Efficient and precise annotation of local structures in data

Efficient and precise annotation of local structures in data

by John Martinsson

LUND
UNIVERSITY

Licentiate Thesis
Thesis advisors: Maria Sandsten, Olof Mogren
Faculty opponent: Annamaria Mesaros

To be presented, with the permission of the LTH of Lund University, for public criticism in the room MH:309A at the Centre for Mathematical Sciences, Mathematical Statistics on Thursday, the 3rd of October 2024 at 10:15.

Organization
LUND UNIVERSITY

Centre for Mathematical Sciences, Mathematical Statistics
Box 118
SE–221 00 LUND, Sweden

Author(s)
John Martinsson

Document name
Licentiate thesis

Date of presentation
2024-10-03

Sponsoring organization The Swedish Foundation for Strategic Research (SSF; FID20-0028), and Sweden's Innovation Agency (2023-01486)

Title and subtitle
Efficient and precise annotation of local structures in data:

Abstract
Machine learning models are used to help scientists analyze large amounts of data across all fields of science. These models become better with more data and larger models mainly through supervised learning. Both supervised learning and model validation benefit from annotated datasets where the annotations are of high quality. A key challenge is to annotate the amount of data that is needed to train large machine learning models. This is because annotation is a costly process and the collected labels can vary in quality. Methods that enable cheap annotation of high quality are therefore needed.

In this thesis we consider ways to reduce the annotation cost and improve the label quality when annotating local structures in data. An example of a local structure is a sound event in an audio recording, or a visual object in an image. By automatically detecting the boundaries of these structures we allow the annotator to focus on the task of assigning a textual description to the local structure within those boundaries. In this setting we analyze the limits of a commonly used annotation method and compare that to an oracle method, which acts as an upper bound on what can be achieved. Further, we propose new ways to perform this kind of annotation that results in higher label quality for the studied datasets at a reduced cost. Finally, we study ways to reduce annotation cost by making the most use of each annotation that is given through better modelling approaches in general.

Key words
Annotation efficiency, Machine learning, Sound event detection

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-028X

ISBN
978-91-8104-199-6 (print)
978-91-8104-200-9 (pdf)

Recipient's notes

Number of pages
42

Price

Security classification

Efficient and precise annotation of local structures in data

by John Martinsson

LUND
UNIVERSITY

Funding information: The Swedish Foundation for Strategic Research (SSF; FID20-0028), and Sweden's Innovation Agency (2023-01486)

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

www.maths.lth.se

Licentiate Thesis in Mathematical Sciences 2024:3
ISSN: 1404-028X

ISBN: 978-91-8104-199-6 (print)
ISBN: 978-91-8104-200-9 (pdf)
LUTFMS-2020-2024

Printed in Sweden by Media-Tryck, Lund University, Lund 2024

Till Lara

Abstract

Machine learning models are used to help scientists analyze large amounts of data across all fields of science. These models become better with more data and larger models mainly through supervised learning. Both supervised learning and model validation benefit from annotated datasets where the annotations are of high quality. A key challenge is to annotate the amount of data that is needed to train large machine learning models. This is because annotation is a costly process and the collected labels can vary in quality. Methods that enable cheap annotation of high quality are therefore needed.

In this thesis we consider ways to reduce the annotation cost and improve the label quality when annotating local structures in data. An example of a local structure is a sound event in an audio recording, or a visual object in an image. By automatically detecting the boundaries of these structures we allow the annotator to focus on the task of assigning a textual description to the local structure within those boundaries. In this setting we analyze the limits of a commonly used annotation method and compare that to an oracle method, which acts as an upper bound on what can be achieved. Further, we propose new ways to perform this kind of annotation that results in higher label quality for the studied datasets at a reduced cost. Finally, we study ways to reduce annotation cost by making the most use of each annotation that is given through better modelling approaches in general.

Publications

Publications concerning the work of this thesis have been made as follows:

A **Modelling the annotation quality and cost of weak labeling of fixed length segments in audio data**
John Martinsson, Olof Mogren, Tuomas Virtanen, Maria Sandsten
Unpublished manuscript

B **From weak to strong sound event labels using adaptive change-point detection and active learning**
John Martinsson, Olof Mogren, Maria Sandsten, Tuomas Virtanen
32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024. (Nominated for best student paper.)

C **DMEL: the differentiable log-Mel spectrogram as a trainable layer in neural networks**
John Martinsson, Maria Sandsten
ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, 2024.

D **Few-shot bioacoustic event detection using an event-length adapted ensemble of prototypical networks**
John Martinsson, Martin Willbo, Aleksis Pirinen, Olof Mogren, Maria Sandsten
Proceedings of the 7th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2022), Nancy, France, 2022, pp. 121-125.

.

Acknowledgements

Firstly, I would like to thank both of my supervisors Olof Mogren and Maria Sandsten. Thank you for creating an academic environment where I feel safe to explore research questions that interest me without the fear of failure, and for your encouragement to do so. Secondly, I would like to thank Tuomas Virtanen for allowing me to visit his research group, and for a very nice ongoing collaboration around topics that have ended up shaping this thesis.

I would also like to thank my colleagues Edvin Listo Zec and Martin Willbo for making our work place a fun and engaging place to be, and for always engaging in scientific discussions.

Finally, I would like to thank Lara. Thank you for always being there for me, and showing your support even in times when you should be focusing on yourself and our future family. I am not very good at celebrating the small achievements in life, but you always show me how. I love you.

Funding

Funding for this research comes from the Swedish Foundation for Strategic Research (SSF; FID20-0028), Sweden's Innovation Agency (2023-01486), and the AI Center at RISE Research Institutes of Sweden.

Popular summary

Machine learning is at the core of the recent success of artificial intelligence. A machine learning model is a form of computer algorithm that learns from data. The most common and reliable way to get the model to learn is by providing supervision. This is done by feeding the model with input data, say an audio recording, and then telling the model how to describe what is in that audio recording. A description of an audio recording given by a model is called a prediction, and a description given by a human that tells the model what to predict is called a label.

As an example, we can feed an audio recording of a bird singing into such a model and then tell it to predict that there is bird singing in the recording by providing an audio recording with an appropriate label. We can do this many times with audio recordings of different animal vocalizations such as dogs barking, whales calling, or birds singing. Eventually the model will learn to predict what is in a given audio recording.

These predictions can then be used to automatically analyze large amounts of recorded audio data to gain new scientific insights and establish policies. As an example, we could detect the vocalizations of different animal species in an acoustically monitored habitat to better understand the biodiversity in that habitat, and to establish policies towards maintaining or improving the biodiversity.

At the core of supervised learning is the human description of the data, the label. If the label is of low quality, for example by indicating that a bird is singing in a part of the audio recording where it is not, then the predictions from the trained model will be of low quality.

What we explore in this research are ways to improve supervised training of machine learning models by helping the human annotator to provide labels of higher quality. The goal is higher quality predictions at a reduced annotation cost. Resulting in higher quality scientific insights and policies, at a reduced cost.

Contents

Introduction

My interest in using machine learning to detect animal vocalizations in audio recordings goes back to a hiking trip in the Himalayas in 2016. During the month-long hike in the mountains my uncle would tell me the names of some different species of birds that were singing along the trail. I had recently come into contact with neural networks, and started to develop the idea that this should be possible to do with a computer. This idea never left me, and when I came home I discovered that there is a whole field with researchers doing just this called *bioacoustics*. This became the topic of my master's thesis, where I classified bird song in audio recordings using convolutional neural networks. And five years later I started my doctoral studies on the topic of machine learning for audio analysis, often called *machine listening*.

At the core of machine listening is the detection and classification of sound events. Sound events are distinct sounds that we can identify and recall based on their descriptions. These events, like "bird singing" or "dogs barking", form the core elements of a sound scene, helping us to interpret the surrounding environment. First we need to notice the onset and offset of the sound event (detection), and then we need to describe the type of event that has occurred (classification). Textual descriptions of these sound events are typically short, capturing the essence of what we hear. When a human is performing the detection and classification we call it annotation, and in this thesis we will call the textual description a class label, and the onset and offset a segment label. Annotated sound events are necessary to train and evaluate machine listening models that perform sound event detection. The speed of annotation can vary depending on the annotation task, and the annotations are inherently subjective, influenced by the annotator's personal experiences and perceptions [1].

What I have learned while developing methods and by discussing and collaborating with ecologists from around the world, is that large scale audio datasets are scarce. Datasets with annotations of thousands of animal sound events are rarely available when these projects start. They often have very few, if any, labeled sound events, and a huge need to annotate months or years of unlabeled audio recordings. What I want to explore in this thesis is therefore ways to make annotation easier, and how to improve the quality of the labels

collected during annotation. Further, I explore ways to make the best use the few initial labels that we may have.

In Paper A, we develop a theory for the label quality and annotation cost of a commonly used labeling method, where the annotator is limited to assigning class labels to fixed length audio data segments, called FIX weak labeling.

In Paper B, we propose a weak labeling method where the annotator assigns class labels to data segments that are adapted to cover the local structures (sound events) of interest. This can save annotation cost by requiring the annotator to give class labels to fewer data segments, and can also make the labels more precise by adapting the data segments to the structures of interest.

In Paper C, we propose a method to learn the time-frequency resolution in the commonly used log-Mel spectrogram as a part of the neural network training process.

In Paper D, we propose a robust method for bioacoustic sound event detection which can learn from only five annotated sound events.

While the motivation for this research mainly comes out of the need for cost efficient and high quality annotations for audio data, a lot of the research may be applicable to other types of data as well. The theory developed, and the methods proposed can in principle be applied to any time series data, and could possibly be extended to annotation of data in 2 or 3 dimensions as well (e.g., images or point clouds). I will therefore talk more broadly about annotation of data with local structures, such as sound events, in this thesis (see chapter 1).

The thesis is divided into three main chapters. First, chapter 1 gives an introduction to annotation of local structures in data and explains in more detail what we mean by a local structure. This chapter puts Paper A and Paper B in perspective. Second, chapter 2 gives an introduction to machine guided annotation of local structures in data, and puts Paper B, Paper C and Paper D in perspective. Finally, chapter 3 conclude the thesis, and discuss interesting future research directions.

Chapter 1

Annotating local structures in data

In this chapter, I will introduce the concept of a local structure in data, why we want to annotate local structures, and what it means to do so. I will then describe a common method for annotating local structures without explicitly asking the annotator to describe the boundaries of the structures. We will compare this method to an oracle method that uses the knowledge of the true boundaries and quantify the gap between them.

The oracle method can be seen as an upper bound on what may be achievable if we were to use the properties of the local structure during the annotation process. Finally, I will discuss some of these properties, and how we may be able to use them to design more precise annotation methods, which will lead us into the next chapter.

1 What is a local structure?

A local structure is a local part of the data that a group of people have given a textual description to. In figure 1.1 a local structure (green) is illustrated for audio (left) and image data (right). In the audio example, the local structure is the sound event (green) associated with the textual description ”bird song”. In the image example, the local structure is the set of pixels that make up the visual object (green) associated with the textual description ”bird”. The other parts of the data (gray) illustrate the other things happening in the background.

An important property of a local structure is that it occurs locally. In the audio example the sound event occurs locally along the time dimension, and in the image example the visual object occurs locally along the spatial dimensions. For audio, the local occurrence is typically associated with some form of temporal coherence where dependencies between previous and future sound samples are strong.

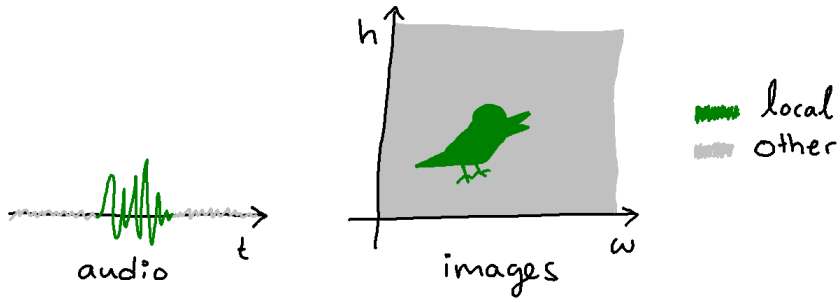

Figure 1.1: An example of a local structure in audio and image data. The audio data is a recording of a bird singing, and the image data is a picture of a bird. The local structures are shown in green. In the audio recording it is the sound event of the bird singing, and in the image data it is the visual object of the bird.

2 Why and how do we annotate local structures?

In supervised learning we train the model to predict the labels given by the annotator, and during evaluation we choose the model that best predict the labels. The prediction quality of the model will therefore necessarily depend on the quality of the annotations. We have seen this in sound event detection, where precise labels for the sound events lead to better performing models [2].

In extent, the (scientific) insights that can be drawn from the model predictions depend on the quality of the model. As an example, in ecology a researcher may be interested in counting the number of animal vocalizations from a certain species in an audio recording. The number of vocalizations, used together with a model of vocalization frequency, can then be used to estimate the number of individuals in a recording [3]. The accuracy of this count will depend on the quality of the annotations.

For evaluation data, a higher label quality means a better specification of what the best model should do, which is always desirable. E.g., if the goal is a model that produces well detected onsets and offsets of sound events, then the labels of the evaluation data need to reflect this. However, when training machine learning models, label noise can act as a form of regularization during training, meaning that sometimes noisy training labels can actually result in a model that generalize better to the evaluation data. Despite this, I will argue that anything that can be achieved with noisy training labels can also be achieved with noise free training labels by simply adding the noise afterwards. The opposite is not true. From this perspective, less noisy labels are strictly better also for the training data.

This is why we want to annotate the local structure; a precise local structure annotation gives a more accurate description of the data that help us develop better machine learning models.

Annotation of a local structure require us to describe the boundaries of the local structure, and to give a textual description of the structure within those boundaries. We therefore consider two label categories: the *segment label* (boundaries) and the *class label* (textual description). The segment label describes the boundaries of a data segment, and the class label is the textual description that we attribute to that data segment.

Labeling data with local structures therefore consists of constructing a set of segment labels and their corresponding class labels. The set of segment labels should partition the data into disjoint segments that cover all of the data, meaning that every part of the data is associated with exactly one class label. In figure 1.2 we show two different sets of segment labels leading to a correctly (top) and incorrectly (bottom) labeled local structure in an audio recording.

In the top image of figure 1.2 we can see that assigning the correct class label to each of the three segments will result in a correctly labeled local structure in the audio recording. In the bottom image, however, we can see that it is impossible to assign a correct class label to the second segment since it covers two different classes of the data, the local structure (green) and background sounds (gray). Further, we can see that three is the minimum number of segments needed to correctly label the audio data in this case, since if we have any fewer we will necessarily have to cover both the green and gray part with one of them. However, there are many ways, using more segments, that would also result in correct labels.

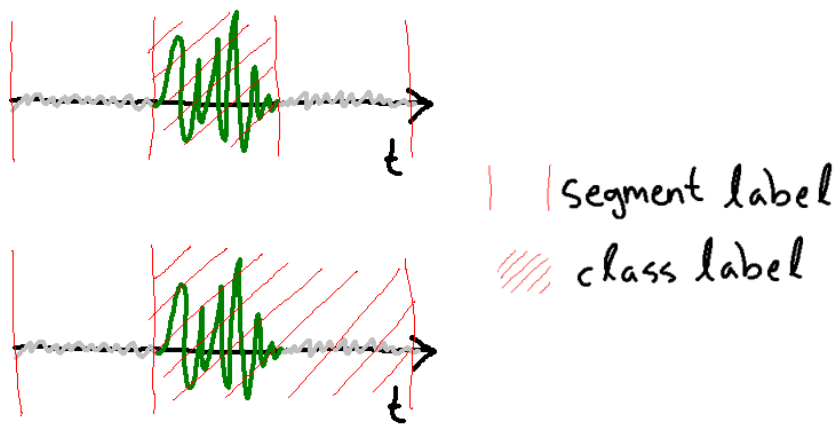

Figure 1.2: An example of segment labels and class labels of a local structure in audio data. The local structure is shown in green. The top image shows a way to partition the audio data into correct segment labels. With correct segment labels we can attribute a class label to each segment without introducing any label errors. In this example the class label indicates the presence of the local structure. The bottom image shows a way to create incorrect segment labels, in this case we will necessarily introduce noise into the labels by assigning a class label to each segment.

We will refer to the data point that we want to annotate, e.g., an audio recording in this example, as x and the label of x as y. The label $y = (s, c)$ consists of a set of segment labels s and a set of corresponding class labels c. We are interested in understanding and reducing

the label noise introduced by incorrect segment labels in this thesis, called *segment label noise*. A segment label is incorrect if it covers data from multiple classes, because then there is no correct class label for that segment. Segment label noise has been shown to lead to decreased performance of sound event tagging models [4, 5], where the goal is to detect if a sound event occurs in a given audio recording.

Another type of label noise occurs when an annotator assigns the incorrect class label for a given data segment, we call this *class label noise*. We do not model class label noise in Paper A, but we do study the effect of it in Paper B. A common way to reduce class label noise is to form a consensus on the class label by asking multiple annotators to label the same data segment [6, 7, 8].

3 Weak labeling of local structures

Annotating local structures is a demanding task that requires the annotator to detail the boundaries for the segment labels and assign the correct class label to each of these segments. The segment labels given by annotators are often inconsistent [2], partly because the interpretation of what constitutes the boundary of a local structure is subjective [9], leading to segment label noise. In addition, annotation of segment labels is demanding and takes more time, increasing the cost of annotation, and if the annotator is not an expert they may misunderstand the annotation task if it is too complex [8]. All these challenges associated with getting segment labels of high quality from annotators has created a need for methods that do not explicitly ask the annotator for the segment labels.

We therefore consider the setting where we only ask the annotator for a class labels of a given data segment, called weak labeling. This means that the segment labels need to be automatically constructed. The automatic construction of labels is often called pseudo-labeling [10]. In this thesis, we are interested in understanding the segment label noise resulting from the weak labeling process, which is a form of pseudo-labeling, and we propose ways to reduce this noise to get more precise annotation of the data.

3.1 FIX and ORC weak labeling

A commonly used weak labeling method is to partition the data into fixed and equal length segments, we will call this FIX weak labeling. The annotator is then asked to provide class labels for the data within each segment, and the corresponding segment label is inferred from the boundaries of the segment. The FIX weak labeling method is illustrated in figure 1.3. For audio data, this means that the annotator assigns class labels to equal sized audio segments. For images, the class labels would be assigned to rectangle segments, and

so on.

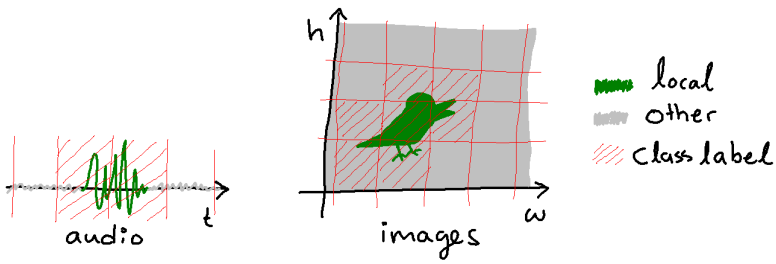

Figure 1.3: The FIX weak labeling method for audio (left) and images (right). The data is partitioned into equally fixed sized segments, and the annotator is asked to assign a class label to these segments. In these examples we can see how FIX introduces false positives, the parts of the class label that do not overlap with the local structure.

Variations of FIX weak labeling have been used to collect many of the large scale audio datasets that exist today. Annotation of AudioSet [11], which is still one of the largest and most used audio datasets today, was done by selecting a subset of 10 second segments to be weakly annotated. A more recently collected dataset called MAESTRO Real [8] was annotated by asking for class labels of 10 second segments with 9 second overlap to reduce segment label noise, and by asking five annotators to annotate each segment to reduce class label noise. They end up with multiple class labels for each part of the audio data and perform a weighed majority vote, based on an estimate of annotator competence, to get a single label for each part of the data.

The segment size as well as the overlap has an effect on the segment and class label noise. Too small segments may result in the annotator missing the presence of a local structure, which can introduce class label noise, but too large segments will introduce segment label noise. In general, smaller segments and larger overlap also mean that the annotator has to assign more class labels which increase the annotation cost. Which segment size and overlap to choose therefore depends on assumptions of the annotators' ability to detect the local structures. For some local structures the annotator may have to observe the whole structure to give a correct class label.

In Paper A we develop a theory for the limits of FIX weak labeling in 1 dimension (e.g., audio). We restrict ourselves to the setting where there is no overlap between the segments, and study the effect that the annotator model has on the resulting segment label noise for varying segment sizes. We introduce a metric called query intersection over union (QIoU), and an intuitive way to think of this metric is that $1 - \text{QIoU}$ roughly correspond to the segment label noise for a given segment. Using this we develop an expression for the segment size that will minimize the segment label noise in expectation for a given annotator model and data distribution.

We compare to an oracle (ORC) weak labeling method, which should be considered as

an upper bound that we can not know in practice. The ORC weak labeling method is illustrated in figure 1.4. It asks the annotator to assign a class label to the ground truth local structures. By construction this method will never introduce segment label noise, it will also always ask the annotator for the fewest possible number of class labels (three in the audio example, and two in the image example).

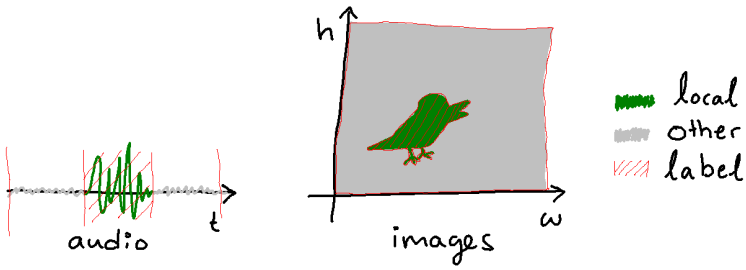

Figure 1.4: The ORC weak labeling method for audio (left) and images (right). There are no false positives, and we ask for the fewest number of class labels, 3 in the audio case, and 2 in the image case.

It may seem counterintuitive to call this ORC *weak* labeling, since this method gets labels that are of equal quality as the best strong labeling method where the class labels and segment labels are given by an oracle annotator. However, it is a weak labeling method in the sense that it only asks the annotator for class labels and never for segment labels. Thus acting as an upper bound on weak labeling, which we can not know, but that we can try to estimate in practice.

This highlights the potential of modelling the ORC weak labeling method by exploiting properties of the local structures, such as local similarities, dependencies and coherence, to automatically construct the segment labels, which is what we do in Paper B. In Paper B we use these properties to create segments that are adapted to the sound events of interest, and show that this can lead to higher quality labels at a reduced annotation cost. This is called machine guided annotation, which is the topic of the next chapter.

Chapter 2

Machine guided annotation of local structures in data

"The real problem is what can man and machine do together and not in competition."— Richard W. Hamming

In this chapter we will look at machine guided ways to reduce the annotation cost of local structures in data. We will introduce the annotation loop and the different ways to reduce annotation cost. Then we will connect Paper B-D to these, and finally discuss some differences of our proposed method and other methods.

1 The data annotation loop

We consider the setting where the annotator can only provide class labels for given data segments. The boundaries of the local structures therefore need to be automatically estimated, and then the annotator is asked to attribute a class label to the data segment contained within the boundaries.

Let us start with the generic annotation loop given in Algorithm 1. The algorithm takes as input a set of unlabeled data $\mathcal{D}_U = \{x_i\}_{i=1}^n$, a model M, and a performance criterion τ. The algorithm iteratively samples the next data point x to be annotated from the set of unlabeled data points, asks an annotator to label that data point, and then updates the model using the new annotation. The new model performance τ_M is then evaluated, and we continue this annotation loop until a satisfactory model performance τ has been reached. The output of the algorithm is the set of m labeled data points $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^m$, that

results in a model M of satisfactory performance $\tau_M \geq \tau$. The set of unlabeled data points $\mathcal{D}_U$ that we want to annotate can for example be a set of audio recordings, or a set of images such as those illustrated in figure 1.1 in the previous chapter, and the goal is a label y for each data point x that describes the local structures in it.

Algorithm 1 Annotation of data

1: **Input:** Unlabeled data $\mathcal{D}_U = \{x_i\}_{i=1}^n$, model M, performance criterion τ
2: **Output:** Labeled data $\mathcal{D}_L$
3: $\mathcal{D}_L \leftarrow \emptyset$
4: $\tau_M \leftarrow$ evaluate model M
5: **while** $\tau_M < \tau$ **do**
6: sample and remove x from unlabeled data $\mathcal{D}_U$
7: annotator gives label y to x and adds (x, y) to labeled data $\mathcal{D}_L$
8: update model M using labeled data $\mathcal{D}_L$
9: evaluate the model to get performance τ_M
10: **end while**
11: **return** $\mathcal{D}_L$

We are interested in ways to reduce the total annotation cost to get a model M with performance $\tau_M \geq \tau$. There are many ways to achieve this goal, each focusing on a separate line in Algorithm 1.

Firstly, looking at line 6, we can sample the next data point x such that the gain in model performance is maximized when x is annotated (using knowledge of M). This is the goal of works in active learning [12, 13].

Secondly, looking at line 7, we can either reduce the annotation cost c associated with the annotator giving the label y to x, or we can improve the quality of the annotation y. A higher quality annotation should lead to a higher gain in model performance. In Paper B we propose a method to increase the label quality of y for a given x at a reduced annotation cost c. We do this by using the model M to guide the annotator towards a higher quality y.

Lastly, looking at line 8, we can design models M that gain more in performance from each update. In Paper D we propose a few-shot learning method which is designed to learn a lot from only a few annotated examples, and in Paper C we propose a differentiable log-Mel spectrogram (DMEL) that can be optimized jointly with the model M.

2 Increasing the label quality at a reduced annotation cost

Increasing the quality of the label y given by the annotator for data point x can be done by increasing the labeling capability of the annotator. This can be done, for example, by choosing an expert annotator, or increasing the annotators' ability to perform the task [14]. Both these are ways of changing the properties of the human annotator, which we will not consider here.

We can also guide the annotator during the annotation task in ways that facilitate higher quality. This can be done, for example, by providing better annotation interfaces [15], or by doing parts of the annotation work automatically [16, 17, 18]. Automating parts of the annotation work has the benefit that label quality can potentially be increased at the same time as the annotation cost is reduced. In Paper B we propose at a weak labeling strategy towards this end.

Let us consider the annotation cost associated with assigning a label y to a data point x (line 7 in Algorithm 1). As described in section 2, the label $y = (s, c)$ consists of a set of segment labels $s = \{s_1, \ldots, s_k\}$ that partition the data point x (e.g., an audio recording) into k disjoint data segments that completely cover x, and a set $c = \{c_1, \ldots, c_k\}$ of the k corresponding class labels given by the annotator. The partitioning of x into k disjoint segments need to be done automatically since we are restricted to only ask the annotator for class labels. The annotation cost can therefore be written as ck where c is the cost of assigning a class label to a data segment. If we need to annotate m data points to achieve model performance τ the total cost therefore becomes mkc. We can reduce this cost by reducing any of the three factors in the product. We will consider m and k in this thesis, as c is a property of the human annotator.

The number of needed annotations can be reduced if the quality of the labels is increased. The quality of the label y can be affected by class label noise and segment label noise (see section 2). Let $Q(x, y)$ be a measure of the quality of the label y given to x with respect to the true class labels and local structures. Let $\bar{Q} = \frac{1}{n} \sum_{(x,y)\in\mathcal{D}_L} Q(x, y)$ denote the average label quality of the annotated dataset $\mathcal{D}_L$ resulting from Algorithm 1.

In Paper B we propose a method that makes use of the model M to partition x into segments that are better adapted to the local structures of interest. We call this machine guided annotation, because the model M is used to guide the annotation towards higher quality segment labels. Further, the model is updated with each new annotation. Initially, this may lead to noisy segment labels, but as the model is updated this noise is reduced. We empirically show that this happens, and that it leads to a higher label quality on average for the same annotation budget k compared to other commonly used methods. An improved average label quality $\bar{Q}$ means that the number of annotation rounds m needed to reach model performance τ is effectively reduced.

Automating parts of the annotation work does come with certain risks. If the automatically constructed segment labels contain a lot of segment label noise, then we may end up with lower quality labels instead. We have not observed this to be a problem in Paper B, but it is important to be aware of this risk. However, a nice property of first constructing the segments and then asking the annotator to give them class labels is that if the segments are noisy, then the annotator can notice this and take appropriate actions.

There is a subtle difference in this setup to other recently proposed pseudo-labeling methods for time series data, where the weak labels are given *before* the pseudo-labeling. In [16, 18] the weak label is first collected for a given point in time and then propagated to cover the local structure according to a temporal coherence criterion, and in [17] the weak labels are used to train a machine learning model which then predicts the pseudo-labels for the local structure, and then another model is trained on the pseudo-labels.

In our setup, by performing the weak-labeling *after* the pseudo-labeling we make sure that the annotator looks at the pseudo-labels, giving a natural quality assurance to the labeling process.

3 Learning more from the annotations

Broadly speaking, this is the goal of most work in supervised machine learning. We want to develop models that learn well from annotated training data, meaning that they generalize to some annotated evaluation data. However, there are specialized research directions such as few-shot learning, where the goal is to learn well from very few training annotations [19, 20, 21, 22]. Few-shot learning methods are, however, typically not designed to scale with more annotations. So, there is a trade-off here, and which way to model the data depends on the budget you have for annotation. If the budget is very low you may consider few-shot learning methods such as the one explored in Paper D, and if the budget is reasonably large then you may consider more complex ways of modelling such as that explored in Paper C.

To realize the ideas in Paper B we need good ways of modelling audio data in general. In Paper D we look at ways to make the most use of a few annotations. We do this by using an event length adapted ensemble of prototypical neural networks [19]. The key idea in the paper is to choose embedding functions for the ensemble that have been trained for certain event lengths based on the event lengths of the few examples that we already have. In Paper C we propose a version of the log-Mel spectrogram where the window length of the underlying short-time Fourier transform can be optimized jointly with the neural network model. The window length defines the resolution in time and frequency of the log-Mel spectrogram, and optimizing this for the classification task at hand can lead to stronger models. The log-Mel spectrogram is a very commonly used input representation

for convolutional neural networks (CNNs) in audio.

4 The difference between active learning and active annotation

In principle, the active updating of the model with each new annotation proposed in Paper B fits into the general framework of *active learning*, which is the reason active learning is in the title of the paper. However, I now believe that it may be reasonable to distinguish between them, and would like to propose the term *active annotation* for machine guided annotation where the model is iteratively updated during the annotation loop.

In active learning, we typically consider the setting where the next data point x is sampled to maximize some uncertainty criteria of the model (line 6 in Algorithm 1 depends on M). The idea is that the data point x that the model is most uncertain about should be most informative to annotate next, and that by biasing the data sampling process in this way we can reduce the number of annotations needed to reach a satisfactory model performance. The label noise is assumed independent of the data point to annotate.

In active annotation, the annotator is guided by the model M during the annotation of a *given data point* x. This means that the label noise will depend on the data point to annotate.

That is, active learning is about changing the sampling of x using knowledge of M, active annotation is about changing the sampling of y given x using knowledge of M.

Chapter 3

Conclusions and future work

1 Conclusions

We have developed a theory for weak labeling of local structures in data measured along 1 dimension, such as time series or audio data, and studied the limits of an approach commonly used in practice. We have compared this to an oracle method that solves the weak labeling task optimally. Knowing the consequences of different choices when performing weak labeling is crucial to make sure that the resulting annotations are of sufficient quality. (Paper A)

The limits frame the weak labeling problem, and can be used to put current methods in context. Further, the developed theory may give insights into ways to develop improved weak labeling methods. Towards this end we have also developed a weak labeling method that aims to model the oracle method by using each new annotation to further improve the annotation quality through machine guidance. We have showed that this method of annotation results in a higher label quality on average on all the studied datasets and for all assumed annotator models. (Paper B)

We have also proposed a method to learn the resolution in time and frequency of the typically used log-Mel spectrogram in audio modelling. We have showed that learning the appropriate resolution for the task at hand as a part of model training can speed up the training process, and lead to better performing models. (Paper C)

Finally, we have explored a modelling method that only requires as few as five annotated local structures to perform well, and have proposed two ways of improving the robustness of that method towards problems where the local structures vary a lot in size. (Paper D)

2 Future work

The papers that have shaped this thesis the most are Paper A and Paper B, which are about annotation of data with local structures, and in particular sound data. There are many future research direction that could be explored on this topic.

2.1 FIX weak labeling in more than 1 dimension

In Paper A we derive a theory for FIX weak labeling of local structures that appear along 1 dimension of the data, such as in time series data. It would be interesting to extend this theory into D dimensions, or at least into 2 and 3 dimensions such as images or point clouds. We rarely annotate in more than 3 dimensions anyway.

The number of class label assignments needed should grow exponentially with D for the studied FIX weak labeling method, but linearly with the number of local structures in the data for ORC weak labeling, making the potential cost gain of adaptive methods higher in more dimensions. Of course, modelling the ORC weak labeling method will also become a harder task with more dimensions. Exploring what happens when more dimensions are introduced is a very interesting research direction.

2.2 Active learning and active annotation in combination

While active learning is about changing the sampling of the data point x using the model, active annotation is about changing the sampling of the label y *given* a data point x using the model.

Clearly there will be a tension between these two processes if they are used jointly.

The best model we can hope to learn is a perfect model of the annotator, meaning that a sample that is hard for the model should also be hard for the annotator to annotate, leading to more label noise. By studying active annotation and active learning jointly, we may gain new insights into this trade-off between label noise and hardness of sample and find that the best way to learn may not be to always be exposed to the hardest sample of the problem, but rather a reasonably hard sample. The question is: what is reasonably hard?

2.3 Model selection in the active annotation loop

The underlying model in Paper B is a prototypical neural network [19], which was developed to learn quickly from only a few annotations. Unfortunately, the way this is done also means

that the learning saturates rather quickly. We have seen this in experiments where the label quality on a held out test set (not part of the paper) saturated after annotation of around 20 to 50 audio recordings.

Note that the label quality after saturation is better than that of the methods we compare with, so it is still beneficial to use this approach. But, it would be even better if the label quality just kept increasing until we eventually learn to model the ORC weak labeling process. We will probably not reach this upper bound, but we should aim to.

A very interesting research direction would be to make the complexity of the model depend on the number of annotations available through some model selection criteria. For example, we have seen in other works on active few-shot learning [21] that a prototypical neural network can have better performance than a linear classifier applied on the same embeddings when only a few annotations are available, but that the linear classifier becomes better after a certain number of annotations. The question is when to make the switch from the simple model (prototypical neural network) to the more complex model (a linear model applied on the embeddings), and going further when to choose models with even higher capacity as we accumulate more annotations.

2.4 Other annotator models

In Paper A we derive the theory for an annotator model that can detect presence of local structures if a fraction $\gamma \in (0, 1]$ of the local structure is contained within a given segment. While this makes sense for some types of sound events, other assumptions may be more applicable for other types of data.

In general, a better understanding of these properties of human annotators in practice would be very beneficial, and empirical studies towards this end are encouraged.

2.5 Adaptive weak labeling of multiple classes

The adaptive weak labeling method proposed in Paper B is developed for presence or absence annotation of a certain sound event class of interest. That is, to annotate multiple classes we need to perform multiple binary annotation tasks. However, the underlying prototypical neural network should be fairly easy to extend to multiple sound event classes of interest to facilitate annotation of multiple classes in a single annotation pass. There are trade-offs between multi-pass binary annotation and single-pass multi-label annotation [23], and being able to choose between these would be beneficial.

References

[1] Annamaria Mesaros, Toni Heittola, and Dan Ellis. Datasets and evaluation. In Tuomas Virtanen, Mark Plumbley, and Dan Ellis, editors, *Computational Analysis of Sound Scenes and Events*, chapter 6, pages 147–179. Springer Cham, 2017.

[2] Shawn Hershey, Daniel P.W. Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R. Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 366–370, 2021. ISSN 15206149. doi: 10.1109/ICASSP39728.2021.9414579.

[3] Tiago A. Marques, Len Thomas, Stephen W. Martin, David K. Mellinger, Jessica A. Ward, David J. Moretti, Danielle Harris, and Peter L. Tyack. Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2):287–309, 2013. doi: https://doi.org/10.1111/brv.12001. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12001`.

[4] Ankit Shah, Anurag Kumar, Alexander G. Hauptmann, and Bhiksha Raj. A closer look at weak label learning for audio events, 2018. URL `https://arxiv.org/abs/1804.09288`.

[5] Nicolas Turpault, Romain Serizel, Emmanuel Vincent, Nicolas Turpault, Romain Serizel, and Emmanuel Vincent. Analysis of weak labels for sound event tagging. 2021. URL `https://hal.inria.fr/hal-03203692`.

[6] Irene Martin-Morato, Manu Harju, and Annamaria Mesaros. Crowdsourcing Strong Labels for Sound Event Detection. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 246–250, 2021. ISSN 19471629. doi: 10.1109/WASPAA52581.2021.9632761.

[7] Irene Martín-Morató, Manu Harju, Paul Ahokas, and Annamaria Mesaros. Training Sound Event Detection with Soft Labels from Crowdsourced Annotations. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2023. ISSN 15206149. doi: 10.1109/ICASSP49357.2023.10095504.

[8] Irene Martin-Morato and Annamaria Mesaros. Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31:902–914, 2023. ISSN 23299304. doi: 10.1109/TASLP.2022.3233468.

[9] Anurag Kumar and Bhiksha Raj. Deep cnn framework for audio event recognition using weakly labeled web data, 2022. URL `https://arxiv.org/abs/1707.02530`.

[10] Eric Arazo, DIego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. *Proceedings of the International Joint Conference on Neural Networks*, 2020. doi: 10.1109/IJCNN48605.2020.9207304.

[11] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 776–780, 2017. ISSN 15206149. doi: 10.1109/ICASSP.2017.7952261.

[12] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[13] Steve Hanneke. Theory of Active Learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

[14] Meng Fang, Xingquan Zhu, Bin Li, Wei Ding, and Xindong Wu. Self-Taught Active Learning from crowds. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 858–863, 2012. ISSN 15504786. doi: 10.1109/ICDM.2012.64.

[15] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):–246, 2017. ISSN 25730142. doi: 10.1145/3134664.

[16] Yooju Shin, Susik Yoon, Sundong Kim, Hwanjun Song, Jae Gil Lee, and Byung Suk Lee. Coherence-Based Label Propagation Over Time Series for Accelerated Active Learning. *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[17] Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, and Yujun Wang. Pseudo Strong Labels for Large Scale Weakly Supervised Audio Tagging. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2022-May:336–340, 2022. ISSN 15206149. doi: 10.1109/ICASSP43922.2022.9746431.

[18] Di Chen, Xin-Yi Li, Ang Li, and Yu-Bin Yang. Representation-Based Time Series Label Propagation for Active Learning. *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1154–1159, 2023. doi: 10.1109/cscwd57460.2023.10152835.

[19] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, pages 4078–4088, 2017. ISSN 10495258.

[20] Yu Wang, Justin Salamon, Nicholas J. Bryan, and Juan Pablo Bello. Few-shot sound event detection. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020-May:81–85, 2020. ISSN 15206149. doi: 10.1109/ICASSP40776.2020.9054708.

[21] Yu Wang, Mark Cartwright, and Juan Pablo Bello. Active Few-Shot Learning for Sound Event Detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1551–1555, 2022. ISSN 19909772. doi: 10.21437/Interspeech.2022-10907.

[22] Ines Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana Strandburg-Peshkin, Ester Vidaña-Vila, Lisa Gill, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, Frants H. Jensen, Joe Morford, Michael G. Emmerson, Elisabetta Versace, Emily Grout, Haohe Liu, Burooj Ghani, and Dan Stowell. Learning to detect an animal sound from five examples. *Ecological Informatics*, 77(May), 2023. ISSN 15749541. doi: 10.1016/j.ecoinf.2023.102258.

[23] Mark Cartwright, Graham Dove, Ana Elisa Méndez Méndez, Juan P. Bello, and Oded Nov. Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–11, 2019. doi: 10.1145/3290605.3300522.

Scientific publications

Author contributions

Co-authors are abbreviated as follows: Maria Sandsten (MS), Olof Mogren (OM), Tuomas Virtanen (TV), Martin Willbo (MW), and Aleksis Pirinen (AP).

Note that ideas for papers almost never come from one person alone, they are mostly a collaborative effort even when we do not realize. Therefore, when I write that I came up with the idea for a paper, I mean that I rather independently thought about the problem to be solved and came up with the main parts of the idea for the methods to try and the experiments and simulations to be done. If a person is a co-author, they probably contributed to the idea of the paper in one way or another.

Paper A: Modelling the annotation quality and cost of weak labeling of fixed length segments in audio data

I came up with the idea for the paper, constructed the proofs to derive the theoretical results, implemented the simulation experiments, and wrote the manuscript (including figures and tables). All co-authors (MS, OM and TV) have provided valuable feedback during the development of the manuscript, including feedback on presentation of proofs and the design of the simulations.

Paper B: From weak to strong sound event labels using adaptive change-point detection and active learning

I came up with the idea for the paper, and developed the idea together with TV during a research visit at his group. Further, I have conducted all the simulations and experiments for the paper under weekly supervision of TV, and I have written the paper (including

figures and tables). All co-authors (MS, OM and TV) have provided valuable feedback during the development of the method and the paper.

Paper C: DMEL: the differentiable log-Mel spectrogram as a trainable layer in neural networks

MS and I came up with the idea for the paper. I have conducted all the simulations and experiments, and I have written the majority of the paper including tables and figures. MS have developed the theory section in the paper, and provided very valuable guidance and insights during the development of the method.

Paper D: Few-shot bioacoustic event detection using an event-length adapted ensemble of prototypical networks

The ideas in the paper were a collaborative effort between all co-authors. I implemented the ideas, performed the training and evaluation of the model, derived all results, and wrote the most of the paper. All co-authors (OM, MS, MW, and AP) have provided feedback during the development of the method and paper.