Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Instytut Informatyki

Praca dyplomowa inżynierska

na kierunku Informatyka
w specjalności Sztuczna Inteligencja

System do analizy dźwięków jelitowych

Igor Matynia
Numer albumu 318693

promotor
dr hab. inż. Robert Nowak, prof. uczelni

WARSZAWA 2024

System do analizy dźwięków jelitowych

Streszczenie. Streszczenie

Słowa kluczowe: dźwięki jelitowe, sztuczne sieci neuronowe, sieci konwolucyjne, wykrywanie wzorców dźwiękowych

System for bowel sound analysis

Abstract. Abstract

Keywords: bowel sounds, neural networks, convolutional neural networks, sound pattern recognition

załącznik nr 3 do zarządzenia
nr 28 /2016 Rektora PW

Politechnika Warszawska

……..
miejscowość i data

……..
imię i nazwisko studenta

……..
numer albumu

…….…….
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta

Spis treści

1. Wstęp

Wstęp.

2. Opis rozwiązania

2.1. Introduction

Detecting bowel sounds in audio recordings presents a unique challenge, as it involves identifying specific patterns within a time series of sound. The goal is to determine precise time ranges within which bowel sounds occur. To address this, a novel two-step approach has been developed, adapting techniques commonly used in image-based object detection to the domain of audio analysis.

The proposed solution divides the problem into two main phases: a broad binary classification to identify potential regions of interest and a precise refinement phase to accurately pinpoint the time ranges of the detected sounds. This is achieved through the cooperation of two convolutional neural networks:

- Window Classification Model - This model performs a preliminary scan of the audio, identifying regions that potentially contain bowel sounds. It acts as a filter, selecting areas that merit further examination.
- Time Range Estimation Model - The second model focuses on the regions highlighted by the first model, estimating a more precise time range within the selected window where the bowel sounds are located.

This method draws inspiration from Region-CNN (R-CNN) architectures commonly used in image object detection. In R-CNN, a model first selects areas of an image for closer analysis and then refines the location of objects within those regions using bounding boxes. Similarly, this approach adapts the concept to one-dimensional time series data found in audio files, enabling efficient detection and localization of bowel sounds.

2.2. Data preparation

Before training can begin, the data must be prepared. The process begins by organizing the data files into a structured directory. A main data folder is set up, containing a subdirectory named raw, which houses all .wav audio files and their corresponding ground truth annotations in .csv format. Each .wav file is paired with a .csv file that contains the time ranges for annotated bowel sounds. These pairs are then assigned to the training, validation, or test datasets during the configuration file creation phase.

A configuration file can be generated using an automated script, which ensures that each file is assigned to the appropriate dataset in correct proportions, according to predefined criteria.

Once the dataset allocation is complete, the .wav files are converted into spectrograms. The spectrograms are normalized individually to ensure consistent scaling, and then concatenated into a single large spectrogram that covers the entire dataset. This combined spectrogram is saved processed folder in local storage for efficient access during the training, validation, and testing phases.

For the spectrogram generation, a low-pass filter is applied to the audio data, allowing only frequencies below 2000 Hz. This threshold was selected based on ground truth annotations, which indicated that bowel sounds rarely exceed this frequency range. The parameters for the spectrogram generation, such as Fast Fourier Transform (FFT) size and hop length, are chosen to ensure that each 0.2-second window has a time resolution of 315 bins and a frequency resolution of 126 bins.

After generation, the processed spectrogram data is saved locally in a processed folder. These pre-generated spectrogram files are used as input during training and inference, significantly reducing processing time and improving efficiency.

2.3. Classification model

The classification model is responsible for identifying regions within the audio that warrant further analysis to accurately pinpoint bowel sounds. This model is implemented as a convolutional neural network (CNN) that processes 0.2-second segments of the audio, represented as monochromatic spectrograms. Each spectrogram segment is of size 315 (time resolution) by 126 (frequency resolution) pixels as mentioned in the data preparation chapter.

The architecture of the model consists of three convolutional layers, each followed by max pooling layers to progressively reduce the spatial dimensions while capturing relevant features. These layers are followed by two fully connected linear layers, which utilize Leaky ReLU activation functions. The output layer consists of a single neuron with a sigmoid activation function, which outputs a confidence score indicating the likelihood of a bowel sound being present in the given 0.2-second window.

This binary output—ranging from 0 to 1—serves as a filter, determining which regions of the audio should be passed on to the subsequent stage for more precise time range estimation.

2.4. Time range estimation model

The Time Range Estimation Model refines the detection by pinpointing the exact time range of the bowel sound within a candidate window. It takes as input the same monochromatic spectrogram segment that the Classification Model identified as potentially containing a bowel sound. The output consists of two neurons whose values define the time range of the bowel sound, represented as a scale and offset relative to the centre of the input window.

The time range is interpreted as follows:

- Scale - Determines the length of the detected region relative to the input window.
- Offset - Represents a shift from the centre of the window, indicating the exact position of the bowel sound within the segment.

For example, a bounding box with a scale of 1.0 and an offset of 0.5 would indicate that the entire window contains a bowel sound, with no adjustment necessary. However, a window with parameters of 0.2 in scale and 0.25 in offset would mean—assuming a 0.2-second window length—that the bowel sound ranges from 0.03 seconds to 0.07 seconds from the beginning of the window.

The architecture of this model mirrors that of the Classification Model, featuring three convolutional layers followed by two fully connected layers. The internal layers utilize Leaky ReLU activation to introduce non-linearity, while the output layer employs a sigmoid activation function to ensure that the scale and offset values remain within a meaningful range.

This configuration allows the model to accurately adjust the detection window, honing in on the precise time range of the bowel sound based on the input spectrogram.

2.5. Training

The training process involves iteratively refining the models through a series of training loops. In each loop, windows are randomly sampled from pre-defined sampling regions, which are calculated prior to training. These regions delineate areas that can be classified as either containing bowel sounds or being free of such sounds. This preparation ensures that both positive (bowel sound) and negative (no bowel sound) examples are available for training. Each positive region as its own best matching bowel sound assigned.

To enhance model robustness, data augmentation is applied in the form of Gaussian noise. This noise is added directly to the spectrogram, with a randomly selected standard deviation ranging between 0 and 0.25. This approach helps the model generalize better by making it resilient to minor variations in the input data.

Both convolutional networks—the Classification Model and the Time Range Estimation Model—are trained independently, albeit on the same dataset.

- Classification Model - This model is trained on both positive and negative samples. Due to a significant imbalance in the dataset—where roughly 90% of the audio recordings lack bowel sounds—more negative samples are used during training. To counteract this imbalance, a weighted binary cross-entropy loss function is employed. The weighting adjusts the importance of bowel sound regions relative to empty regions in a 3:1 ratio, ensuring that the model does not default to always predicting the absence of bowel sounds. This weighted approach is consistent in both the sampling of the training data and the calculation of the loss.
- Time Range Estimation Model - In contrast, this model is trained exclusively on windows that contain bowel sounds. The underlying assumption is that the Time Range Estimation Model only processes windows when bowel sounds are present, as filtered by the Classification Model. The loss function utilized here is a modified Intersection over Union (IoU) loss, adapted for one-dimensional data. This modi-

fication allows the model to more accurately predict time ranges within the 1-D spectrogram input, capturing the precise boundaries of bowel sounds.

Overall, this two-model training approach ensures that each model is specialized for its respective task: the Classification Model for identifying potential regions of interest and the Time Range Estimation Model for fine-tuning those detections to determine precise time ranges.

2.6. Inference

The inference phase involves using the trained models to detect bowel sounds in new audio recordings.

This process is controlled by three key parameters that influence how the algorithm behaves:

- Detection Threshold - Defines the minimum confidence level that the Classification Model must achieve to mark a window for further analysis. If the model's confidence score for a window exceeds this threshold, the window is considered likely to contain a bowel sound.
- Window Overlap - This parameter specifies the degree of overlap between consecutive windows as the algorithm scans through the audio recording. A higher overlap means that each part of the recording will be analysed multiple times from different perspectives, enhancing detection accuracy.
- Vote Fraction - This is the final decision-making threshold, determining which regions are ultimately marked as containing bowel sounds. It is calculated as the sum of detection scores across overlapping windows, normalized by the number of overlaps. This parameter decides the threshold at which a region is classified as containing a bowel sound.

The inference process begins by converting the audio recording into a sequence of spectrogram windows. These windows are staggered according to the Window Overlap parameter, ensuring that each segment of the recording is reviewed multiple times.

The Classification Model is then applied to each window, producing a confidence score that indicates the likelihood of the presence of a bowel sound. If the confidence exceeds the specified Detection Threshold, the window is marked for further analysis. At this point, the Time Range Estimation Model is used to estimate the precise location of the bowel sound within the selected window.

Each estimated region is treated probabilistically—the model's confidence score is distributed over the entire detected area. These scores are accumulated for overlapping windows, providing a composite measure of confidence for each part of the recording.

In the final step, the Vote Fraction threshold is applied to these accumulated scores to determine the definitive locations of bowel sounds within the audio. Regions where

the average confidence meets or exceeds this threshold are classified as containing bowel sounds.

This multi-step process, leveraging overlapping windows and probability-based confidence, ensures that bowel sounds are detected accurately and minimizes the risk of false positives or missed detections.

Wykaz symboli i skrótów

EiTI – Wydział Elektroniki i Technik Informacyjnych
PW – Politechnika Warszawska

Spis rysunków

Spis tabel

Spis załączników