

Jarkko Piilola

CAPTION GUIDED SOUND EVENT TAGGING

Bachelor of Science thesis
Faculty of Information Technology and Communications Sciences
Parthasaarathy Ariyakulam Sudarsanam
June 2025

ABSTRACT

Jarkko Piilola: Caption guided sound event tagging
Bachelor of Science thesis
Tampere University
Information technology, signal processing and machine learning
June 2025

Sound event tagging (SET) has seen progress through deep learning models, achieving high levels of accuracies. However, these models can face limitations, especially when there is ambiguity in classification. Therefore, it can be beneficial to assist these SET models utilizing multimodality for more contextual information. This thesis presents a system for utilizing audio clip related captions and natural language processing (NLP) to enhance confidence scores produced by a baseline SET model. Both the SET and caption tags are pre-processed, extracted and embedded into a shared vector space using Sentence-BERT (SBERT) model. Cosine similarity is then calculated between tags to assess their contextual correspondence. The adjusted SET confidences are derived from both positive and negative boosting using a mathematical formula that considers caption tag matches related to SET tags and number of captions into account with variable caption weighting. The goal is to produce more accurate and semantically coherent classification confidences. Experimental results show gradual decreasing trend in F1-scores with increasing caption tag weight. However, there are cases where improvements in F1-scores happen given low weight to caption tags, with moderate acceptance and semantic similarity thresholds. The highest F1-score achieved 0.696 with parameters $\alpha = 0.1$, $\tau_{conf} = 0.3$, $\tau_{sim} = 0.5$. These findings suggest that multimodality through caption tagging can complement but not replace sound event tagging, especially when boosting is based on SET model's confidences.

Keywords: sound event tagging (SET), deep learning, multimodality

The originality of this thesis has been verified using the Turnitin Originality Check service.

TIIVISTELMÄ

Jarkko Piilola: Caption guided sound event tagging
Kandidaatintyö
Tampereen yliopisto
Tietotekniikka, signaalinkäsittely ja koneoppiminen
Kesäkuu 2025

Äänitapahtumien tunnistaminen (SET) on kehittynyt merkittävästi syväoppimismallien myötä, saavuttaen korkeita tarkkuuksia. Näillä malleilla on kuitenkin rajoitteita, erityisesti tapauksissa, joissa äänitapahtuman luokittelu on epäselvää. Tämän vuoksi multimodaalisen lähestymistavan hyödyntäminen voi tuoda lisätukea tunnistuksille ja niihin liittyville varmuuksille. Tässä kandidaatintyössä esitellään järjestelmä, joka hyödyntää ääniklippeihin liittyviä kuvauksia ja luonnollisen kielen käsittelyä (NLP) SET-mallin tuottamien tunnistusvarmuuksien parantamiseksi. Sekä SET-että kuvauksien tunnisteet esikäsitellään, uutetaan ja upotetaan vektoriavaruuteen Sentence-BERT (SBERT) -mallin avulla. Kosinietäisyyttä käytetään tunnisteiden kontekstuaalisen vastaavuuden arviointiin. Järjestelmä säätää SET-mallin tunnistusvarmuuksia positiivisen ja negatiivisen vahvistuksen avulla. Säätö perustuu matemaattiseen kaavaan, joka ottaa huomioon kuvauksista johdetut tunnisteet niiden määrät, kuvauksien määrät ja säädettävän painokertoimen. Tavoitteena on tuottaa tarkempia ja semanttisesti johdonmukaisia luokittelutuloksia. Kokeelliset tulokset osoittavat, että F1-arvot laskevat yleisesti kuvauksien painon kasvaessa. On kuitenkin tapauksia, joissa havaitaan parannuksia. Nämä esiintyvät erityisesti, kun kuvauksille annettu paino on matala ja hyväksymisen sekä semanttisen samankaltaisuuden kynnyksarvot ovat kohtuulliset. Paras saavutettu F1-arvo on 0.696 parametreillä $\alpha = 0.1$, $\tau_{conf} = 0.3$, $\tau_{sim} = 0.5$. Tulokset viittaavat siihen, että kuvauksiin perustuva multimodaalisen lähestymistavan integraatio voi täydentää, mutta ei korvata äänitapahtumien tunnistamista, erityisesti kun säätö perustuu SET-mallin alkuperäisiin varmuuksiin.

Avainsanat: äänitapahtumien tunnistus, syväoppiminen, multimodaalisuus

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin Originality Check -ohjelmalla.

USE OF AI IN THESIS

I have utilised AI tools in my thesis:

No

Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools: OpenAI GPT-4o mini

Purpose of using AI tools: Code debugging for the experiment. Reference formatting for the thesis.

Sections where AI tools were used: Experiment code. Thesis section 6. References.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

CONTENTS

USE OF AI IN THESIS	III
1.INTRODUCTION	1
1.1 Background and significance	1
1.2 Objectives and scope of the study	1
1.3 Related works	2
2.THEORETICAL BACKGROUND.....	3
2.1 Feature representation	3
2.2 Machine learning and deep learning in audio processing.....	4
2.3 Sound event tagging with CNNs	5
2.4 Natural Language Processing (NLP)	6
2.5 Contextual information in audio tagging	7
3.RESEARCH METHODS AND DATA	9
3.1 Description of datasets: audio datasets and their captions.....	9
3.2 Proposed framework.....	10
3.2.1 Preprocessing	10
3.2.2 PANNs tags	10
3.2.3 Caption tags.....	12
3.2.4 Tag comparison	12
3.2.5 Confidence boosting	13
4.RESULTS AND ANALYSIS.....	15
4.1 Parameter and threshold analysis.....	15
4.2 Class-specific results	18
5.SUMMARY AND CONCLUSIONS	21
5.1 Key findings	21
5.2 Limitations	22
5.3 Future research	22
6.REFERENCES	24

1. INTRODUCTION

1.1 Background and significance

Sound event tagging refers to identifying and labelling specific events or sound categories within a clip, predicting the presence without determining temporal boundaries (Kong et al., 2019). Sound event tagging has a wide range of applications including sound monitoring, assistive technologies, media indexing and much more.

Deep learning is a part of machine learning which attempts to follow the structure of human brains using neural networks (Goodfellow et al., 2016). Modern deep learning models have advanced in audio classification tasks, with high accuracies (Kong et al., 2020).

The focus of this thesis is in improving sound event tagging with a pretrained audio neural network model as a baseline and boosting tagging confidences based on their contextual importance through natural language processing and clip related captions. The experiment uses modern audio tagging and NLP techniques and models to combine a multi-modal approach for sound event tagging.

1.2 Objectives and scope of the study

A large amount of audio data is unlabelled, and their use cases can be limited (Wang et al., 2018). While pre-trained sound event tagging models can be used to extract tags, these models may be limited in accuracy as there may be a domain mismatch on audio data used compared to what models have been trained on (Yao et al., 2020). Sound event tagging models may not be able to identify non-prominent classes and rely on a fixed pre-defined label set, which might not cover the entire acoustic content (Hamaguchi, Sakurada, & Nakamura, 2018).

Human-annotated captions commonly refer to the labelling of raw data with relevant information, useful for improving machine learning model capabilities (Tan et al., 2024). The assumption is that human-annotated captions provide meaningful semantic information which decreases ambiguity in audio tagging and allows for more accurate and semantically coherent results. The use of textual information through captions provides multimodality and more context describing the acoustic scene which helps improve

confidence adjustment for sound event tagging. This raises the question: “How can sound event tagging be improved by utilizing contextual information from captions?”.

1.3 Related works

Applications such as automatic tagging for audio content have improved due to advancements in machine learning and deep learning. This part focuses on modern techniques using convolutional neural networks, pre-trained audio neural networks, natural language processing (NLP) for audio analysis as well as multimodal approaches for utilizing contextual information in addition to the audio signal. In addition to CNNs, transformer-based models have made advancements in the audio processing field.

Gong, Chung, & Glass (2021) introduce Audio Spectrogram Transformer (AST), utilizing transformer-based architectures able to capture long-range global context in sequential data. These advancements have proved to perform better than convolutional neural networks, indicating its ability to learn audio representations through context.

Related to AST, Koutini et al. (2022) proposed Patchout Audio Spectrogram Transformer (PaSST), which involves dropping patches of the input spectrogram randomly during training to improve training efficiency and generalization. The use of PaSST reduces computational requirements, suitable for large-scale audio processing tasks.

Dinkel et al. (Interspeech 2022) introduce UniKW-AT, an audio classifier combining keyword spotting and audio tagging into one model. The approach combines general audio tagging model MobileNetV2 and additional output labels for keywords into a single convolutional model. The model can handle predicting wake-word presence and sound-event classification simultaneously through multi-label loss training. The model performs well with both tasks showing that keyword spotting and audio tagging can be combined into a single framework without significant degradation.

By contrast to UniKW-AT, this work does not train a new model but rather utilizes natural language captions as support to adjust tagging confidence of a pretrained audio tagging model. The audio caption is used to reassign a new boosted confidence based on possible tag alignment. Unified models improve within the audio domain, while caption guided approach can be applied to different audio tagging models and adds a language based multimodal approach, therefore complementing audio tagging through semantic context.

2. THEORETICAL BACKGROUND

This section covers all concepts that serve as the basis for the research carried out in the thesis. There are various ways of analysing audio tagging through machine learning methods such as convolutional neural networks, various long short-term memory architectures as well as different neural network models specifically for audio tasks (Purwins et al., 2019).

Contextuality is particularly useful where there is ambiguity with sound events and might be difficult to differentiate between events unimodally. Captions for audio clips gives textual descriptions from which NLP can be used to extract semantic information. The use of data from captions can provide more accuracy and relevance to tagging, resulting in improved tagging confidences. (Wu, Dinkel, & Yu, 2019).

2.1 Feature representation

Audio signals have many possible representations for feature extraction in machine learning. Some common representations include Mel Frequency Cepstral Coefficients (MFCCs), spectrograms and more specific forms and scales of spectrograms such as logarithmic-mel spectrograms. Due to the thesis using a specific model utilizing logarithmic-mel spectrogram, the focus will be on this representation.

A spectrogram is the Time-Frequency representation of an audio signal containing spectral content over time. It represents the time-frequency-intensity of the short time spectrum (A.V Oppenheim). Feature extraction from spectrograms means capturing specific patterns in the time-frequency distribution, highlighting relevant distinctions between audio events (Wolf-Monheim, 2024). Applying convolutional filters in deep learning allows for finding significant local structures in spectrograms for feature extraction (Purwins et al., 2019).

Figure 2.1 shows an example of a logarithmic mel spectrogram computed using PANNs settings. PANNs uses preprocessing parameters: 64 mel bands, sampling rate 32 kHz, window size 1024 samples, hop length 320 sample and mel frequency range 50 Hz to 14 kHz (Kong et al., 2020).

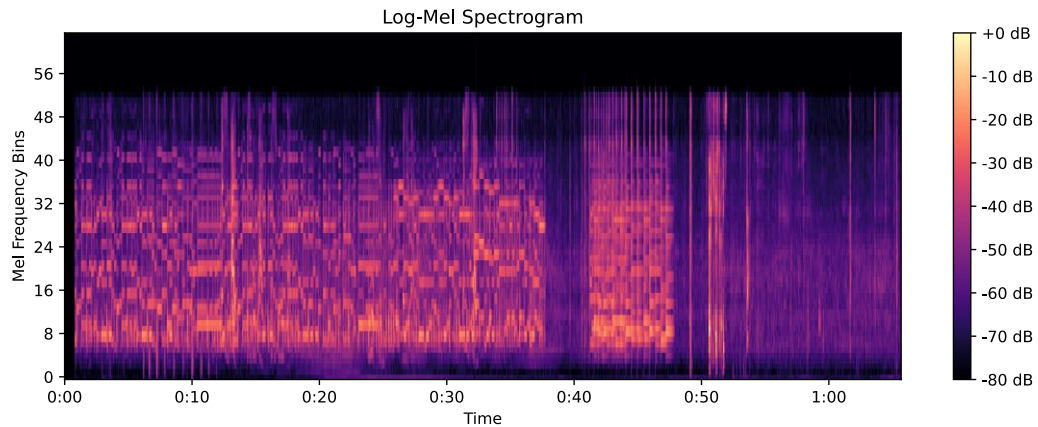


Figure 2.1. *Log-mel spectrogram of sample clip from AVCaps dataset with PANNs parameters*

2.2 Machine learning and deep learning in audio processing

Audio processing through machine learning methods usually requires transforming raw audio signals into different representations for more clarity and structure. A common approach, that for example PANNs model uses, is a Short-Time Fourier Transform (STFT). This allows conversion of time-domain audio into time-frequency domain. (Bellanger & Engel, 2024). Spectral patterns can be captured through the time-frequency domain. (Fu et al., 2024). Windowing functions are necessary in the STFT process to reduce the risk of spectral leakage and allows for preserving partial or local continuity. The time-frequency domain is represented through a 2-D spectrogram, which shows energy intensity at times and frequencies. This can be represented in different ways, for example changing frequency to mel-scale and taking the logarithm of it, which produces the log-mel spectrogram. Preprocessing is especially important for deep learning models such as Pretrained Audio Neural Networks (PANNs) as they rely on consistent features for training. (Purwins et al., 2019)

Deep learning is related to models with multiple neural network layers. Convolutional neural networks are a class of deep learning for processing spatial data. The models utilize convolutional layers by applying learning filters in local regions to detect features. Activation functions such as Rectified Linear Unit (ReLU) are used for non-linearity in the model. This allows for more complex mappings and more advanced learning. Pooling layers reduce dimensionality for reducing computational cost and ensure stable fitting. Fully connected layers interpret the extracted features to provide predictions. (Purwins et al., 2019).

Transformers are also deep learning models utilized especially in natural language processing. Transformers can process elements of input simultaneously through self-attention mechanisms. They are particularly useful due to parallelization and ability to process long sequences effectively. Due to the self-attention mechanism, the model can weigh relevance of input tokens compared to all others, no matter the sequence. (Vaswani et al., 2017).

CNNs have advanced the field of automatic tagging by learning hierarchical representations for features within waveform or spectrograms. The automatic learning and tagging allows for the models to capture relevant features and patterns for classification. Pre-trained Audio Neural Networks (PANNs) have made advancements in audio tagging through large-scale training with datasets with broad ontologies. (Kong et al. 2019) Pre-trained models extract acoustic features from audio clips, and they are effective in uni-modal tasks.

2.3 Sound event tagging with CNNs

Sound event tagging is the prediction and labelling of an event in an audio clip, while detection refers to also predicting the start and end times of the event. The general goal is to identify which tags are present in an audio clip or segment. In sound event tagging, the classification is usually multi-labelled due to multiple sound events happening simultaneously. (A. Mesaros et al. 2021). Figure 2.2 shows an example of what sound event tagging models produce: predicted tags and their corresponding confidences. The accuracy of a model or system can be evaluated through metrics such as f1-scores combining precision and recall when compared to ground truth.

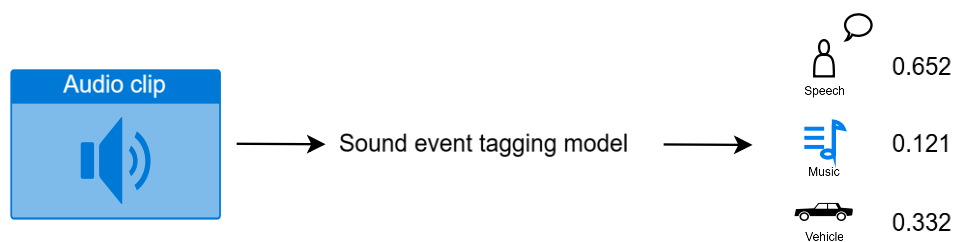


Figure 2.2. Example sound event tagging model

2.4 Natural Language Processing (NLP)

Natural language processing is a part of artificial intelligence. It allows for natural language data processing through computers. Natural language processing provides possibilities for handling natural language in terms of sentence structures, linguistic features as well as finding meaningful information. (Chen et al., 2024).

Integrating NLP techniques to introduce multimodal approaches can improve and enhance audio tagging through utilizing semantic relationships and context between tags. Contextual tagging using transformers allows for modelling dependencies between tags. (Özkaya Eren & Sert, 2021)

Natural language processing is especially useful as human-annotated captions are used as contextual information. NLP libraries make finding meaningful contextual information from natural language possible by for example removing unnecessary words such as common stop words. This is especially useful in this experiment since the unnecessary words could impact the results. (Chen et al., 2024). NLP also allows for data encoding into word vectors which allows for comparing similarities between different words and in this case tags.

Word and sentence vectors are numerical vector representations or embeddings. Models such as SBERT introduce Siamese network structures for fine-tuning BERT, enabling more semantic sentence embeddings. (Reimers & Gurevych, 2019). Word and sentence embeddings can be compared using similarity calculations such as cosine similarities which allows for semantic search.

Utilizing NLP-based processing and vectorization for text or captions can be beneficial for multimodal systems. It allows for integration of contextual information through different modalities, useful for classification tasks.

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based pre-trained model used to understand language and context of a word within a sentence. It uses transformer encoding stack with layers of self-attention and feedforward networks. BERT contextualizes word representations, therefore embedding based on entire sentences. (Devlin et al., 2019). This allows for more complex and semantic analysis compared to more common and simple word embeddings which has single vector representations for all words without context.

BERT processes text bidirectionality, therefore both sides of the text. It is pretrained with two main objectives. The first objective is Masked Language Masking for masking and

predicting random tokens. The second objective is Next Sentence Prediction, allowing the model to learn relationships between sentences. (Devlin et al., 2018).

The base version of BERT model is made up of 12 transformer layers (encoder blocks), each with multi-head self-attention mechanisms as well as feed-forward neural networks with a hidden size of 768. This represents the model's dimensionality. Contextual meaning within a sentence can be encoded simultaneously through the model's 12 attention heads. With 110M parameters, BERT can model complex language patterns. BERT can create significant, meaningful and contextualized embeddings for sentence similarity tasks. To optimize for fixed-size sentence embeddings, useful for sentence similarity, SBERT can be utilized. (Reimers & Gurevych, 2019).

SBERT, a sentence transformer can be utilized for not only words but has been trained to capture semantic relationships in sentences. BERT uses token-level embeddings while SBERT uses sentence-level embeddings. SBERT introduces Siamese or triplet networks with BERT encoding, meaning multiple BERT models share weights but encode input sentences independently to the fixed vector embeddings. This allows for more accurate contextual comparison. (Reimers & Gurevych, 2019).

Even with single words, SBERT is a trained model to handle semantic relationships, which suits the context. This means that we have more complex NLP utilization. SBERT provides better results in general for Semantic Textual Similarity (STS) tasks compared to other word embedding models (Reimers & Gurevych, 2019).

SpaCy is a Python library which allows for diverse use in NLP tasks such as extracting information, classification and text analysis. SpaCy can be utilized for extracting information from captions, such as linguistic structures, various types of phrases and individual words. The phrases and individual words can be processed for use by removing punctuation and predefined stop words using SpaCy. SpaCy can use CNNs, RNNs for models. Transformer-based architectures have recently been introduced to SpaCy which can improve accuracy. (Explosion AI, n.d.).

2.5 Contextual information in audio tagging

Multimodality is the use of more than one modality such as audio, visual or textual. In this work, the baseline for predicting tags comes from audio and the PANNs model. To be able to give tagging more confidence, multimodality is used through giving more contextual information to either boost or decrease confidence. The use of multimodality can

provide good contextual information which might be missed with unimodal approaches (Pahal et al., 2015).

Contextual information is important in audio tagging as to ensure accurate and reliable results. This means that only relevant data is utilized in the analysis of data. Contextual and relevant information is essential in comparing tags since it allows for accurate and justified boosting calculations for tagging confidence. It is possible to check relevance of words and sentences by calculating the similarity of two tags with a certain threshold.

3. RESEARCH METHODS AND DATA

This section introduces the proposed framework for improving audio tagging with contextual tags from audio captions. The proposed framework is illustrated in the block diagram in Figure 3.1.

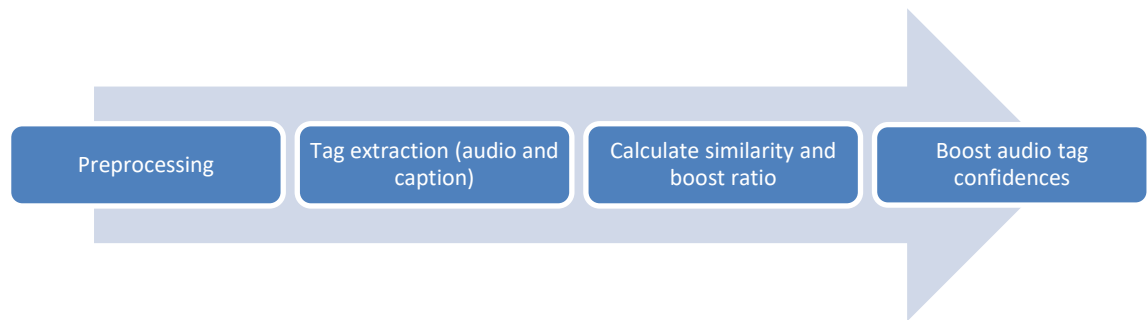


Figure 3.1. *Overview of stages in the work*

The goal of the study is to improve an audio event tagging model by boosting its confidence using tags extracted from audio captions. As a baseline we have the PANNs model's classification confidence. By using tags gathered from the captions, it is possible to compare the similarities between the methods of tagging to either boost confidence either positively or negatively.

3.1 Description of datasets: audio datasets and their captions

AVCaps is an audio-visual dataset with videos as well as corresponding modality-specific captions with video quantity of approximately 2000 videos and a total time of 28.8 hours. The dataset is derived from VidOR. The dataset is a resource designed for research in multimodal machine perception. AVCaps dataset provides crowdsourced captions for audio, visual, audio-visual as well as GPT-4 generated captions. For each audio clip there can be up to 5 different captions. (Sudarsanam et al.) This thesis limits and focuses on only audio captions to ensure that the captions are relevant to audio when comparing tags.

3.2 Proposed framework

In this part the chosen methods for the experiment are presented. The baseline audio tagging model is PANNs. SpaCy is an open-source library for caption NLP analysis in Python. SBERT is a sentence-transformer utilized for sentence vector representations. These tools combined provide the experiment base.

3.2.1 Preprocessing

The files from AVCaps are in video format (mp4), which require extraction to waveform file for audio processing. The files are processed with sampling rate 32kHz aligning with PANNs model. The audio clips are split into segments of 5 seconds with an overlap of 2 seconds or 40% using Python librosa audio analysis library. Overlapping is essential in audio processing as it reduces the risk of cutting off a classification within a clip.

The top three tags with highest confidence from each segment are extracted. After processing the whole audio clip, all unique tags with the highest corresponding confidence are used as the audio tags.

Processing text is important especially in this work as irrelevant words or excessive punctuation might influence results with tag similarity calculations (Rahimi & Homayounpour, 2023). All words from caption sentences are extracted excluding filler and stop words as well as punctuation. An example of extracted caption tags could be “music, playing, people, talking” extracted from sentence “music playing while people are talking”. This however might also take irrelevant tags into the analysis.

3.2.2 PANNs tags

The deep learning model used in the experiment is PANNs, short for Pretrained Audio Neural Networks which is a state-of-the-art convolutional neural network model pre-trained on AudioSet dataset. The chosen architecture is CNN14. (Kong et al., 2020)

The use of pretrained models rather than training a model from scratch allows for avoiding computationally demanding training processes. The models are trained on AudioSet dataset. (Kong et al., 2020) Specifically, “Cnn14_mAP=0.431” is the pretrained model used in the experiment with mean average precision (“mAP”) of 0.431.

PANNs CNN14 is a 14-layer convolutional neural network which takes an input of raw waveform and transforms it to log-mel spectrogram. The model has 6 convolutional layers with filter size 3x3. Each layer uses batch normalization and ReLU activation for stability and speed. (Kong et al., 2020). The architecture can be seen in Figure 3.2.

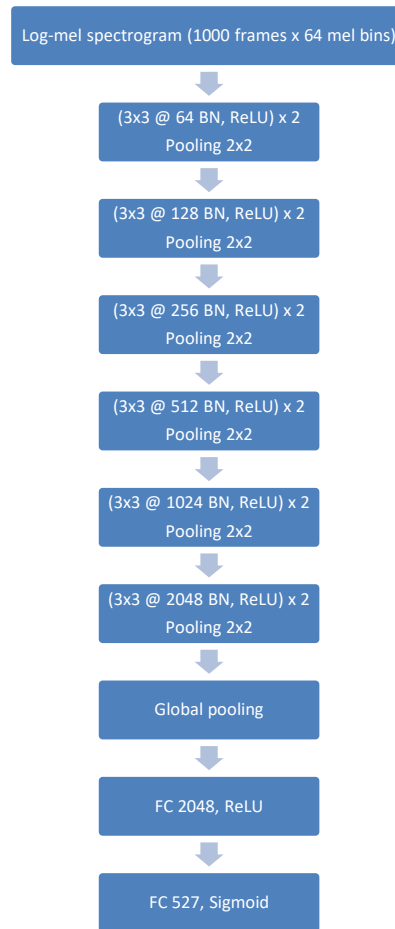


Figure 3.2. *PANNs CNN14 base architecture*

PANNs model accepts waveform input and transforms to spectrogram representation. The used model uses STFT with a hamming window to convert the waveform input into time-frequency domain. This results in a spectrogram for which a mel filterbank and logarithmic transformation is utilized for feature extraction.

The model extracts features from the logarithmic-mel spectrogram. It provides tag predictions and their corresponding confidences. This provides the baseline model for the experiment. The use of the PANNs model is based on it being pretrained on a large-scale dataset, providing an accurate baseline for the experiment. This model provides specific tags which come from AudioSet data. AudioSet consists of 527 classes seen through AudioSet ontology (Gemmeke et al., 2017). This provides good common tags to utilize in this experiment.

3.2.3 Caption tags

Caption tags are pre-processed and extracted using SpaCy library to gather all words within the audio captions. This results in words excluding stop and filler words as well as punctuation. All words are considered to ensure possible matches in tags used later in boosting. This might however introduce noise through irrelevant tags. Table 3.1 shows examples of extracted caption tags from captions.

Original caption	Extracted tags
A family is having a conversation while some are singing and others are laughing.	'family', 'having', 'conversation', 'singing', 'laughing'
The child says something cute and the parents listen to it and enjoy his cute voice.	'child', 'says', 'cute', 'parents', 'listen', 'enjoy', 'cute', 'voice'
People are laughing and chatting, and the child is singing a song.	'People', 'laughing', 'chatting', 'child', 'singing', 'song'

Table 3.1. Captions and corresponding extracted caption tags from clip 10433664864 from AVCaps dataset

3.2.4 Tag comparison

The tags extracted from the audio clip captions in the experiment are human-annotated and are originally in sentence form for which the use of SBERT can be beneficial. With sentences, SBERT can utilize dynamic word embeddings, effectively capturing and encoding contextual information, which is not possible with simpler tools word vector tools such as Word2vec.

By representing both audio and caption tags as vectors in the same vector space, it is possible to calculate their cosine similarity to determine if they are semantically similar. This is important in the experiment as tags extracted from the PANNs audio model differ to the human-annotated captions, therefore cannot be directly compared. It is necessary to calculate the similarities between the tags to ensure that relevant information through similar meaning is not missed in the analysis. The threshold for semantic similarity needs to be tested as a threshold too high will not find semantically similar tags and a threshold too low will take irrelevant tags into account.

This work utilizes the 'all-MiniLM-L6-v2' SBERT model. This model is a sentence-transformer model which maps sentences in 384 dimensional dense vector space, the SBERT architecture can be seen in Figure 3.3. The model can be utilized for clustering and semantic search. (Reimers & Gurevych, 2019)

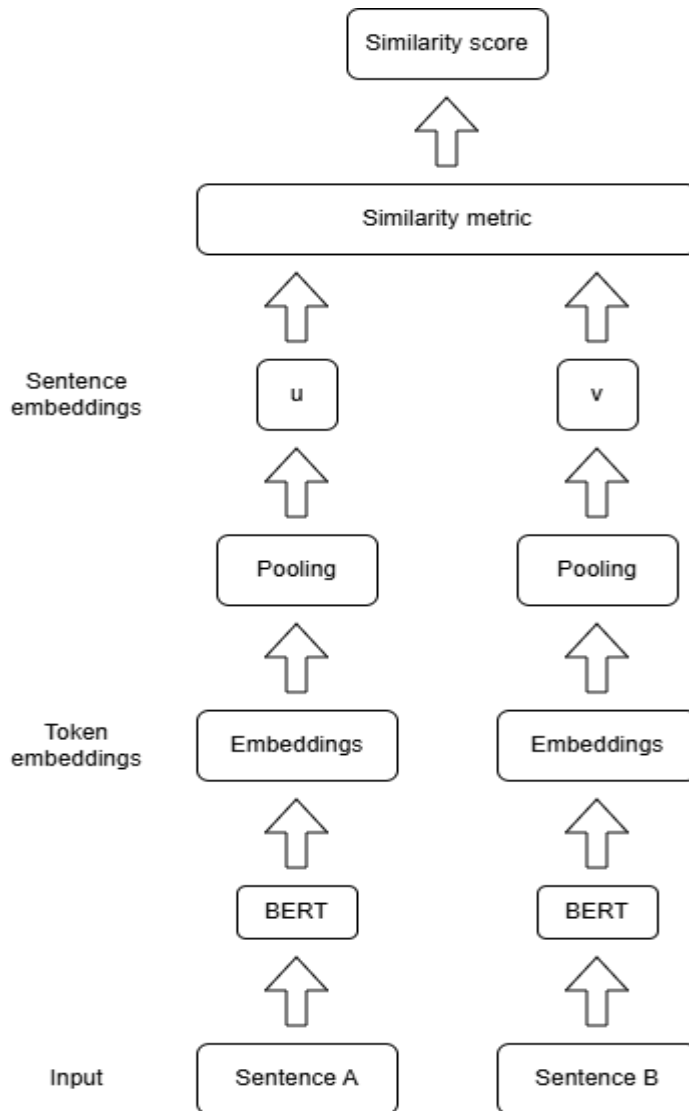


Figure 3.3. SBERT architecture

There are two different cases for using similarity checking: firstly, for calculating boosting ratios and secondly for comparing predicted sound event tags to ground truth tags.

3.2.5 Confidence boosting

The goal of this thesis is to boost confidence of tagging using tags obtained by sound event tagging model and captions. The baseline for tags and confidence comes from the

PANNs-model. For confidence boosting, caption tags are utilized. The boosting function is in a form which boosts tags with high correspondence and penalizes tags with low correspondence which aims for consistency and reliability for more accurate adjusted confidence values.

$$Conf_{boosted} = \alpha \times \left(\frac{\text{num in caption}}{\text{total num captions}} \right) + (1 - \alpha) \times \{\text{audio tagging confidence}\} \quad (3.1)$$

In order to use the mathematical boosting, it is necessary to gather information on the number of tags in captions as well as number of tags. In ratio $\frac{\text{num in caption}}{\text{total num captions}}$ from Equation 3.1, the numerator refers to the number of caption tags matching the processed sound event tag above a certain similarity threshold. The denominator refers to the number of captions the processed clip has.

Due to human-annotated tags differing from each other in captions, the similarity between extracted tags from captions and the main audio tag need to be compared and checked through tag similarity. This is done to ensure important information is not left out when processing since the tags may differ in literal word but might have the same meaning. To calculate the number of similar caption tags to the audio tag, a threshold is required to consider only meaningfully impacting words and ensure they correspond to the audio tag. The α value determines the weight of caption tags within the boosting method. The boosting ratio in caption weight is maxed out at value 1, meaning that higher ratios do not get more weight than a full ratio would.

4. RESULTS AND ANALYSIS

The objective of this work is to evaluate effectiveness of integrating natural language captions to improve tagging accuracy. Comparing baseline unimodal audio tagging and the proposed boosting method based on semantic similarity. For this work, 10 audio clips have been manually annotated to gather ground truth tags to use for evaluating the performance of the presented system.

4.1 Parameter and threshold analysis

A range of parameter combinations are tested to assess the performance of semantic similarity boosting on audio tagging. The parameters involved in the process are explained here:

The alpha parameter controls the weight of caption-based information impact on the confidence scores. This means that when alpha is 0, all the weight is in audio tagging model confidences, resulting in the original confidence. The tested alpha values are

$$\alpha \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}.$$

Combinations of confidence or acceptance thresholds $\tau_{conf} \in \{0.3, 0.5, 0.7\}$ for the predicted tags. This threshold dictates the low-bound confidence a predicted tag must achieve after boosting to be selected. The acceptance threshold is required to compare predicted tags after boosting against ground truth tags to evaluate system performance.

The semantic thresholds of $\tau_{sim} \in \{0.3, 0.5, 0.7\}$ are used to check semantic relatedness between tags in two different ways. Firstly, sound event tags and caption tags for boosting. Secondly for comparing predicted tags and ground truth tags to ensure words with similar meaning but different wording do not get discarded. An example of this would be ground truth tag “Music” and predicted tag “Musical instrument” or other specific instruments such as “guitar”, which are semantically similar.

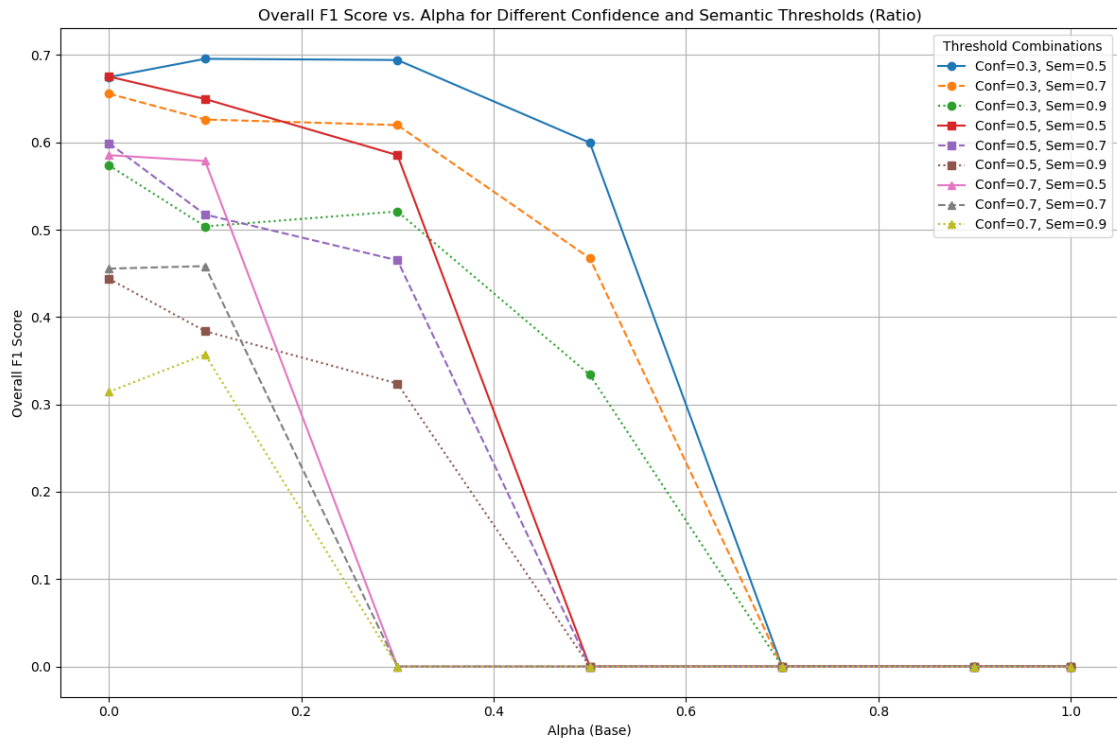


Figure 4.1. Average F1-score between clips using combinations of alpha and thresholds

Alpha	Confidence threshold	Similarity threshold	Precision	Recall	F1
0	0.3	0.5	0.657	0.781	0.675
0	0.3	0.7	0.651	0.767	0.656
0	0.3	0.9	0.576	0.669	0.574
0	0.5	0.5	0.845	0.648	0.675
0	0.5	0.7	0.829	0.55	0.599
0	0.5	0.9	0.77	0.388	0.444
0	0.7	0.5	0.825	0.525	0.585
0	0.7	0.7	0.807	0.385	0.455
0	0.7	0.9	0.69	0.262	0.314
0.1	0.3	0.5	0.689	0.781	0.696
0.1	0.3	0.7	0.667	0.692	0.626
0.1	0.3	0.9	0.602	0.544	0.504
0.1	0.5	0.5	0.925	0.583	0.65
0.1	0.5	0.7	0.9	0.443	0.517
0.1	0.5	0.9	0.783	0.32	0.384
0.1	0.7	0.5	0.775	0.5	0.579
0.1	0.7	0.7	0.767	0.36	0.458
0.1	0.7	0.9	0.75	0.262	0.357
0.3	0.3	0.5	0.775	0.723	0.694
0.3	0.3	0.7	0.749	0.633	0.62
0.3	0.3	0.9	0.706	0.519	0.521
0.3	0.5	0.5	0.825	0.525	0.585
0.3	0.5	0.7	0.817	0.385	0.465
0.3	0.5	0.9	0.7	0.262	0.324
0.3	0.7	0.5	0	0	0
0.3	0.7	0.7	0	0	0

0.3	0.7	0.9	0	0	0
0.5	0.3	0.5	0.825	0.55	0.6
0.5	0.3	0.7	0.8	0.41	0.467
0.5	0.3	0.9	0.683	0.287	0.334
0.5	0.5	0.5	0	0	0

Table 4.1. Average F1-scores between clips using combinations of alpha and thresholds

The parameter combinations and their corresponding precision, recall and F1-scores can be seen in Figure 4.1 as plotted and Table 4.1 in table format. Using these results, it is possible to evaluate the performance of the system with different parameters as well as analyse how the parameters affect results.

Baseline performance ($\alpha = 0$), meaning all system weight is placed on audio tagging model’s confidences shows F1-scores ranging from (0.3-0.7), averaging approximately 0.6. Light weight (0.1-0.3) for caption yields the best results, pushing highest F1-scores over baseline with some combinations especially with threshold values around 0.5.

Excess reliance on captions ($\alpha > 0.5$) causes significant decreases in F1-scores. At alpha values over 0.7, where most weight is on captions, F1-scores plummet to zero. Values above combination $\alpha = 0.5$, $\tau_{conf} = 0.5$, $\tau_{sim} = 0.5$ are discarded from Table 4.1 due to all resulting f1-scores being zero. This can happen due to two reasons; the boosting is based on audio tagging confidences as well as captions not providing enough value to substitute for audio confidences.

There is a trade-off between thresholds. With loose relatedness thresholds ($\tau_{sim} = 0.3$), the system ensures relevant tags are accounted for while also accepting some noise. With strict relatedness ($\tau_{sim} = 0.7$), valid synonyms and related tags are filtered out, meaning that not all relevant tags are correctly accepted. To handle both situations where trade-offs occur, $\tau_{sim} = 0.5$ is an acceptable middle-ground.

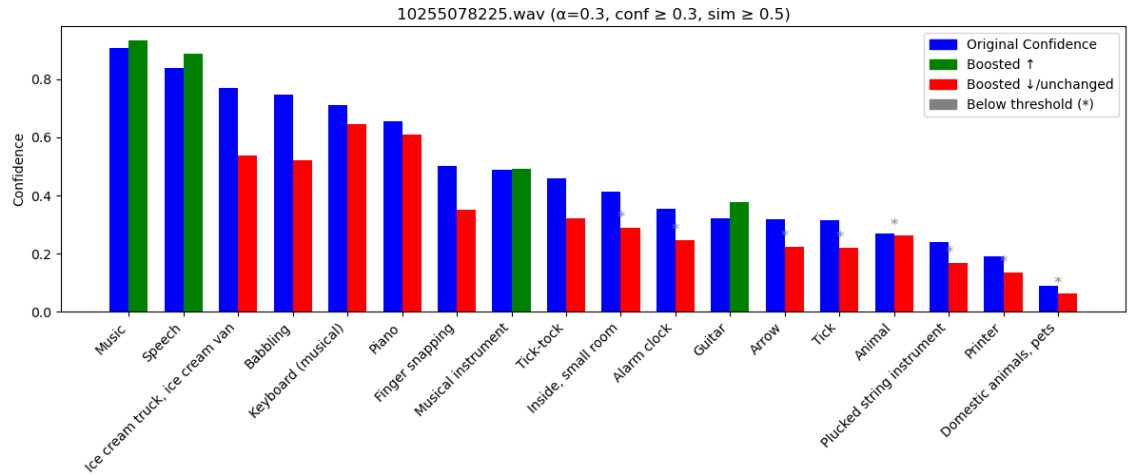
The best overall result is achieved with light weight $\alpha = 0.1$ with $\tau_{conf} = 0.3$ and $\tau_{sim} = 0.5$, with an F1-score of 0.696. With same threshold values and a slightly higher weight value of $\alpha = 0.3$, the F1-score is near the previous score with 0.694.

Overall, the slight weighting of semantically filtered captions ($\alpha = 0.1 - 0.3$) with moderate acceptance thresholds yield the best results. This indicates that contextual captions as a multimodal approach can complement sound event tagging under specific conditions.

4.2 Class-specific results

In this part, class-specific results are analysed to evaluate connections and patterns between tag classes. Three example clips are used to showcase how boosting with low α values and moderate τ_{conf} and τ_{sim} influences confidence scores between classes.

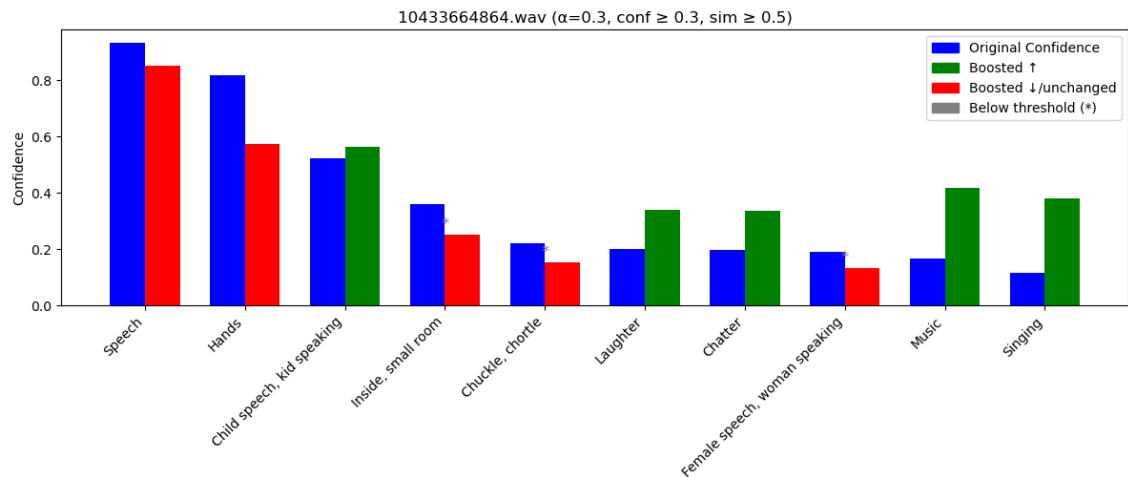
Ground Truth: music, speech



Caption Words: mild, music, playing, background, man, speaking, man, speaking, loud, voice, audio, playing, voices, kids, hilarious, music, background, child, playing, Christmas, song, toy, dad, comments

Figure 4.2. Original vs boosted confidence scores between classes for clip 10255078225 from AVCaps dataset

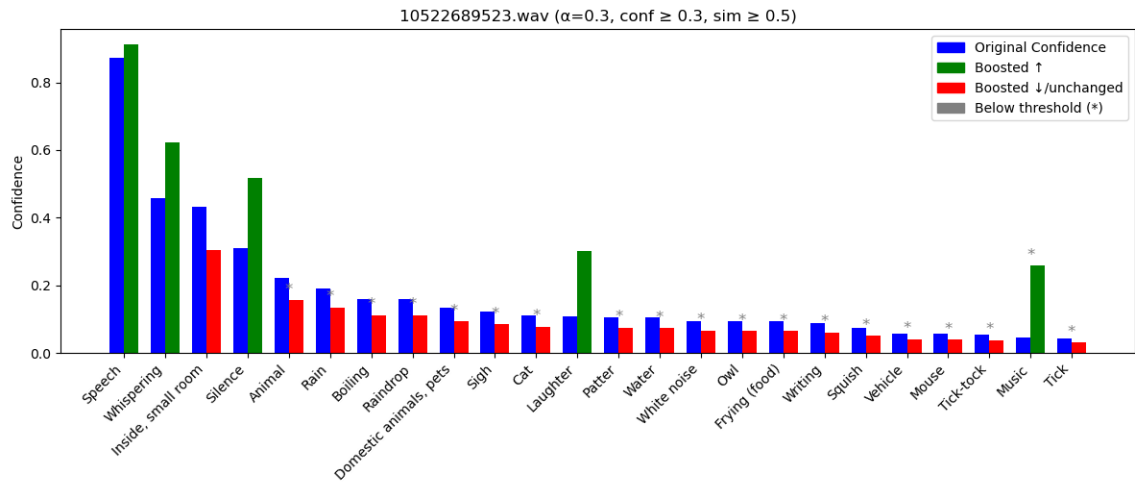
Ground Truth: speech, child speech, kid speaking, laughter, chatter, chuckle, chortle, music



Caption Words: family, having, conversation, singing, laughing, child, says, cute, parents, listen, enjoy, cute, voice, People, laughing, chatting, child, singing, song

Figure 4.3. Original vs boosted confidence scores between classes for clip 10433664864 from AVCaps dataset

Ground Truth: speech, whispering, silence, laughter



Caption Words: man, lady, speaking, quietly, eventually, audio, audible, end, lady, burst, loud, laughter, people, whispering, laughing, quiet, environment, sound, beginning, middle, Suddenly, man, spoke, softly, brief, moment, followed, period, silence, long, time, woman, spoke, softly, end, laughed, man, woman, talking, quietly, background

Figure 4.4. Original vs boosted confidence scores between classes for clip 10522689523 from AVCaps dataset

Tag	Original	Matches	Ratio	Boosted	Present
Music	0.908	14	1.000	0.954	1
Piano	0.656	5	1.000	0.828	0
Keyboard (musical)	0.710	4	0.800	0.755	0
Speech	0.838	3	0.600	0.719	1
Guitar	0.323	5	1.000	0.662	0
Musical instrument	0.490	4	0.800	0.645	0
Ice cream truck, ice cream van	0.770	0	0.000	0.385	0
Babbling	0.747	0	0.000	0.373	0
Finger snapping	0.502	0	0.000	0.251	0
Tick-tock	0.460	0	0.000	0.230	0
Inside, small room	0.414	0	0.000	0.207	0
Alarm clock	0.353	0	0.000	0.176	0
Arrow	0.317	0	0.000	0.158	0
Tick	0.316	0	0.000	0.158	0
Animal	0.269	0	0.000	0.135	0
Plucked string instrument	0.238	0	0.000	0.119	0
Printer	0.191	0	0.000	0.096	0
Domestic animals, pets	0.088	0	0.000	0.044	0

Table 4.2. Boosted scores and match ratios for clip 10255078225 from AVCaps dataset

Class-specific boosting figures show examples of label specific boosting. In three different example clips using parameter values $\alpha = 0.3$, $\tau_{\text{conf}} = 0.3$, $\tau_{\text{sim}} = 0.5$, it is possible to see how the confidence adjustment relates to captions and ground truth.

More high-level, general and frequent tags (e.g. “Music” or “Speech”) tend to receive largest boosts most commonly due to common appearances or semantic similarities to different tags (e.g. “Musical instrument”, “Talking”).

Table 4.2 has same results as in Figure 4.2, represented in table format. It shows how similar words get similar boosting ratios, indicating semantic similarity. For example, “music” tag has 14 semantic matches and other similar words such as “piano”, or “guitar” get

5 semantic matches, both still resulting in a boost ratio of 1. This approach gives the same boost to both cases.

Failure cases with the system can arise from captions which are based on background events can lead to false positives with loose semantic thresholds. Rare and specific tags (e.g. "Printer", "Arrow") might be over-generalized depending on semantic similarity threshold as seen in Figure 4.2. Another failure case might arise when tag classes are multiple words (e.g. "Domestic animals, pets") in Figures 4.2 and 4.4. Another example of this can be seen in Figure 4.2 where most "music" related tags are boosted, a longer and more specific "plucked string instrument" tag is not boosted.

Figures 4.2, 4.3 and 4.4 provide examples of two clips where boosting can be seen based between different classes. They showcase semantic connections between classes. These figures show that caption quality is significant for boosting as captions with direct similarity or close synonyms, recall is reliably increased without harming precision. In addition to caption quality, threshold tuning is critical, as semantic thresholds too low introduce noise and irrelevant data, and semantic thresholds too high limit, filter and miss out on relevant tags.

5. SUMMARY AND CONCLUSIONS

5.1 Key findings

Overall results show that utilizing this caption based boosting method has a trend of not improving sound event tagging, although the best results were obtained with slight caption weight. This is due to the boosting mechanism being strict with when it boosts therefore penalizing most labels. The boosting mechanism works well with more general and frequent labels, highlighting their presence.

The best overall f1-score (0.696) from processing the 10 clips is obtained when $\alpha = 0.1$, $\tau_{conf} = 0.3$, $\tau_{sim} = 0.5$. This combination of thresholds produces the best overall performance over α values. Compared to $\alpha = 0$, this combination of thresholds improves the f1-scores over values $\alpha = 0.1$, $\alpha = 0.3$. The slight weight given to captions shows that there are semantically meaningful tags within the captions which can be obtained for boosting. There are some other combinations which boost as well with higher alpha values.

In terms of label or class boosting, tagging accuracy generally improves with more common or general labels such as “Music” or “Speech” as they appear in audio tagging and captions frequently. Another reason for this may be their possibly high semantic similarity to other words.

Tags with adjusted confidences are also adjusted for semantically similar tags as they are contextually the same with different labels. Confidence adjustment is correlated between semantic similarity between tags. The confidence adjustment method can also decrease confidences in tags which do not appear in captions, considering possible semantic similarity. This would refer to possible contextual irrelevance, therefore not disregarding but rather lowering prediction confidence. This illustrates the model’s ability to leverage semantic context beyond direct tag matches. The confidence adjustment method’s ability to adjust tagging confidence based on contextual significance reduces false positives and negatives.

The main reason for the overall confidence adjustment f1-scores decrease is due to boosting being based on the audio tagging confidences. As stated, there are possibilities for relevant tags discarded and irrelevant tags considered based on how the model’s handle and predict the tags.

Overall, the method showcases the advantages of a multimodal approach to audio tagging, combining information through audio and captions to produce more accurate and semantically coherent results.

5.2 Limitations

In this work, 10 audio clips were manually annotated, which is not enough for obtaining the best set of parameters but gives an understanding of how parameters affect the results and estimates for possible parameters.

Another limitation for tagging adjustment is that PANNs tags cannot be directly compared to caption tags as PANNs has own AudioSet ontology while captions are annotated and not following any specific form. Since the comparison is indirect, both tags need to be encoded, and their similarities checked in vector space. This issue can be resolved using SBERT for sentence embedding for tag comparison with a specific similarity threshold which can cause relevant words to be missed and irrelevant words to be considered. Cosine similarity assumes modalities aligning in embedding space.

When dealing with ground truth tags and self-annotation, it is impossible to ensure full validity of the tag, especially for very specific tags. An example of this would be when "Music" tag is certainly present, but it is difficult to say the same about and differentiate specific musical instruments. For ground truth, a semantic similarity threshold is added to check for predicted tags compared to ground truth through cosine similarity as the tags are not directly comparable.

Low confidences in PANNs might lead to relevant tags being missed out since only top 3 tags from each segment are taken into consideration. There might be more overpowered audio events which suppress other sounds.

There are possible trade offs in the using models in terms of embedding dimensionality, accuracy and speed. The used models are pretrained and could possibly be fine-tuned to suit the task more accurately. The general pretrained models are good enough for baseline tasks.

5.3 Future research

Currently only auditory and textual information are utilized. Introducing yet another modality such as video could provide more contextual information for tagging confidence adjustments and give another perspective for boosting.

Different ways to gather caption tags could provide more accurate information compared to gathering all words. Subject-Verb-Object triplets can be obtained using NLP, which was used in earlier stages of this experiment, but changed due to complexity.

Currently sentence embeddings are done using SBERT and similarities checked using cosine similarity. A more sophisticated approach for this could be to use cross-modal transformers or multimodal contrastive learning for alignment.

6. REFERENCES

Bellanger, M., & Engel, B. A. (2024). The discrete Fourier transform. In *Digital signal processing: Theory and practice* (pp. 35–54). <https://doi.org/10.1002/97811394182695.ch2>

Chen, K., Fei, C., Bi, Z., Liu, J., Peng, B., Zhang, S., Pan, X., Xu, J., Wang, J., Yin, C. H., Zhang, Y., Feng, P., Wen, Y., Wang, T., Li, M., Ren, J., Niu, Q., Chen, S., Hsieh, W., Yan, L. K. Q., Liang, C. X., Xu, H., Tseng, H.-M., Song, X., & Liu, M. (2024). *Deep learning and machine learning – Natural language processing: From theory to application* (arXiv preprint arXiv:2411.05026). <https://arxiv.org/abs/2411.05026>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. <https://arxiv.org/abs/1810.04805>

Dinkel, H., Wang, Y., Yan, Z., Zhang, J., & Wang, Y. (2022, September). *UniKW-AT: Unified keyword spotting and audio tagging*. In *Interspeech 2022* (pp. 3238–3242). ISCA. <https://doi.org/10.21437/Interspeech.2022-607>

Explosion AI. (n.d.). *spaCy 101: Everything you need to know*. Retrieved March 24, 2025, from <https://spacy.io/usage/spacy-101>

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 776–780). IEEE. <https://doi.org/10.1109/ICASSP.2017.7952261>

Gong, Y., Chung, Y.-A., & Glass, J. (2021). *AST: Audio Spectrogram Transformer*. arXiv preprint arXiv:2104.01778. <https://arxiv.org/abs/2104.01778>

Hamaguchi, R., Sakurada, K., & Nakamura, R. (2018). Rare event detection using disentangled representation learning. *arXiv preprint arXiv:1812.01285*. <https://arxiv.org/abs/1812.01285>

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2019). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition (pretrained models). *Zenodo*. <https://doi.org/10.5281/zenodo.3987831>

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2880–2894.

Koutini, K., Schlüter, J., Eghbal-zadeh, H., & Widmer, G. (2022). Efficient training of audio transformers with patchout. In *Interspeech 2022*. ISCA. <https://doi.org/10.21437/Interspeech.2022-227> *

Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5), 67–83. <https://doi.org/10.1109/MSP.2021.3090678>

Oppenheim, A. V. (1970). Speech spectrograms using the fast Fourier transform. *IEEE Spectrum*, 8(7), 57–62.

Pahal, N., Chaudhury, S., Lall, B. (2015). Context-Based Semantic Tagging of Multimedia Data. In: Kryszkiewicz, M., Bandyopadhyay, S., Rybinski, H., Pal, S. (eds) Pattern Recognition and Machine Intelligence. PReMI 2015. Lecture Notes in Computer Science(), vol 9124. Springer, Cham. https://doi.org/10.1007/978-3-319-19941-2_17

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2), 206–219. <https://arxiv.org/abs/1905.00078>

Rahimi, Z., Homayounpour, M.M. The impact of preprocessing on word embedding quality: a comparative study. *Lang Resources & Evaluation* **57**, 257–291 (2023). <https://doi.org/10.1007/s10579-022-09620-5>

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*. <https://arxiv.org/abs/1908.10084>

Sudarsanam, P., Kumar, A., Yu, J., Kumar, N., & Wu, H. (2024). AVCaps: An audio-visual dataset with modality-specific captions. *Zenodo*. <https://doi.org/10.5281/zenodo.14536325>

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). *Large language models for data annotation and synthesis: A survey* (arXiv preprint arXiv:2402.13446). <https://arxiv.org/abs/2402.13446>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. <https://arxiv.org/abs/1706.03762>

Wang, D., Zhang, L., Bao, C., Xu, K., Zhu, B., & Kong, Q. (2018). *Weakly supervised CRNN system for sound event detection with large-scale unlabeled in-domain data*. arXiv preprint arXiv:1811.00301. <https://arxiv.org/abs/1811.00301>

Wolf-Monheim, F. (2024). Spectral and rhythm features for audio classification with deep convolutional neural networks. *arXiv preprint arXiv:2410.06927*. <https://arxiv.org/abs/2410.06927>

Wu, M., Dinkel, H., & Yu, K. (2019). Audio caption: Listen and tell. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 830–834). IEEE. <https://doi.org/10.1109/ICASSP.2019.8682377>

Yao, Y., Hua, X., Gao, G., Sun, Z., Li, Z., & Zhang, J. (2020). Bridging the web data and fine-grained visual recognition via alleviating label noise and domain mismatch. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1735–1744). Association for Computing Machinery. <https://doi.org/10.1145/3394171.3413851>

Özkaya Eren, A., & Sert, M. (2021). Audio captioning with composition of acoustic and semantic information.