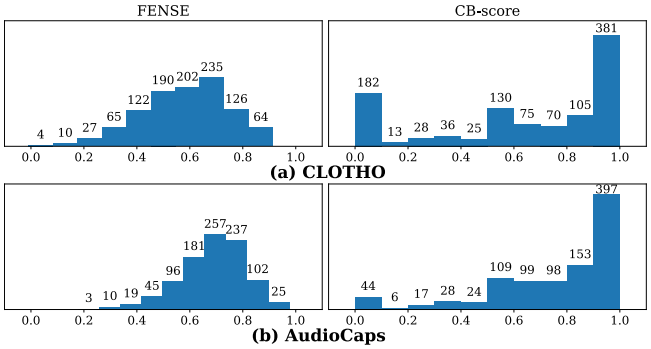

Figure 2: Distribution of FENSE and CB-score for evaluating one reference caption against the others for each clip in the test set.

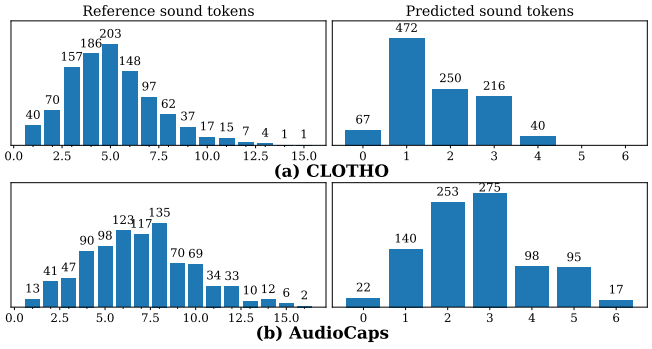

Figure 3: Statistics of the extracted tokens that refer to sound events in the reference captions and in the automatically generated ones.

baseline system. For both Clotho and AudioCaps, our token extraction process results in a large number of terms per clip. On the other hand, 45% of the captions produced for Clotho result in only one token, while 6% do not match any sounds. For AudioCaps, the same system produces captions which are richer in sound tokens, and our token extraction method results in 2 or 3 tokens for 54% of them.

3.4. Discussion and further development

Based on its formulation in Section 2, it can be noticed that this is a precision-type of metric, reflecting how many of the sound events that appear in a caption are as highly weighted as possible. This scoring penalizes the presence of sounds if there are others more highly weighted but not included in the caption. A recall-oriented score can be formulated by defining the optimal caption as containing the *average number* of sound events in the reference annotations (instead of the same number K as the candidate). It can also be observed that the formulation in eq.3 is not sensitive to extra information, which in effect means that inserted sounds are not penalized, even if they were not present in the reference captions at all.

The idea of scoring the caption using consensus is not new: CIDEr also uses consensus, but is based on n-grams, while our proposal is based on sound events. We find the use of sound events as units of information more relevant to the audio captioning than the use of n-grams, even though this ignores the lexical structure. Because the definition of the sound relevance is independent from the CB-score itself, therefore the CB-score is flexible in allowing sound relevance score formulations depending on the target application. Additionally, given the relatively short sentences available as captions, we hypothesize that a higher number of reference captions would provide more reliable relevance estimates.

4. CONCLUSIONS

We introduced CB-score, a metric to evaluate how well the sound events mentioned in automatically produced captions correspond to the sounds mentioned in the reference captions. We also introduced a simple method to estimate relevance of reference sounds based on multiple captions, which can be extended depending on the target application to give more importance to the most commonly mentioned sounds or to rare sounds that are highly specific to a clip. The proposed metric lacks the ability to measure lexical or grammar structure of the caption, therefore it is not a sufficient metric for evaluating audio captions, but it successfully summarizes the content of the acoustic description, which means that it can be used as such for evaluating captioning tasks focused on the correctness rather than syntactic richness of the produced sentences. Similar to SPIDEr being a linear combination of a semantic and a syntactic fluency score, the CB-score can further evolve into more complex compound measurements, with additional components measuring for example syntactic and semantic aspects of the produced caption.

5. REFERENCES

[1] 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 on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318.

[2] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 228–231.

[3] 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.

[4] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Int. Res.*, vol. 47, no. 1, p. 853–899, May 2013.

[5] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 382–398.

[6] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation." in *CVPR*. IEEE Computer Society, 2015, pp. 4566–4575.

[7] 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)*, 2017, pp. 873–881.

[8] 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)*, 2017, pp. 374–378.

[9] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 211–215.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the NAACL HLT, Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 7871–7880.

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

[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," in *EMNLP/IJCNLP*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 3980–3990.

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

[15] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2004, pp. 145–152.

[16] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

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

[18] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the NAACL HLT, Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132.

[19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135. [Online]. Available: https://arxiv.org/abs/1609.09430

[20] C. Fellbaum, *WordNet: An Electronic Lexical Database*. The MIT Press, 05 1998. [Online]. Available: https://doi.org/10.7551/mitpress/7287.001.0001

[21] I. Montani, M. Honnibal, M. Honnibal, S. V. Landeghem, A. Boyd, H. Peters, P. O. McCann, M. Samsonov, J. Geovedi, J. O'Regan, D. Altinok, G. Orosz, S. L. Kristiansen, L. Miranda, D. de Kok, Roman, E. Bot, L. Fiedler, G. Howard, Edward, W. Phatthiyaphaibun, Y. Tamura, S. Bozek, murat, R. Daniels, M. Amery, B. Böing, B. Vanroy, and P. K. Tippa, "explosion/spaCy: v3.3.0: Improved speed, new trainable lemmatizer, and pipelines for Finnish, Korean and Swedish," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6504092

[22] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.