Thesis

Kilian Callebaut

May 29, 2021

Abstract

Swag

Contents

Chapter 1

Introduction

Multi-task learning (MTL) is a machine learning paradigm where multiple different tasks are learned at the same time, exploiting underlying task relationships, to arrive at a shared representation. While the principle goal was to improve generalization accuracy of a machine learning system [Caruana, 1997], over the years multitask learning has found other uses, including speed of learning, improved intelligibility of learned models [Caruana, 1997], classification fairness [Oneto et al., 2019] and as a means to compress multiple parallel models [Georgiev, 2017]. This led to the paradigm finding its usage in multiple fields, including audio recognition.

The field of audio recognition is varied and ever expanding, due to a growing number of large public and non-publicly available datasets (e.g. AudioSet [Gemmeke et al., 2017]) each with their own variations like sources, lengths and subjects. The tasks in the field can roughly be divided into three categories: Speech recognition tasks, Environmental Sound recognition tasks and Music recognition tasks, along with tasks that combine multiple domains [Duan et al., 2014]. These domains inherently have a different structure from each other, which requires different processing and classification schemes. Speech for example, is inherently built up out of elementary phonemes that are internally dependent, the tasks linked to which have to deal with the exact differentiation and characterization of these, to varying degrees. Environmental sounds in contrast, do not have such substructures and cover a larger range of frequencies. Music then has its own stationary patterns like melody and rhythm [BoreGowda, 2018]. A general purpose audio classification system, dealing with real life audio, would have to deal with

the presence of each of these types of audio though, regardless if its task is only in one of the domains.

Usually, in order to achieve high performance, it is necessary to construct a focused detector, which targets a few classes per task. Only focusing on one set of targets with a fitting dataset however, ignores the wealth of information available in other task-specific datasets, as well as failing to leverage the fact that they might be calculating the same features, especially in the lower levels of the architecture [Tagliasacchi et al., 2020]. This does not only entail a possible waste of information (and thus performance) but also entails a waste of computational resources, as each task might not require its own dedicated model to achieve the same level of performance. Originally conventional methods like Gaussian Mixture Models (GMM) and State Vector Machines (SVM) were the main focus, but due to the impressive results in visual tasks deep learning architectures have seen a lot of attention. The emergence of deep learning MTL set-ups is still fairly recent in audio recognition. While it has seen both successful [Tonami et al., 2019] applications and less successful [Sakti et al., 2016] when combining different tasks, very little is known about the exact circumstances when MTL works in audio recognition. In fact, this question has not been investigated often in MTL in general.

This thesis will aim to investigate the relationship of multiple audio tasks in a MTL set-up. Following previous work in investigating the MTL set-up in different tasks in the Natural Language Processing field [Alonso and Plank, 2016] [Bingel and Søgaard, 2017], this work will focus on experimenting with different MTL set-ups, where the multi-task runs are compared to the single task runs, evaluating the differences in results. The tasks examined will be unrelated, each having their own dedicated dataset. The goal is to identify connections between tasks in single-task learning that predict results in MTL.

The rest of this chapter will define the exact problem this work is addressing, then address this in specific research questions, declare the contributions of this project and finally outline the rest of the document.

1.1 Problem Statement

Tasks with dedicated datasets bring together

- Little research about when multi-task works

- power in audio comes from being able to harness the power multiple dedicated datasets
- Multiple different tasks in 1 classifier
- Question is both which/how many can you combine while maintaining predictive power and what their effect is on each other for further performance development
- Do these effects remain when switching classifier

While most audio classifier research focuses on building systems optimized on dedicated datasets, general purpose classification needs require multiple tasks to be performed at once. While it is possible to build a separate classifier per task, trained on their dedicated dataset, this entails that a lot of work is done in parallel, which perhaps should not be. Not only does this waste computational resources (which is important for on-device inference), but also ignores to leverage potential useful information in other tasks. Building a multi-head model with shared embedding layers addresses this problem, as the low-level features are learned simultaneous saving space and time as well as sharing information coming from every annotated training set. While this has seen some application [Tagliasacchi et al., 2020], [Georgiev, 2017], [Lee et al., 2019] proving the viability of this approach in audio, no in depth look has been taken in regards to the effects of this combinatory approach. The main questions that remain unanswered are when this works and what the effects of the joint learning set-up in audio classification are, beyond general performance evaluation.

1.2 Research Questions

In this thesis different combinations of tasks are evaluated in a MTL setup. The example of [Bingel and Søgaard, 2017] is followed in that this work will try to identify data characteristics and patterns both in single-task and dual-task learning that predict task interactions in MTL deep neural networks.

Table 1.1: Research questions

Question
When do Acoustic Classification tasks learned in a deep learning multi-task hard parameter sharing set-up improve each other?
Subquestions
Which unrelated Acoustic Classification tasks, learned in a deep learning multi-task, hard parameter sharing set-up, improve each other?
What measurements can predict success in the multi-task set-up?
Do the task relationships demonstrate similarities when changing the classifier?
Are the task combination effects consistent when combining further?

1.3 Contributions

1.4 Outline

Chapter 2

Literature Study

2.1 Methodology

The systematic search for different approaches happened on a combination of Google Scholar, ACM Digital Library and Web of Science. Using the following keywords and combinations as well as variations "Acoustic Event Detection", "Acoustic Scene Classification", "Multi-task Speech Classification", "Deep learning Audio", "Joint learning acoustic", "Multi-task audio classification", "Multi-task learning", "Multi-task relations", "When is Multi-task effective", a collection of papers was found which was expanded by referenced and referencing papers about Deep learning audio tasks in a multi-task framework, as well as general research on when multi-task learning is effective. The original focus was pure on AED, which expanded into its combinations with other, unrelated audio classification tasks. Special attention was given to papers coming from the DCASE audio classification competitions, a yearly event where a number of different audio classification goals are defined with both returning as well as varying tasks.

2.2 Deep Learning Audio classification

A growing amount of

2.3 Multi-task Learning

2.4 The Interest in Multi-task Learning in Audio Classification

Chapter 3

Experiment Design

3.1 Tasks

3.2 Data Properties

3.3

Chapter 4

Implementation

Chapter 5

Evaluation

Chapter 6

Discussion & Future Directions

6.1 Discussion

6.2 Future Directions

Chapter 7

Conclusion

Chapter 8

Journal

From my research, I saw the potential of multi-task learning in audio classification tasks. The original focus was on AED tasks, which had been appended with different tasks to varying degrees of success.

I started off following Georgiev, implementing the feature extraction as a summary of logmel features, with a DNN as a classifier. There was a discrepancy between train and test and bad results on e.g. AED. I reimplemented using a cnn and mel spectrogram features. Now I have to test it all. Testing is going slow. I can't figure out how to use the vm without uploading my data.

I also have to figure out a way to compare the data and get some useful conclusions out of it.

Bibliography

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.

HB BoreGowda. Environmental sound recognition: A survey. 2018.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Shufei Duan, Jinglan Zhang, Paul Roe, and Michael Towsey. A survey of tagging techniques for music, speech and environmental sound. *Artificial Intelligence Review*, 42(4):637–661, 2014.

Jort F Gemmeke, Daniel PW 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. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

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

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.

Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings*

of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 227–237, 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.

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.