Thesis

Kilian Callebaut

April 29, 2021

Contents

Chapter 1

Introduction & Background

1.1 Audio Sensing

There is a growing and varied amount of labelled audio datasets available, which has made it possible for a large amount of audio classes to be recognised, for different purposes. Audio sensing tasks usually have a sense-transform-classify organisation with the purpose of categorising activity and behaviour Georgiev [2017]. Tasks are performed by taking a labelled audio fragment, extracting features and then run a classification algorithm like SVM, GMM or deep neural networks like CNN, RNN, CRNN or DNN. The tasks aim to correctly identify sound events (e.g. a word, the sound of an object, etc) or contexts (e.g. the scene where the audio was captured).

In order to reliably perform audio sensing tasks, complex and expensive networks are usually built, to perform the specific task. However, there are problems that arise with this, in certain contexts. For one, focusing on single tasks may ignore valuable discriminative information which can improve the task, in terms of correct detection of its target classes. Another one is that performing a number of tasks in parallel is expensive both on the memory and computational performance of a device.

1.2 Multi-task learning

Multitask learning has progressively seen more interest in the past few years in the field of acoustic analysis tasks. This interest has seen two main sides: performance improvement and resource efficiency. While the prevailing ap-

proach is still to build a single model addressing multiple classes at once, recent work has seen success when learning a shared representation between two tasks leading to a more robust representation of the signal. On the resource efficiency side, Multi-task learning has been proposed and used as a way to combine multiple different analysis tasks with little increase in resource usage, while maintaining performance.

In work done by (2017 Georgiev et al. Low-resource Multi-task Audio Sensing for Mobile and Embedded Devices via Shared Deep Neural Network Representations), diverse tasks have been combined in an adaptive way, optimizing for defined resource efficiency criteria. They found that emotion recognition, ambient scene analysis, keyword spotting and speaker verification could successfully be combined outperforming the single task models sometimes even when all tasks are combined with shared layers. However, they did not report on which tasks improved each other, the change in individual class performances or try to define why their set-up works.

1.3 Research gap

This signifies a wider lack of investigation in MTL for audio sensing tasks. When MTL works is an active research question in Multi-task learning. Some work has been done to try to answer this in NLP in work done by (2017 Martinez Alonso and Plank. When is multitask learning effective? Semantic sequence prediction under varying data conditions) and (2017 Bingel and Sogaard. Identifying beneficial task relations for multi-task learning in deep neural networks). In these works, the focus is mainly in finding out why, which and when one task improves another, for performance improvement. However in audio sensing, the fact should be pointed out again that there is significant interest in combining tasks for efficiency purposes.

Table 1.1: Tried Combinations

Title	Tasks	Classifier
Lu et al. [2004]	Automatic Speech Recognition; Speech Enhancement; Gender Detection	RNN
Panchapagesan et al. [2016]	Keyword Spotting; Large Vocabulary Continuous Speech Recognition Senones Targets Recognition	
Sakti et al. [2016]	Automatic Speech Recognition; Acoustic Event Detection	DNN
Georgiev [2017] Georgiev et al. [2017]	Speaker Identification; Emotion Detection; Stress Detection; Acoustic Scene Classification	DNN
Kim et al. [2017]	Emotion Detection; **Auxiliary tasks:** Arousal Level; Valence Level; Gender Detection	CNN
Nwe et al. [2017]	Acoustic Scene Classification (Grouped scenes as different tasks)	
Sun et al. [2017]	Keyword Spotting; Large Vocabulary Continuous Speech Recognition Phone Targets Recognition	
Kremer et al. [2018]	Word Error Rate and Character-Level Automatic Speech Recognition	CNN
Morfi and Stowell [2018]	Audio Tagging; Event Activity Detection	DNN
Lee et al. [2019]	**Main Tasks:** Audio Tagging; Speaker Identification; Speech Command Recognition (Keyword Spotting); **Auxiliary Tasks:** Next-Step prediction; Noise Reduction; Upsampling	
López-Espejo et al. [2019]	Keyword Spotting; Own-voice/External Speaker Detection	
Pankajakshan et al. [2019]	Sound Activity Detection (Event Activity Detection); Sound Event Detection (Audio Tagging)	CRNN
Tonami et al. [2019]	Acoustic Event Detection; Acoustic Scene Classification	CRNN for AED, CNN for ASC

Table 1.2: Tried Combinations (Continued)

Title	Tasks	Classifier
Xia et al. [2019]	Acoustic Event Type Detection (Audio Tagging); Predict frame position information (Event Activity Detection)	CNN
Xu et al. [2019]	Acoustic Event Detection; Acoustic Scene Classification	
Zeng et al. [2019] (1)	Emotion Detection; Music/Speech Classification	DNN
Zeng et al. [2019] (2)	Accent Recognition; Speaker Identification	DNN
Abrol and Sharma [2020]	Fine and Coarse Labels Acoustic Scene Classification	DNN
Deshmukh et al. [2020]	Acoustic Event Detection; Reconstruct Time Frequency Representation of Audio	CNN
Fernando et al. [2020]	Acoustic Event Type Detection (Audio Tagging); Predict Frame Position Information (Event Activity Detection)	LSTM
Huang et al. [2020a]	Audio Tagging; Temporal Detection (Event Activity Detection)	CNN PT/PS model
Huang et al. [2020b]	Audio Tagging; Event Boundary Detection (Event Activity Detection)	CNN
Tagliasacchi et al. [2020]	Keyword Spotting; Speaker Identification; Language Identification; Music/Speech Classification; Bird Audio Detection; Urban Acoustic Scene Classification; Music Instrument Pitch Detection; Music Instrument Detection	CNN
Wu et al. [2020]	Keyword Spotting; Domain Prediction	

Chapter 2

Research goal

In this thesis different combinations of tasks are evaluated in a MTL setup. The example of (2017 Bingel and Sogaard. Identifying beneficial task relations for multi-task learning in deep neural networks) is followed in that this work will try to identify data characterstics and patterns both in single-task and dual-task learning that predict task interactions in MTL deep neural networks.

Table 2.1: Research questions

Question
When do Acoustic Classification tasks learned in a deep learning multi-task hard parameter sharing set-up improve each other?
Subquestions

Chapter 3

Method

3.1 Experiment 1

Research Question: When do unrelated Acoustic Classification tasks, learned in a deep learning multi-task, hard parameter sharing set-up, improve each other?
Research Method: Implement different multi-class acoustic classification tasks from different corpi and different domains. The classification model will be a DNN without any adjustments for any specific task, to generalize the observations for task interaction in a multi-task learning setting. Different measurements are taken from the datasets beforehand. Each task is run in a single task set up, with different measurements being kept from the learning process and results. The goal of the investigation is to assess whether improved performance for one of the tasks in a dual task learning set-up can be achieved, or to which degree these tasks can be combined without a significant performance decrease for efficiency purposes.

Task choice:

Main Tasks

- General Acoustic Event Detection: Acoustic event detection or Sound event detection refers to the task of locating and classifying Sound Events in an audio from real life environments. General purpose AED means no specific optim

- Acoustic Scene Classification: Acoustic Scene Classification involves the automatic detection of the environment in an audio stream. Both indoor as well as outdoor environments can be included, with the length duration of a scene being long.
- Speaker Identification is the task of identifying the person speaking in an audio clip.
- Keyword Spotting aims at detecting predefined keywords in audio streams.

These are the main tasks, chosen to represent a variety of audio sensing tasks in the research. The tasks are more often then not the primary focus in audio sensing research, and are more complex in nature, with a multi-class classification goal. The choice was made for this specific set of tasks for multiple reasons. First of all, Gurrin et al. identified the main interests for audio sensing tasks for lifelogging purposes as being the identification of activities, audio events, location, people and keywords in dialogue, each of which is represented as a main task here. Futhermore, in work done by Georgiev, a system is built which combined Acoustic Scene Classification, Speaker Identification, Emotion Recognition and Stress detection, finding that all could be combined in a single multi-task DNN set-up without significant performance decrease. In this work the same DNN structure will as it has shown to perform well on multiple audio sensing tasks as well as allow direct verification for its findings. Furthermore are all (main and auxiliary) tasks chosen so that they would span multiple domains (Speech, Environment and Music), as well as multiple different datasets.

Neural Network:

The architecture is based on the system, described by Georgiev [2017], which is a Deep Neural Network (DNN) feed-forward propagation classifier. As seen in figure 3.1, in a multi-task setting, all tasks share the same input and hidden layers, with task-specific output layers. This leads to these layers being used as a universal feature transformation which captures acoustic observations from different tasks. For the input, features have to be extracted from the audio. In the original work this is implemented by statistical summaries of log filter banks, extracted from each audio frame. They use summaries both for reducing the input feature complexity as well as allowing a

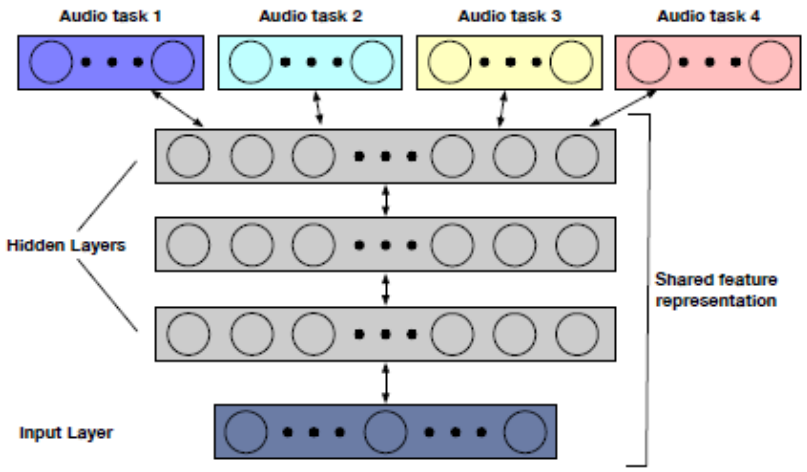

Figure 3.1: Multi-task DNN model

representation of the same size across multiple tasks.

To allow for comparison and analysis of the interaction between tasks, this work follows the approach from Alonso and Plank [2016] and Bingel and Søgaard [2017] the hyperparameters (this includes a fixed random seed set) and the architecture will be kept the same, regardless of which and the number of tasks learned. The set-up Georgiev [2017] had was 3 hidden layers which were tested at 128, 256 or 512 in each layer, with a soft-max pooling layer at the output layers. The output layers are updated separately depending on whether the training input instance is one for their own task. Furthermore, no task-specific features, nor any pre-trained embeddings will be used for initialisation to avoid any task bias.

The training is done by mini-batch Stochastic Gradient Descent (SGD) on all audio tasks simultaneously, which is the key to succesfully training the multi-task DNN [Georgiev, 2017]. The samples are randomised across tasks before being fed into the DNN system. Backpropagation is then used for the fine-tuning of the DNN, which again only trains the softmax layer for the task at hand with the other softmax layers being kept intact.

This set-up is chosen for the same reasons as Georgiev, namely due to

the fact that it proved to be successful in multiple audio analysis tasks, as well as providing a good comparison opportunity for analysing the results. As Bingel and Søgaard [2017] noted in their work as well however, the fixed hyper-parameters make the results only applicable to the scenario where one wants to know whether MTL works in the chosen parameter configuration.

Dataset measurements:
One way of structurally figuring out when multiple tasks can improve each other is described in Alonso and Plank [2016] and Bingel and Søgaard [2017], where they try to predict which dataset and task features correlate to improvements in the multi-task setting for NLP tasks. Bingel and Søgaard [2017] builds a regression analysis system for this to find the Pearson Coefficiënt between pre-measured data and the change in F1 score from single to multi-task systems. There are 3 groups of dataset measurements which will be examined: Dataset features, Information Theory derived features and Single task inference features.

Dataset features

- Number of labels
- Number of clips
- Number of labeled frames
- Type/token ratio
- Labels per domain
- Dataset size

Information Theory derived features

In Alonso and Plank [2016], these were found to be the most informative features. The interest is in determining whether these results are transferable to the audio domain.

- label entropy: Entropy of the label distribution
- label kurtosis: Indicates the skewness of a distribution

Learning curve features

In Bingel and Søgaard [2017] these were one of the most informative features. This involves taking measurements of the learning curve for the single tasks at different stages.

- Curve gradients: The gradients of the loss curve, taken at 10, 20, 30, 50 and 70 percent of the total number of batches
- Fitted log-curve: A logarithmic functio is fitted to the loss curve values, where the function is of the form L(i) = a*ln(c*i+d) + b where a and c are features that describe the steepness of the learning curve.

Audio features

Along with the features from the NLP research, additional audio related features are taken in account, which have been mentioned in the literature.

- Average sample length
- Total length of dataset (in seconds)
- Sampling rate
- Signal to Noise Ratio

Non-numetric Factors

Finally there are also non-numetric Factors that might have an impact.

- Overlapping
  Non-overlapping events
- Unlabeled, weakly labeled and strongly labeled dataset
- Mono or Stereo channel

Evaluation:

The evaluation will
For evaluation these are the general metrics

- Averaged F1 score: The event categories are considered equally important Phan et al. [2019]

- Overall F1 score: The event instances are considered equally important Phan et al. [2019]
- Accuracy: Used in contexts where false positives and false negatives are less important than the rate of true positives and true negatives, which is often the case in audio sensing tasks.
- Detection Error Rate (ER): Used for evaluating correct segmentation of events from a continuous test signal Phan et al. [2019]
- Area Under the Receiver Operating Characteristic Curve (ROC AUC): plots the true positive rate against the false positive rate, with the metric being calculated from the area under the plot. Random guesses have an AUC of 0.5 Deshmukh et al. [2020]
- Intermediate output variances: Used in Xia et al. [2019] to prove they can average respective noise patters
- Efficiency: Combining tasks in a multi-task setting is used both for performance as efficiency purposes. Two metrics will be evaluated, following Georgiev [2017]. Both of these will be compared to running a single inference task and the combined time of running all combined tasks separately.
  - Runtime: The time needed for performing audio classification
  - Memory: The size of the trained multi-task DNN compared to the single task and combined single tasks variants.

Lastly, this work will take a look at class relationships in the context of multi-task. In the research, a few observations have already been made, in regards to the effect of the combination of tasks on certain classes. Specifically in this experiment, two effects will be specifically investigated.

- **Speech related tasks for AED**: (REFERENCE) multiple observations have been made that for a general AED classifier, the Speech class is often mislabeled and causes mislabels of events like coughs. (CITE) notes that this is due to the fact that speech signals and environmental audio signals have a different structure. Therefore, this experiment especially interested in whether and/or which tasks in the speech domain can improve the speech class detection in AED.

- **AED events only happening in certain scenes for ASC**: (REFERENCE) has noticed, when combining AED and ASC that AED can improve ASC, especially when certain events only happen in specific scenes. This was also found to go in the other direction, namely that events happening in multiple scenes will lead to a worse performance in ASC. This means the overview will be made of which events are in which scenes and also which ones are unique.

3.2 Experiment 2

Research Question: What measurements can predict success in the multi-task set-up?

Research Method: Build a regression analysis system to compute the pearson correlation between the dataset and single task measurements and the multi-task performance shift. The goal of the investigation is to assess whether improved performance for one of the tasks in a dual task learning set-up can be predicted from the dataset and single inference task measurements through regression analysis.

Dataset measurements:
The dataset and single task measurements are described in Experiment 1 (except for the non-numetric factors). The same structure as Bingel and Søgaard [2017] will be used for the regression as it is a good evaluation template, as well as making it possible to compare whether similar measurements are good predictors between NLP and Audio tasks.

Regression Analysis

Evaluation:

Any consistency in measurements that are good predictors for multi-task success between tasks will be looked for. The logistic regression will take the binarized results of the micro-averaged F1 scores from Experiment 1 (and the results from subsequent experiments) and be used to predict the benefits or detriments of MTL setups based on the computed features. An observation from the NLP research to verify is that multi-task gains seemed

to be more likely for main tasks that quickly plateau with non-plateauing tasks, which might hold for audio tasks. Bingel and Søgaard [2017] uses the mean performance of 100 runs of randomized five-fold cross-validation.

- Pearson coefficient: Measure for correlation between two variables. In this case, this will be the correlation of the numeric measurements from the dataset. For this a regression analysis system will be built.

3.3 Experiment 3

Research Question: Can auxiliary tasks be added to the multi-task set-up, to improve classification performance of (certain) main tasks?

Research Method: Implement different auxiliary tasks specifically aimed at improving performance of one or more main tasks. The classification model will be the same DNN without any extra adjustments. The auxiliary tasks are chosen from the multi-task audio sensing literature, where they have proven to show positive results when combined with one of the main tasks. The aim is to find whether and which results are somewhat replicable in the current set-up, as well as again figuring out which factors can be related to positive improvements. All auxiliary tasks will be treated as main tasks in this experiment, in the sense that they will also be tested in combination with each other without any main task necessarily being present.

Task choice:

Auxiliary Tasks

- Music/Speech Detection
- Voice Activity Detection
- Emotion Detection
- Gender Detection
- Audio Tagging

- Event Activity Detection
- Stress Detection
- Music Instrument detection

The previous mentioned main tasks are supplemented by a list of auxiliary tasks, which are generally simpler tasks that are not often the main research goal but added as a side task to explicitly improve the main task. They often can be performed on the same dataset as the main task. Each task is treated as a main task and combined with all the other tasks. These are still mentioned as separate, as they are explicitly chosen for a specific relationship with at least one of the main tasks. Some tasks - e.g. Emotion detection - are also only considered auxiliary tasks in this context therefore.

The reason for each inclusion will be disclosed next. There are three main categories the reasons of inclusion fall in, which relation to the main tasks this work aims to investigate: Simple differentiation, semantic detection in a specific domain and different specificity level.

- With simple differentiation, the Music/Speech detection, Voice Activity Detection and Event Activity Detection tasks fall under this. These tasks simply try to differentiate if activity of a certain type is or isn't happening at a certain point. The hypothesis is that simple differentiation helps to build a representation that would make less errors for tasks where wrongfully differentiating the target labels of these tasks leads to prediction errors in its own results.
  - Music/Speech Detection: The task of classifying audio in a stream between Music, Speech or neither. This is included as it learns to differentiate between different domains which might be beneficial for all (main) tasks
  - Voice Activity Detection: The task of detecting voice activity in a stream. This learns a representation that can differentiate specifically between what sound is or is not coming from a voice, which might lower the number of classification errors happening in SI and KS when the sound is not a voice, as well as AED for discriminating the speech label.

  - Event Activity Detection: The task of detecting whether or not a sound event is happening at a certain point. In the literature (REFERENCE) AED has often been defined as a multi-task problem, by splitting the tasks of detecting an event and determining the type of event (Audio Tagging) in two, while still learning them simultaneously and combining the results afterwards. This has lead to better performance to baseline single task AED systems. This inclusion might be beneficial for lowering detection errors when no (target) sound event is present AED, SI and KS.

- Next is semantic detection, which here means the Emotion detection, stress detection, gender detection in the speech domain and music instrument detection in the music domain. These try to identify a certain property of events in the audio. The idea of including this is to check whether identifying more complex properties of speech and music having a less direct link to the purposes of the main tasks can be beneficial for performance.

  - Emotion Detection: The task of detecting emotion from speech and music. This was included in the work done by Georgiev as well as others. Emotion Detection is not expected to have any direct link to any of the main tasks, except perhaps a loose one with the speech domain tasks.
  - Gender Detection: The task of detecting gender of a speaker. This is directly linked to speaker identification, as the same speaker will have the same gender.
  - Stress Detection: The task of detecting stress in speech audio. This was also included in the work done by Georgiev. Stress detection has a loose connection with the speech domain, but a direct one with Emotion Detection. It has a smaller target labelset size than Emotion Detection, in which we are interested due to previous observations that tasks with smaller target label sets make for better auxiliary tasks.
  - Music Instrument detection: The task of detecting which music instrument is played in an audio clip. This has a loose connection with the music domain, which might be interesting for better detection of music audio, leading to less wrongful labels for events that are out of target for the main tasks.

- The last one is different specificity level, under which only audio tagging falls directly, as it is a simpler version of AED (you only need to identify what is happening in the whole clip and not when it is or is not present). The first category can be seen as a more specific subset of this with a small target label set though. The interest for this class is different, in the sense that it is more about reporting the difference adding a simplified version of the same task makes as well as comparing its interaction with the other tasks to the specific version.

  - Audio Tagging: In AED research, improved results have been achieved by defining the task of AED as two separate tasks, namely Audio Tagging and Event Activity Detection, learned simultaneously in a multi-task framework. Besides this, work done by Huang et al. achieved the best results in the 2019 AED challenge by adding additional branches of audio tagging tasks (with different pooling methods) besides the main Audio Tagging/Event Detection branch. Therefore the interest lies in its comparison with the performance compared to single task AED as well as its interaction with other tasks in a multi-task framework.

Evaluation:

Aside from the Evaluation metrics described in the previous experiment, this will also include qualitative comparisons with results from the literature, including the difference in set-up, in order to evaluate how the difference in set-up affects multi-task performance.

3.4 Experiment 4

Research Question: What is the effect of the choice of dataset in auxiliary tasks?

Research Method: Implement and compare the performance of combining tasks when training on the same dataset as a main task, compared to using a different dataset.

Possible:

- AED & ASC
- Speaker Identification & Keyword detection (Kaggle2018)
- AED & Music/Speech Detection (verify)
- AED & Voice Activity Detection
- AED & Audio Tagging
- AED & Event Activity Detection
- (if AED and ASC contain the same classes in the combined subset, then everything above can be combined with ASC as well)
- Speaker Identification & Music/Speech Detection (verify)
- Speaker Identification & Voice Activity Detection (verify)
- Speaker Identification & Emotion Detection (verify)
- Speaker Identification & Gender Detection
- Speaker Identification & Stress Detection (verify)
- Keyword Detection & Music/Speech Detection (verify)
- Keyword Detection & Voice Activity Detection
- Keyword Detection & Emotion Detection (verify)
- Keyword Detection & Gender Detection
- Keyword Detection & Event Activity Detection
- Keyword Detection & Stress Detection (verify)

3.5 Experiment 5

Introduction In previous research, observations have been made that the multi-task setting is affected by the label overlap between classes. One example of this is in (REFERENCE), where they observe that, when combining AED and ASC, that the accuracy of the ASC task improved when an output label (i.e. a scene) did not share any output labels (i.e. sound events) with other scenes. When 2 scenes did share a number of sound events, the set-up had more difficulty differentiating the scenes correctly. In the same vein, it has been observed (REFERENCE) that including an Environmental Sound Classification task (i.e. AED task, but without speech-related labels) can improve speech domain tasks. This experiment will investigate the effect of keeping certain instances out of a task, compared to the case where the tasks have overlapping label relationships.

Research Question: How does class overlap affect combinations in the multi-task setting?

Research Method: Identify the task relationships in terms of overlap between labels and compare the results when training the network on a dataset where instances with these labels are left out.

Chapter 4

Schedule

The first experiment will be about implementing and evaluating the main tasks and their combination in a multi-task setting.

- DNN classifier done 15/1
- Main task implementation done (4) 24/1
- Task inference experiments done 31/1

The second experiment

- done 14/2

The third experiment

- done 21/3

The fourth experiment

- done 11/4

The fifth experiment

- done 2/5

Finishing report done 30/5
Done

Table 4.1: Datasets

Tasks	Datasets	Count	Available	Count
Acoustic Scene Classification	LITIS ROUEN; TUT Acoustic Scenes 2016	2	TUT Acoustic Scenes 2016	1
Acoustic Event Detection	TUT SED 2016 for overlapping AED; DCASE 2020 Task 4; TUT Sound Events 2016; TUT Sound Events 2017; AudioSet adaptation	5	TUT SED 2016 for Overlapping AED; DCASE 2020 Task 4; AudioSet Adaptation	3
Speaker Identification	Automatic Speaker Verification Spoofing and Counter measures challenge; LIBRISPEECH; Voice Cloning Toolkit; FSD Kaggle 2018; LibriAdapt	5	Automatic Speaker Verification Spoofing and Counter measures challenge; Librispeech; Voice Cloning Toolkit; FSD Kaggle 2018; LibriAdapt	5
Keyword detection	TIMIT continuous speech corpus; FSD Kaggle 2018	2	FSD Kaggle 2018	1
Music/Speech detection	Musan MUS dataset; RAVDESS	2	Musan MUS dataset; RAVDESS	2
Voice Activity Detection	AMI Meeting; NIST OPEN SAT	2	AMI Meeting;	1
Emotion Detection	Emotional Prosody Speech and Transcript Library; RAVDESS	2	RAVDESS	1
Gender Detection	LibriAdapt	1	LibriAdapt	1
Audio Tagging	DCASE 2020 Task 4; Urban SED; DCASE 2018 Task 4; FSD Kaggle 2018	4	DCASE 2020 Task 4; Urban SED; FSD Kaggle 2018	3

Bibliography

Vinayak Abrol and Pulkit Sharma. Learning hierarchy aware embedding from raw audio for acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1964–1973, 2020.

Héctor Martínez Alonso and Barbara Plank. When is multitask learning effective? semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251*, 2016.

Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*, 2017.

Soham Deshmukh, Bhiksha Raj, and Rita Singh. Multi-task learning for interpretable weakly labelled sound event detection. *arXiv preprint arXiv:2008.07085*, 2020.

Tharindu Fernando, Sridha Sridharan, Mitchell McLaren, Darshana Priyasad, Simon Denman, and Clinton Fookes. Temporarily-aware context modeling using generative adversarial networks for speech activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1159–1169, 2020.

Petko Georgiev. Heterogeneous resource mobile sensing: computational offloading, scheduling and algorithm optimisation. 2017.

Petko Georgiev, Sourav Bhattacharya, Nicholas D Lane, and Cecilia Mascolo. Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–19, 2017.

Yuxin Huang, Liwei Lin, Shuo Ma, Xiangdong Wang, Hong Liu, Yueliang Qian, Min Liu, and Kazushige Ouch. Guided multi-branch learning systems for dcase 2020 task 4. *arXiv preprint arXiv:2007.10638*, 2020a.

Yuxin Huang, Xiangdong Wang, Liwei Lin, Hong Liu, and Yueliang Qian. Multi-branch learning for weakly-labeled sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 641–645. IEEE, 2020b.

Nam Kyun Kim, Jiwon Lee, Hun Kyu Ha, Geon Woo Lee, Jung Hyuk Lee, and Hong Kook Kim. Speech emotion recognition based on multi-task learning using a convolutional neural network. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 704–707. IEEE, 2017.

Jan Kremer, Lasse Borgholt, and Lars Maaløe. On the inductive bias of word-character-level multi-task learning for speech recognition. *arXiv preprint arXiv:1812.02308*, 2018.

Tyler Lee, Ting Gong, Suchismita Padhy, Andrew Rouditchenko, and Anthony Ndirango. Label-efficient audio classification through multitask learning and self-supervision. *arXiv preprint arXiv:1910.12587*, 2019.

Iván López-Espejo, Zheng-Hua Tan, and Jesper Jensen. Keyword spotting for hearing assistive devices robust to external speakers. *arXiv preprint arXiv:1906.09417*, 2019.

Youyi Lu, Fei Lu, Siddharth Sehgal, Swati Gupta, Jingsheng Du, Chee Hong Tham, Phil Green, and Vincent Wan. Multitask learning in connectionist speech recognition. In *Proceedings of the Australian International Conference on Speech Science and Technology*. Citeseer, 2004.

Veronica Morfi and Dan Stowell. Deep learning for audio event detection and tagging on low-resource datasets. *Applied Sciences*, 8(8):1397, 2018.

Tin Lay Nwe, Tran Huy Dat, and Bin Ma. Convolutional neural network with multi-task learning scheme for acoustic scene classification. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1347–1350. IEEE, 2017.

Sankaran Panchapagesan, Ming Sun, Aparna Khare, Spyros Matsoukas, Arindam Mandal, Björn Hoffmeister, and Shiv Vitaladevuni. Multi-task learning and weighted cross-entropy for dnn-based keyword spotting. In *Interspeech*, volume 9, pages 760–764, 2016.

Arjun Pankajakshan, Helen L Bear, and Emmanouil Benetos. Polyphonic sound event and sound activity detection: A multi-task approach. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 323–327. IEEE, 2019.

Huy Phan, Oliver Y Chén, Philipp Koch, Lam Pham, Ian McLoughlin, Alfred Mertins, and Maarten De Vos. Unifying isolated and overlapping audio event detection with multi-label multi-task convolutional recurrent neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 51–55. IEEE, 2019.

Sakriani Sakti, Seiji Kawanishi, Graham Neubig, Koichiro Yoshino, and Satoshi Nakamura. Deep bottleneck features and sound-dependent i-vectors for simultaneous recognition of speech and environmental sounds. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 35–42. IEEE, 2016.

Ming Sun, David Snyder, Yixin Gao, Varun K Nagaraja, Mike Rodehorst, Sankaran Panchapagesan, Nikko Strom, Spyros Matsoukas, and Shiv Vitaladevuni. Compressed time delay neural network for small-footprint keyword spotting. In *Interspeech*, pages 3607–3611, 2017.

Marco Tagliasacchi, Félix de Chaumont Quitry, and Dominik Roblek. Multi-task adapters for on-device audio inference. *IEEE Signal Processing Letters*, 27:630–634, 2020.

Noriyuki Tonami, Keisuke Imoto, Masahiro Niitsuma, Ryosuke Yamanishi, and Yoichi Yamashita. Joint analysis of acoustic events and scenes based on multitask learning. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 338–342. IEEE, 2019.

Haiwei Wu, Yan Jia, Yuanfei Nie, and Ming Li. Domain aware training for far-field small-footprint keyword spotting. *arXiv preprint arXiv:2005.03633*, 2020.

Xianjun Xia, Roberto Togneri, Ferdous Sohel, Yuanjun Zhao, and Defeng Huang. Multi-task learning for acoustic event detection using event and frame position information. *IEEE Transactions on Multimedia*, 22(3):569–578, 2019.

Kuilong Xu, Shilei Huang, Gang Cheng, and Xiao Song. A multi-task learning approach based on convolutional neural network for acoustic scene classification. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 23–27, 2019.

Yuni Zeng, Hua Mao, Dezhong Peng, and Zhang Yi. Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 78(3): 3705–3722, 2019.