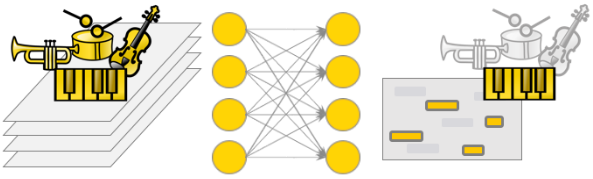
Report from Dagstuhl Seminar 22082

Deep Learning and Knowledge Integration for Music Audio Analysis

Meinard Müller*[1], Rachel Bittner*[2], Juhan Nam*[3], Michael Krause†[4], and Yigitcan Özer†[5]

1 **Friedrich-Alexander-Universität Erlangen-Nürnberg, DE.** meinard.mueller@audiolabs-erlangen.de
2 **Spotify – Paris, FR.** rachelbittner@spotify.com
3 **KAIST – Daejeon, KR.** juhan.nam@kaist.ac.kr
4 **Friedrich-Alexander-Universität Erlangen-Nürnberg, DE.** michael.krause@audiolabs-erlangen.de
5 **Friedrich-Alexander-Universität Erlangen-Nürnberg, DE.** yigitcan.oezer@audiolabs-erlangen.de



Abstract

Given the increasing amount of digital music, the development of computational tools that allow users to find, organize, analyze, and interact with music has become central to the research field known as Music Information Retrieval (MIR). As in general multimedia processing, many of the recent advances in MIR have been driven by techniques based on deep learning (DL). There is a growing trend to relax problem-specific modeling constraints from MIR systems and instead apply relatively generic DL-based approaches that rely on large quantities of data. In the Dagstuhl Seminar 22082, we critically examined this trend, discussing the strengths and weaknesses of these approaches using music as a challenging application domain. We mainly focused on music analysis tasks applied to audio representations (rather than symbolic music representations) to give the seminar cohesion. In this context, we systematically explored how musical knowledge can be integrated into or relaxed from computational pipelines. We then discussed how this choice could affect the explainability of models or the vulnerability to data biases and confounding factors. Furthermore, besides explainability and generalization, we also addressed efficiency, ethical and educational aspects considering traditional model-based and recent data-driven methods. In this report, we give an overview of the various contributions and results of the seminar. We start with an executive summary describing the main topics, goals, and group activities. Then, we give an overview of the participants' stimulus talks and subsequent discussions (listed alphabetically by the main contributor's last name) and summarize further activities, including group discussions and music sessions.

Seminar February 20–25, 2022 – http://www.dagstuhl.de/22082
2012 ACM Subject Classification Information systems → Music retrieval; Applied computing → Sound and music computing
Keywords and phrases Audio signal processing, deep learning, knowledge representation, music information retrieval, user interaction and interfaces
Digital Object Identifier 10.4230/DagRep.12.2.103

* Editor / Organizer
† Editorial Assistant / Collector

Deep Learning and Knowledge Integration for Music Audio Analysis, *Dagstuhl Reports*, Vol. 12, Issue 2, pp. 103–133
Editors: Rachel Bittner, Meinard Müller, Juhan Nam, Michael Krause, and Yigitcan Özer
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)
Rachel Bittner (Spotify – Paris, FR)
Juhan Nam (KAIST – Daejeon, KR)

This executive summary gives an overview of our discussions on the integration of musical knowledge in deep learning approaches while summarizing the main topics covered in this seminar. We also describe the seminar's group composition, the overall organization, and the seminar's activities. Finally, we reflect on the most important aspects of this seminar and conclude with future implications and acknowledgments.

Overview

Music is a ubiquitous and vital part of our lives. Thanks to the proliferation of digital music services, we have access to music nearly anytime and anywhere, and we interact with music in a variety of ways, both as listeners and active participants. As a result, music has become one of the most popular categories of multimedia content. In general terms, music processing research aims to contribute concepts, models, and algorithms that extend our capabilities of accessing, analyzing, understanding, and creating music. In particular, the development of computational tools that allow users to find, organize, analyze, generate, and interact with music has become central to the research field known as Music Information Retrieval (MIR). Given the complexity and diversity of music, research has to account for various aspects such as the genre, instrumentation, musical form, melodic and harmonic properties, dynamics, tempo, rhythm, timbre, and so on.

As in general multimedia processing, many of the recent advances in MIR have been driven by techniques based on deep learning (DL). For example, DL-based techniques have led to significant improvements for numerous MIR tasks including music source separation, music transcription, chord recognition, melody extraction, beat tracking, tempo estimation, and lyrics alignment. In particular, major improvements could be achieved for specific music scenarios where sufficient training data is available. A particular strength of DL-based approaches is their capability to extract complex features directly from raw audio data, which can then be used for making predictions based on hidden structures and relations. Furthermore, powerful software packages allow for easily designing, implementing, and experimenting with machine learning models based on deep neural networks (DNNs).

However, DL-based approaches also come at a cost, being a data-hungry and computing-intensive technology. Furthermore, the design of suitable network architectures (including the adaption of hyper-parameters and optimization strategies) can be cumbersome and time-consuming – a process that is commonly seen more as an art rather than a science. Finally, the behavior of DL-based systems is often hard to understand; the trained models may capture information that is not directly related to the core problem. These general properties of DL-based approaches can also be observed when analyzing and processing music, which spans an enormous range of forms and styles – not to speak of the many ways music may be generated and represented. While one aims in music analysis and classification problems at capturing musically relevant aspects related to melody, harmony, rhythm, or instrumentation, data-driven approaches often capture confounding factors that may not directly relate to the target concept (e.g., recording conditions in music classification or loudness in singing voice detection).

One main advantage of classical knowledge-based engineering approaches is that they result in explainable and explicit models that can be adjusted intuitively. On the downside, such hand-engineered approaches not only require profound signal processing skills as well as domain knowledge, but also may result in highly specialized solutions that cannot be directly transferred to other problems.

As mentioned earlier, one strong advantage of deep learning is its ability to learn, rather than hand-design, features as part of a model. Nowadays, it seems that attaining state-of-the-art solutions via machine learning depends more on the availability of large quantities of data rather than the sophistication of the approach itself. In this seminar, we critically questioned this statement in the context of concrete music analysis and processing applications. In particular, we explored existing approaches and new directions for combining recent deep learning approaches with classical model-based strategies by integrating knowledge at various stages in the processing pipeline.

There are various ways how one may integrate prior knowledge in DL-based MIR systems. First, one may exploit knowledge already at the input level by using data representations to better isolate information known to be relevant to a task and remove information known to be irrelevant (e.g., by performing vocal source separation before transcribing lyrics). Next, one may incorporate musical knowledge via the model architecture in order to force the model to use its capacity to characterize a particular aspect (e.g., limited receptive fields to prevent a model from "seeing" too much or introducing constraints that mimic DSP systems). Furthermore, the hidden representations can be conditioned to provide humans with "musically sensible control knobs" of the model (e.g., transforming an embedding space to separate out different musical instruments). Knowledge can also be exploited in the design of the output representation (e.g., structured output spaces for chord recognition that account for bass, root, and chroma) or the loss function used for optimization. During the data generation and training process, one may use musically informed data augmentations techniques to enforce certain invariances (e.g., applying pitch shifting to become invariant to musical modulations). Exploiting musical knowledge by combining deep learning techniques with ideas from classical model-based approaches was a core topic of this seminar.

The success of deep learning approaches for learning hidden structures and relations very much depends on the availability of (suitably annotated and structured) data. Therefore, as one fundamental topic, we discussed aspects of generating, collecting, accessing, representing, annotating, preprocessing, and structuring music-related data. These issues are by far not trivial. First of all, music offers a wide range of data types and formats, including text, symbolic data, audio, image, and video. For example, music can be represented as printed sheet music (image domain), encoded as MIDI or MusicXML files (symbolic domain), and played back as audio recordings (acoustic domain). Then, depending on the MIR task, one may need to deal with various types of annotations, including lyrics, chords, guitar tabs, tapping (beat, measure) positions, album covers, as well as a variety of user-generated tags and other types of metadata. To algorithmically exploit the wealth of these various types of information, one requires methods for linking semantically related data sources (e.g., songs and lyrics, sheet music and recorded performances, lead sheet and guitar tabs). Temporal alignment approaches are particularly important to obtain labels for automatic music transcription and analysis tasks. As for data accessibility, copyright issues are the main obstacle for distributing and using music collections in academic research. The generation of freely accessible music (including music composition, performance, and production) requires considerable effort, experience, time, and cost.

Besides the quantity of raw music data and its availability, another crucial issue is the input representation used as the front-end of deep neural networks. For example, log-frequency or Mel spectrograms are often used as input representations when dealing with music signals. We discussed recent research efforts where one tries to directly start with the raw waveform-based audio signal rather than relying on hand-engineered audio representations that exploit domain knowledge. In this context, we discussed how one might resolve phase shift issues by using carefully designed neural network architectures. Further recent research directions include the design of network layers to mimic common front-end transforms or incorporate differentiable filter design methods into a neural network pipeline.

Another central topic we discussed during our seminar was how to exploit musical structures via *self-supervised* and *semi-supervised learning*. Instead of relying on large amounts of labeled data, this technique exploits known variants and invariants of a dataset, using lots of *unlabeled* data. For example, without knowing the transcription of a musical piece, we know how the transcription would change if we shift the whole audio signal by some number of semitones. As another example, we can learn a notion of audio similarity by exploiting the fact that samples from a single musical audio signal are more similar than two samples drawn from different musical audio signals. We also discussed using multimodal data to give implicit labels, such as text, image, video, and audio correspondences. On the semi-supervised learning side, representations learned in a self-supervised way can be fine-tuned to a particular task with a small amount of labeled data. In this vein, we discussed model generalization, model adaptability, active learning, few-shot learning, and human-in-the-loop systems.

Finally, we addressed topics related to the evaluation of MIR systems. In particular, we discussed the gap between loss functions typically used for optimizing deep learning pipelines and evaluation metrics designed for evaluating specific MIR tasks. In this context, we pointed out the vulnerability of standard metrics to slight variances irrelevant to the perceived output quality, expressing the need for more reliable evaluation metrics. Furthermore, we envisioned the possibility of closing the gap by designing more meaningful loss functions that may be used in the context of end-to-end learning systems.

Participants and Group Composition

In our seminar, we had 22 participants, who came from various locations around the world, including North America (2 participants from the United States), Asia (2 participants from South Korea), and Europe (18 participants from France, Germany, Netherlands, Sweden, United Kingdom). The number of participants and international constellation are remarkable considering the ongoing pandemic. (Note that many of the invited participants, particularly from overseas, were not allowed to go on business trips.) More than half of the participants (12 out of 22) came to Dagstuhl for the first time and expressed enthusiasm about the open and retreat-like atmosphere. Besides its international character, the seminar was also highly interdisciplinary. While most of the participating researchers are working in music information retrieval, we also had participants with a background in musicology, signal processing, machine learning, mathematics, computer vision, and other fields. Our seminar stimulated cross-disciplinary discussions by having experts working in technical and non-technical disciplines while highlighting opportunities for new collaborations among our attendees. Furthermore, the number of participants from the industry (6 out of 22) was relatively high, which also underlines the relevance of the seminar's topic beyond fundamental research. Most of the participants had a strong musical background, some of them even having a dual career in an engineering discipline and music. This led to numerous social activities, including playing music together. In addition to geographical locations and research disciplines, we tried to foster variety in terms of seniority levels (e.g., we had three Ph.D. students and six participants on the postdoc/junior/assistant professor level) and in terms of gender (6 out of 22 of the participants identify as female). Besides scientific questions, we discussed in our seminar also various challenges that younger colleagues typically face when setting up their research groups and scientific curriculum at the beginning of their academic careers.

Overall Organization and Schedule

Dagstuhl Seminars have a high degree of flexibility and interactivity, allowing participants to discuss ideas and raise questions rather than presenting research results. Following this tradition, we fixed the schedule during the seminar asking for spontaneous contributions with future-oriented content, thus avoiding a conference-like atmosphere, where the focus tends to be on past research achievements. After the organizers gave an overview of the Dagstuhl concept, we started the first day with self-introductions, where all participants introduced themselves and expressed their expectations and wishes for the seminar. We then continued with short (15 to 20 minutes) stimulus talks, where specific participants addressed some critical questions related to the seminar's overall topic in a non-technical fashion. Each of these talks seamlessly moved towards an open discussion among all participants, where the respective presenters took over the role of a moderator. These discussions were well received and often lasted for more than half an hour. The first day closed with a brainstorming session on central topics covering the participants' interests while shaping the overall schedule and format for the next day. We continued having stimulus tasks interleaved with extensive discussions on the subsequent days. On the second day, we split into smaller groups, each group discussing a more specific topic in greater depth. The results and conclusions of these parallel group sessions, which lasted between 60 to 90 minutes, were then presented and discussed with the plenum. However, since the overall seminar size of 22 participants was relatively small, it turned out that the division into subgroups was not necessary. Thanks to excellent group dynamics and a fair distribution of speaking time, all participants had their say and were able to express their thoughts in the plenum while avoiding a monotonous conference-like presentation format. On the last day, we enjoyed a tutorial by Umut Simsekli on some theoretical concepts behind deep learning (a topic unanimously desired by the group). We concluded the seminar with a session we called "self-outroductions" where each participant presented their personal view on the seminar's results.

While working in technical engineering disciplines, most participants also have a strong background and interest in music. This versatility significantly impacted the seminar's atmosphere, leading to cross-disciplinary intersections and provoking discussions and resulting in intensive joint music-making during the breaks and in the evenings. One particular highlight was a concert on Thursday evening organized by Cynthia Liem and Christof Weiß, where various participant-based ensembles performed a wide variety of music, including classical music, Irish folk music, and jazz.

Conclusions and Acknowledgment

There is a growing trend toward building more interpretable deep learning systems, from the data collection and generation stage, to the input and output representations, to the model structure itself. On the other hand, classical model-based approaches bring a wealth of expertise on techniques for knowledge integration in system design. The Dagstuhl Seminar gave us the opportunity for connecting experts from classical model-based approaches, deep learning-based approaches, and related interdisciplinary fields such as music perception and human-computer interaction in order to generate discussion and spark new collaborations. The generation of novel, technically oriented scientific contributions was not the main focus of the seminar. Naturally, many of the contributions and discussions were on a conceptual level, laying the foundations for future projects and collaborations. Thus, the main impact of the seminar is likely to take place in the medium and long term. Some more immediate results, such as plans to share research data and software, also arose from the discussions. As further measurable outputs from the seminar, we expect to see several joint papers and applications for funding.

Besides the scientific aspect, the social aspect of our seminar was just as important. We had an interdisciplinary, international, and interactive group of researchers, consisting of leaders and future leaders in our field. Many of our participants were visiting Dagstuhl for the first time and enthusiastically praised the open and inspiring setting. The group dynamics were excellent, with many personal exchanges and shared activities. Some scientists expressed their appreciation for having the opportunity for prolonged discussions with researchers from neighboring research fields, which is often impossible during conference-like events. At this point, we would like to let some of the participants have their say:

- Stefan Balke (pmOne – Paderborn, DE): *“Dagstuhl is always a wonderful experience, having time to think, talk, and play music. All in a relaxed atmosphere, the seminar feels like a family meeting – especially in these times.”*
- Alice Cohen-Hadria (IRCAM – Paris, FR): *“Now I feel like a part of a community.”*
- Dasaem Jeong (Sogang University – Seoul, KR): *“Full of insightful discussions, music, and friends in a beautiful place.”*
- Cynthia Liem (TU Delft, NL): *“Dagstuhl is the one place in the world where one effectively can have a week long unconference. More deeply talking about research and new ideas, enjoying time with academic friends, with much less distraction than one would have at home, or even in a ‘regular’ conference. Especially coming out of a pandemic, I am realizing this is among the most valuable things in our professional life.”*
- Daniel Stoller (Spotify – Bonn, DE): *“Dagstuhl brings perspectives on the big issues.”*
- Yu Wang (New York University – Brooklyn, US): *“Discussion is like music: the live version is always better.”*

In conclusion, our expectations for the seminar were not only met but exceeded, in particular concerning networking and community building. We want to express our gratitude to the Dagstuhl board for giving us the opportunity to organize this seminar, the Dagstuhl office for their exceptional support in the organization process, and the entire Dagstuhl staff for their excellent service during the seminar. In particular, we want to thank Susanne Bach-Bernhard and Michael Gerke for their assistance during the preparation and organization of the seminar.

2 Table of Contents

Executive Summary

Stimulus Talks and Further Topics

Working Groups

3 Stimulus Talks and Further Topics

3.1 A Perspective for Machine Learning Education from (Non-Musical) Data Science Projects

Stefan Balke (pmOne – Paderborn, DE)

A current trend in the industry is to put a strong focus on data and its potential impact on the business side. However, companies tend to underrate the importance of data quality and integration. In fact, 80% of the project time is often spent on these two issues, while only 20% remain for modeling and other issues. However, the expectations on these models are high. Often, there is little knowledge on the business side regarding the mathematical concepts of optimization or machine learning (ML) in general. In addition, the amount of (annotated) data is often very limited. Deep learning approaches are increasingly used, especially for computer vision tasks where pre-trained models are available. In practice, ensemble or boosting methods (e.g., random forests or gradient boosting machines) are still used as the de facto standard in many other applications. Domain knowledge is often integrated into the model via custom features that resemble the underlying business or production processes.

Besides the aforementioned challenges, it is sometimes not clear whether the data resembles or measures the relevant factors (e.g., a missing temperature sensor in a process-critical position will undoubtedly impact the model's performance). Communicating and explaining these challenges to the customer is often more important than trying out the most recent ML approaches. Educators in this field should be aware that many data science projects currently do not fulfill the business side's high expectations. Many stakeholders in such projects still consider machine learning as some "magic" procedure that can solve any problem. Against this background, students should gain experience (e.g., in a class project) in building a dataset from scratch, including the (cumbersome) annotation process, cleaning the data, and developing a model. This experience will help them develop an intuition on the underlying challenges when using algorithms initially designed and tested under lab conditions (e.g., using benchmark datasets). Finally, students should learn to present the results (especially when they are not as expected) in a structured and transparent manner.

3.2 Toward a New Generation of Source Separation Metrics

Rachel Bittner (Spotify – Paris, FR)

The task of source separation has relied on what has become a standard set of metrics for the past 15 years. These metrics have been the basis for evaluating the latest "state-of-the-art," and determining when design choices are good or bad. Nevertheless, many articles have shown that these metrics are limited in how well they correlate with human perception. The metrics themselves have several undesirable properties, such as being extremely sensitive to phase and unstable at the origin. Moving forward, what could the next generation of metrics look like? The old metrics are not all bad – what can we keep or learn from them? How can we move beyond "one metric to rule them all" and account for a wider variety of applications and properties to measure? How do we move a community that is very used to ranking all algorithms using one metric away from the concept of "best" or "state-of-the-art"?

References

1 Emmanuel Vincent, Rémi Gribonval, Cédric Févotte. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech & Language Processing, 14(4): 1462–1469 (2006)

2 Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, Justinian P. Rosca. First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. ICA 2007: 552–559

3 Valentin Emiya, Emmanuel Vincent, Niklas Harlander, Volker Hohmann. Subjective and Objective Quality Assessment of Audio Source Separation. IEEE Transactions on Audio, Speech & Language Processing, 19(7): 2046–2057 (2011)

4 Emmanuel Vincent. Improved Perceptual Metrics for the Evaluation of Audio Source Separation. LVA/ICA 2012: 430–437

5 Estefanía Cano, Derry FitzGerald, Karlheinz Brandenburg. Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics. EUSIPCO 2016: 1758–1762

3.3 What is the Best Task to Learn a Generic Music Audio Representation?

Simon Durand (Spotify – Paris, FR) and Daniel Stoller (Spotify GmbH – Berlin, DE)

In the last few years, the fields of computer vision and natural language processing have witnessed a new generation of general-purpose models that can be used for a wide variety of downstream applications [1]. These are pre-trained models using a domain-specific pretext task before applying them to the downstream task of interest. The pretext task is usually conceptually simple and does not require side information such as expensive and often strongly biased labels, but powerful in the sense that a comprehensive understanding of the input is needed to perform well. These approaches also enormously benefit from scaling up model expressivity (such as Transformer models) and dataset size, which are enabled by hardware improvements such as Tensor Processing Units (TPUs).

If such an approach was applied to the music domain, we could obtain models that perform many different music audio analysis tasks even when only a few annotations for a task are available. However, music audio is different from language and vision, and we have not yet found a suitable music processing task with the same simplicity and strength as existent in other domains. Therefore, we review some pretext tasks used in existing approaches and discuss which tasks might be suitable for the music domain.

In natural language processing, randomly masking input tokens and asking the model to reconstruct them was used successfully as a pretext task [2]. Here, one needs to understand the context and semantic meaning of a sentence and its words to predict masked tokens. Similarly, in computer vision, we can predict randomly masked image patches [3]. However, it is questionable whether applying this directly to music audio spectrograms would result in an equally powerful pretext task due to the differing nature of spectrograms and images.

Other generative learning methods are promising pretext tasks, as they encourage a model to learn many semantically relevant features in the process. Generative models can be constructed in different ways, one of which is by using an autoregressive approach. Assuming that the input data can be represented as a sequence of elements, the model must predict the next element in the sequence, given the previous ones. In the NLP domain, [4] used a Transformer-based sequence model to predict the next word given the previous ones, which turned out to acquire many features useful for various downstream tasks. Similarly, making a model continue a given music piece could also yield semantically meaningful features – the usefulness of the Jukebox [5] features can be seen as an early example of such an approach. However, it is more difficult to identify where the relevant features from the model need to be extracted since auto-regressive models do not yield an explicit latent representation. This fact is in contrast to reconstruction-based generative models, such as variational auto-encoders (VAEs) and, in particular, VAEs using vector quantization (VQ-VAEs) that were shown to be effective at compressing music signals into short sequences of tokens [6]. Finally, the flow-based approach has shown some promise in computer vision, where the data is mapped to an equally high-dimensional but ordered latent space using a bijective function. Since the input is not compressed in this approach when moving to the latent space, it might be better suited for tasks where input details are relevant but worse for more abstract tasks.

An alternative task, called contrastive self-supervised learning, is to learn to compare different views of the same input such that the representation puts similar inputs closer together [7]. While it can learn a representation invariant to musical properties through data augmentation, it can be counterproductive for tasks dependent on these properties.

Our discussion is grounded on the specificities of music, on the desirable properties we would like a self-supervision task to highlight, and on the most promising musical applications that could come out of this.

References

1 Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, Jie Tang. Self-supervised Learning: Generative or Contrastive. CoRR abs/2006.08218 (2020)

2 Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171–4186

3 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

4 Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. Language Models are Unsupervised Multitask Learners. OpenAI Blog, 2019

5 Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, Ilya Sutskever. Jukebox: A Generative Model for Music. CoRR abs/2005.00341 (2020)

6 Aäron van den Oord, Oriol Vinyals, Koray Kavukcuoglu. Neural Discrete Representation Learning. NeurIPS 2017: 6306–6315

7 Michael Gutmann, Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. AISTATS 2010: 297–304

3.4 What is Actually our Problem? Generalization, Causality, and Error Surfaces

Sebastian Ewert (Spotify GmbH – Berlin, DE)

As outlined in the Dagstuhl Seminar's description, data-driven approaches tend to employ confounding factors to achieve results. Based on this observation, I raised a variety of questions in my talk.

- When is the use of confounders actually a problem, and when is it just fine?
  - In other words: Let us say a causal relationship between input and output exists, but the underlying information is hard to extract (which is usually the case). At the same time, some other information is easy to extract but only correlates with the output. Depending on the problem, we mix the uncertainty we get from the difficulty of extracting the proper information with the certainty that we can derive proper conclusions once we have that information. On the other hand, we could use correlations (where we often cannot measure how correlated or causal a factor is) and do everything to optimize average performance. If the latter is a problem, why is 95% of the ML community doing just that and humanity finds it useful?
  - One could argue that we might want to exclude confounding factors if we want to build a causal model of (some aspect of) music, where application performance does not matter, but performance is only helpful to compare different causal models to say which explains the data better. One provocative question is whether we, as MIR researchers, are interested in causal models. Do we understand the underlying principles of music theory and sound production practice? In other words, which aspect of music can not be broken down yet into causal components? Is there something about music we have not yet understood?
- In some cases, the use of confounders is why models do not generalize to new data domains (recording conditions, types of music, and so on). From that perspective, let us discuss when the integration of musical knowledge is actually the best solution to increase generalization capabilities and what "best" means in this case.
- To answer the former question, we should look into alternatives to improve generalizability.
  - In this context, we should discuss what adversarial attacks could solve. What if we encourage submissions that take existing methods and code and try to break it in interesting ways. What is the average Euclidean distance when modifying the input until we observe a confusion or perfect result (for classification problems)? What happens with out-of-domain data and unobserved data? Alternatively, could we encourage authors to attack their own systems? For example, the availability of adversarial attacks on speaker ID systems has led to tons of work to make such systems more robust. What can we learn from that?
  - Is there enough work to understand the error surfaces of neural networks? We have regularizers of all sorts, aiming to make the error more Lipschitz in one way or another. What does that mean? One aspect often discussed in this context is flat areas in error, where small parameter changes only lead to small changes in the output. As music is so structured, is there something here to help us characterize and identify such areas and potentially inspire us to create corresponding regularization approaches?

- Differentiable signal processing components have recently spawned some interest in the community. Do we understand how those should be seen and used in the context of confounders? There are also new stochastic tangent methods to estimate a gradient for non-differentiable components. Is this an option?
- Before the neural network hype, Bayes nets and probabilistic modeling were all the rage. Some of these could be seen as causal models. If the goal is to build and verify proper/causal models for aspects of music, should we not simply ditch neural networks and application performance and go back to these approaches? Alternatively, stochastic nets, such as Bayes neural networks, have recently attracted more and more attention in the theoretical ML world. Is this getting good enough that we should investigate if and how such approaches could build a bridge between the two worlds?

3.5 Closing the Gap Between Models and Users

Magdalena Fuentes (NYU – Brooklyn, US)

Machine learning (ML) now touches all of our lives. Such systems recommend which videos to watch, where to go on holiday, and how to get there. Though progress has been astounding, these ML models fall short when it comes to non-standard examples and subjective applications, such as analyzing underrepresented music genres, understanding people with accents or speech impediments, or recognizing faces from underrepresented ethnicities. ML-based models fall short because they are designed and trained to work well on "averages". In this context, I discussed in my talk how we could take steps to develop the next generation of ML models to close the gap between the models' output and the users' need. I argued how we could leverage several recent advancements in self-supervision, representation learning, and multimodal data analysis to move towards more user-centered ML.

3.6 Performance in Audio/MIDI/Features: Do we Need to Model Underlying Features?

Dasaem Jeong (Sogang University – Seoul, KR)

Joint work of Dasaem Jeong, Juhan Nam, Taegyun Kwon

In expressive performance modeling or performance style analysis (e.g., for piano music), it is common to use hand-crafted performance features instead of directly using audio or MIDI representations. One of the reasons is that it makes the problem more focused on *how* someone played, and not *what* or in *which* acoustic condition someone played. Researchers who are using performance features believe that these hand-crafted features somehow better capture the musical intention of the performer. Even when the output of a performance is encoded as audio or MIDI, we assume that hidden underlying features (e.g., related to musical tempo) are present. Explaining a musical performance as a function of an underlying musical tempo assumes that each note's position in time depends on the performer's internal musical tempo. Using this strong assumption, computational systems for expressive performance modeling have succeeded in mimicking human performances [1, 2].

Most recent deep learning (DL) approaches aim to model outputs in an end-to-end fashion. However, there are examples where those approaches fail, such as generating structured musical pieces with WaveNet trained on piano recordings [3]. On the other hand, the success of DDSP [4] with a limited amount of training data is an example of the effectiveness of modeling underlying features of instrumental sound (which is a combination of harmonic oscillations) instead of modeling the final observable output (i.e., waveform samples).

Still, counterexamples show that only modeling observations without explicit assumptions may be enough to model the complex structure of musical sounds and music. OpenAI's Jukebox [5] models music audio without explicitly modeling underlying features, such as notes, chords, or instrumentation. Also, recent research shows that neural networks that were trained to model observations also learned characteristics of underlying features. Examples are the usage of audio representations obtained by Jukebox for retrieval tasks [6], or the usage of 2D GANs to infer 3D depth [7].

There are further examples in DL-based MIR research, such as using explicit modeling of drum sounds for transcription [8], or using MIDI transcriptions instead of audio for composer identification [9]. However, it is still an open question whether it is beneficial to explicitly model underlying features or model end-level observations with larger-scale training data.

References

1 Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Kyogu Lee, Juhan Nam, VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance. ISMIR 2019: 908–915

2 Gerhard Widmer, Sebastian Flossmann, Maarten Grachten, YQX Plays Chopin. AI Mag. 30(3): 35–48 (2009)

3 Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, Koray Kavukcuoglu, WaveNet: A Generative Model for Raw Audio. SSW 2016: 125

4 Jesse H. Engel, Lamtharn Hantrakul, Chenjie Gu, Adam Roberts. DDSP: Differentiable Digital Signal Processing. ICLR 2020

5 Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, Ilya Sutskever. Jukebox: A Generative Model for Music. CoRR abs/2005.00341 (2020)

6 Rodrigo Castellon, Chris Donahue, Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. ISMIR 2021: 88–96

7 Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, Ping Luo. Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs. ICLR 2021

8 Keunwoo Choi, Kyunghyun Cho. Deep Unsupervised Drum Transcription. ISMIR 2019: 183–191

9 Qiuqiang Kong, Keunwoo Choi, Yuxuan Wang. Large-Scale MIDI-based Composer Classification. CoRR abs/2010.14805 (2020)

3.7 Soft Alignment Techniques for Music Information Retrieval

Michael Krause (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE) and
Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

Joint work of Michael Krause, Yigitcan Özer, Christian Dittmar, Frank Zalkow, Meinard Müller

The problem of aligning two time series is central to many areas of music information processing (MIR), including music synchronization, lyrics alignment, and music transcription. At the same time, more and more MIR tasks are approached using deep-learning (DL) models that consist of differentiable building blocks. In traditional music processing, the primary technique used by MIR researchers for aligning sequences of music data is *dynamic time warping* (DTW) [1], which finds a minimum cost alignment between two sequences, given appropriate feature representations and a suitable cost measure. However, the standard DTW recursion involves the computation of hard minima and, therefore, is not differentiable everywhere. Recently, this technique has been modified to be fully differentiable by using a "soft" way to compute minima in an approximative fashion [2, 3]. This strategy opens up the possibility of utilizing a soft DTW variant inside DL systems and backpropagate gradients through an alignment.

In this talk, we focused on music synchronization as a motivating scenario to demonstrate how soft alignment techniques might be helpful in MIR. The idea is to replace a classical pipeline based on chroma features and DTW with a DL-based approach using learned features and a soft DTW variant. We discussed several challenges arising from this related to trivial solutions and the efficiency of the gradient computation. The subsequent discussion focused on other application scenarios, in particular lyrics alignment and lyrics-informed source separation [4]. Furthermore, we discussed the relationship between the soft DTW variant and other techniques such as the connectionist temporal classification (CTC) function, the Viterbi algorithm, and the Baum–Welch algorithm. We found that the soft DTW variant allows for a cleaner formulation of the lyrics alignment problem than commonly used approaches such as CTC, but several technical challenges remain to be solved.

References

1 Meinard Müller. Fundamentals of Music Processing – Using Python and Jupyter Notebooks. Second Edition, Springer 2021

2 Marco Cuturi, Mathieu Blondel. Soft-DTW: a Differentiable Loss Function for Time-Series. ICML 2017: 894–903

3 Isma Hadji, Konstantinos G. Derpanis, Allan D. Jepson. Representation Learning via Global Temporal Alignment and Cycle-Consistency. CVPR 2021: 11068–11077

4 Kilian Schulze-Forster, Clément S. J. Doire, Gaël Richard, Roland Badeau. Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation. IEEE/ACM Transactions on Audio, Speech and Language Processing, 29: 2382–2395 (2021)

3.8 Towards Detecting Musical Patterns in Audio Recordings: A Study on Leitmotifs

Michael Krause (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE),
Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE),
Christof Weiß (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

During the Dagstuhl Seminar, we discussed some of our recent work on automatically detecting musical patterns in audio recordings with a particular focus on leitmotifs. These motifs are specific patterns associated with certain characters, places, items, or feelings occurring in an opera or a movie soundtrack [2]. Detecting such leitmotifs is particularly challenging since their appearance can change substantially throughout a musical work. Based on a dataset of 200 hours of audio with over 50 000 annotated leitmotif instances, we explored in [1] the benefits and limitations of deep-learning techniques for detecting leitmotifs. To investigate the robustness of the trained systems, we tested their sensitivity to different modifications of the input. We found that our deep-learning systems seem to work well in general. However, a closer investigation reveals that they capture confounding factors such as pitch distributions in leitmotif regions rather than identifying characteristic musical properties such as rhythm and melody. Thus, our in-depth analysis demonstrates some challenges that may arise from applying deep-learning approaches for detecting complex musical patterns in audio recordings. In personal conversations with Dagstuhl participants, we discussed how deep learning systems need to be adapted to detect musical patterns robustly. In particular, we found that these systems need to explicitly model the underlying musical structures (such as melody and rhythm) to prevent them from exploiting confounding factors.

References

1 Michael Krause, Meinard Müller, Christof Weiß. Towards Leitmotif Activity Detection in Opera Recordings. Transactions of the International Society for Music Information Retrieval, 4(1): 127-140 (2021)

2 Matthew Bribitzer-Stull. Understanding the Leitmotif. Cambridge University Press 2015

3.9 What Does it Mean to "Work as Intended"?

Cynthia Liem (TU Delft, NL)

With Dagstuhl being a good place for joint reflection, I took the audience along in a central question I have increasingly been pondering: What does it mean to "work as intended"? Considering the nature of our academic research and the applied nature of our work, do we succeed in the way we substantiate and communicate in our papers that our contributions are relevant and impactful? Are our contributions indeed relevant and impactful? What would a contribution need for this? In our discussions throughout the Dagstuhl Seminar, these considerations turned up several times when colleagues raised questions on current methodologies and metrics (associated with the question on "what would (not) get someone a paper").

Looking at topics surrounding deep learning and its applications, I also engaged in discussions on how to teach these topics to our future generations. I have been wondering whether we currently show and teach the proper examples to them and whether we encourage them to "work as intended" in responsible ways in the future. To illustrate this, in a joint talk with Bob Sturm, I illustrated how misconceptions and bad modeling choices (that a naive student may easily make) led to the childcare benefits scandal in The Netherlands. Next to this, I engaged in discussion groups on MIR teaching, where the broad consensus was that music uniquely gives opportunities to really engage with and consciously perceive data. As part of the discussion following my joint presentation with Bob, I finally raised a controversial question: Is the work we present as "science" actually scientific? Methodologically, I think we are mixing aspects of design, science, and engineering while still seemingly upholding "science" as the most important of these three. However, is this really justified? As a group, we did not reach a consensus on this matter, but lively discussions were held that will hopefully extend beyond this seminar.

3.10 How can we jointly represent the global and local aspects of music at multiple hierarchical levels?

Gabriel Meseguer Brocal (Deezer – Paris, FR)

Integrating musical knowledge into neural networks can be viewed from two perspectives:

- Using our prior knowledge about a task to guide computation ("telling the model *what* to obtain").
- Identifying our underlying perceptual mechanisms to construct blocks or learning paradigms to guide the model toward knowledge extraction without enforcing a specific output (i.e., "telling the model *how* to obtain it").

Both views raise challenging questions. In my presentation, I mainly discussed the latter view, even if it is unclear what our knowledge extraction mechanisms are or how they can be transferred to machine learning models. This hodgepodge includes exploring how we codify patterns and repetitions, develop a sense of hierarchical connections and structures, identify timbral spaces and textures, create context-dependent analysis, or enjoy music through a balance between predictability and surprise.

I find the human capacity to distill complex and compact representations to form high-level ideas at several hierarchical levels and our capacity to work with them fascinating. For instance, we can not only associate a new song to an existing artist (a pure classification task) but also recall it in a few instants without any external musical trigger. This capability shows that we have a deep understanding of what that artist is. The ability to develop hierarchical connections from low-level local elements (i.e., how we process information as a part of a whole) is essential to our knowledge distillation. How to mimic this ability when developing "intelligent" models [1, 2] is a fundamental research question. One main challenge is how to induce this part–whole behavior in our systems to let them derive complex but compact high-level musical ideas from local acoustic observations? These ideas open many other stimulating research questions:

- What are the learning paradigms as well as model and task definitions that induce the right invariances needed to create these capsules of knowledge?

- What is the ideal final high-level representation: a single one, a parse tree, or a graph?
- How do we overcome the problem of dynamically allocating model resources (e.g., to account for the number of essential elements and hierarchical levels that will undoubtedly vary from one song to another)?
- Can we distill knowledge by "listening to" the audio several times and take advantage of overfitting?

By finding answers to these questions (and raising many more), we may be able to capture compact but rich representations and better understand knowledge extraction and model behavior.

References

1 Anirudh Goyal, Yoshua Bengio. Inductive Biases for Deep Learning of Higher-Level Cognition. CoRR abs/2011.15091 (2020)

2 Geoffrey E. Hinton. How to represent part-whole hierarchies in a neural network. CoRR abs/2102.12627 (2021)

3.11 Combining NMF-Based Decomposition and Neural Network Techniques

Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE) and Yigitcan Özer (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

Nonnegative Matrix Factorization (NMF) is a powerful technique for factorizing, decomposing, and explaining data [4]. Thanks to its nonnegativity constraints and the multiplicative update rules that preserve these constraints in the training stage, it is easy to incorporate additional domain knowledge that guides the factorization to yield interpretable results. On the other hand, deep neural networks (DNNs), which can learn complex non-linear patterns in a hierarchical manner, have become omnipresent thanks to the availability of suitable hardware and software tools. However, deep learning (DL) models are often hard to interpret and control due to the massive number of trainable parameters. In this talked, we reviewed and discussed current research directions that combine the advantages of NMF-based and DL-based learning approaches. To make our discussion more concrete, we considered an audio decomposition application with the objective to decompose a music recording's magnitude spectrogram into musically meaningful spectral and activation patterns [1]. In this context, we addressed the following questions:

- How can one redraft an NMF-based decomposition as a DNN-based learning problem using autoencoder-like network architectures? Such an approach was originally proposed in [7].
- How can one retain nonnegativity constraints during the optimization process? Such approaches may be based on projected gradient descent methods [5] and other optimization schemes [10].
- How can one incorporate additional (score-based) prior knowledge into DNN-based models? In this context, one may apply ideas from structured regularization [9] and structured dropout [2].
- How can one simulate stacked NMF-like decomposition technique [8] within the DL framework, thus extending traditional flat NMF architectures?

- What is the effect of invertibility constraints of specific layers on the interpretability of the resulting models? How can such conditions be enforced at the training stage? A first approach was described in [3].

By systematically combining and transferring ideas between NMF-based and DL-based learning approaches, our goal was to understand better the interaction between various regularization techniques and DL-based learning procedures while improving the interpretability of DL-based decomposition results. Furthermore, using audio decomposition as a concrete example, we showed how education in machine learning might be supported by considering motivating and tangible music processing applications [6].

References

1 Sebastian Ewert, Meinard Müller. Using score-informed constraints for NMF-based source separation. ICASSP 2012: 129–132

2 Sebastian Ewert, Mark B. Sandler. Structured dropout for weak label and multi-instance learning and its application to score-informed source separation. ICASSP 2017: 2277–2281

3 Rainer Kelz, Gerhard Widmer. Towards Interpretable Polyphonic Transcription with Invertible Neural Networks. ISMIR 2019: 376–383

4 Daniel D. Lee, H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. NIPS 2000: 556–562

5 Chih-Jen Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. Neural Computation, 19(10): 2756–2779 (2007)

6 Meinard Müller, Brian McFee, Katherine M. Kinnaird. Interactive Learning of Signal Processing Through Music: Making Fourier Analysis Concrete for Students. IEEE Signal Processing Magazine, 38(3): 73–84 (2021)

7 Paris Smaragdis, Shrikant Venkataramani. A neural network alternative to non-negative audio models. ICASSP 2017: 86–90

8 George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, Björn W. Schuller. A Deep Semi-NMF Model for Learning Hidden Representations. ICML 2014: 1692–1700

9 Martin J Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. Annual Review of Statistics and Its Application 1:233–253 (2014)

10 Scott Wisdom, Thomas Powers, James W. Pitton, Les Atlas. Deep recurrent NMF for speech separation by unfolding iterative thresholding. WASPAA 2017: 254–258

3.12 What is the Future of Audio Representations for Music?

Juhan Nam (KAIST – Daejeon, KR)

The gist of deep learning (DL) is to learn representations from data to better predict the output or better structure the data within an embedding space. These representations can be learned through neural networks as part of an entire processing pipeline for a given task. For many music information retrieval (MIR) tasks, state-of-the-art neural network models still use time–frequency audio representations designed by explicitly exploiting domain knowledge. Examples are perceptually motivated Mel spectrogram or log-frequency spectrograms obtained by applying a musically motivated constant-Q transform. A recent research trend aims to obtain data-driven audio representations directly from raw audio waveforms (e.g., in sparse coding and convolutional neural networks). While these approaches have shown promising results in, e.g., music classification tasks, they require large-scale datasets to be successful (e.g., one million songs). Furthermore, these methods do not significantly outperform Mel spectrograms, and the learned (convolution) filters are often hard to interpret. As a compromise between hand-designed and fully-learned audio representations, researchers have attempted to learn filters with constraints (e.g., phase invariance) or learn only parameters of pre-designed filter prototypes. While reviewing recent work on this topic, we discussed possible research directions for learning audio representations and their potential for music processing.

References

1 Evan Smith, Michael Lewicki. Efficient Audio Coding. Nature, 2006

2 Sander Dieleman, Benjamin Schrauwen. End-to-end learning for music audio. ICASSP 2014: 6964–6968

3 Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, Juhan Nam. Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms. CoRR abs/1703.01789 (2017)

4 Taejun Kim, Jongpil Lee, Juhan Nam. Comparison and Analysis of SampleCNN Architectures for Audio Classification. IEEE Journal of Selected Topics in Signal Processing, 13(2): 285–297 (2019)

5 Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, Xavier Serra. End-to-end Learning for Music Audio Tagging at Scale. ISMIR 2018: 637–644

6 Minz Won, Andres Ferraro, Dmitry Bogdanov, Xavier Serra. Evaluation of CNN-based Automatic Music Tagging Models. CoRR abs/2006.00751 (2020)

7 Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS 2020

8 Minz Won, Sanghyuk Chun, Oriol Nieto, Xavier Serra. Data-Driven Harmonic Filters for Audio Representation Learning. ICASSP 2020: 536–540

9 Frank Cwitkowitz, Mojtaba Heydari, Zhiyao Duan. Learning Sparse Analytic Filters for Piano Transcription. CoRR abs/2108.10382 (2021)

3.13 Source Separation of Piano Concertos with Test-Time Adaptation

Yigitcan Özer (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE) and Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

Music source separation (MSS) aims at decomposing a music recording into constituent sources, such as a lead instrument and the accompaniment. Despite the difficulties in MSS due to the high correlation of musical sources in time and frequency, deep neural networks (DNNs) have led to substantial improvements to accomplish this task [1, 2]. For training supervised machine learning models such as DNNs, isolated sources are required. In the case of popular music, one can exploit open-source datasets which involve multitrack recordings of vocals, bass, and drums. For western classical music, however, isolated sources are generally not available. In this talk, we considered the case of piano concertos, which are composed for a pianist typically accompanied by an orchestra. The lack of multitrack recordings makes training supervised machine learning models for the separation of piano and orchestra challenging. To overcome this problem, we suggest generating artificial training material by randomly mixing sections of the solo piano repertoire (e.g., piano sonatas) and orchestral pieces without piano (e.g., symphonies) to train state-of-the-art DNN models for MSS. Furthermore, we propose a test-time adaptation (TTA) procedure, which exploits random mixtures of the piano-only and orchestra-only parts in the test data to finetune the separation quality. To this end, one may first train on an artificially generated dataset. Then, the idea is to exploit that piano concertos comprise long piano-only (e.g., in the *Cadenza*) and orchestra-only (e.g., in the *Exposition*) sections. Using these sections, one can generate synthetic mixtures for the test item at the testing stage to adapt the model and enhance the separation. In this context, we discussed the following points:

- Although random mixes are not musically plausible, they already yield an acceptable separation quality. Would it be feasible to generate further data for this task using an orchestra part (e.g., provided by Music Minus One) of piano concertos and mix them with the piano part played by various pianists on various instruments?
- How can one design an additional loss term to better capture the onsets of the piano?
- Can generative models, e.g., GANs, help to reduce the interference between the separated piano and orchestral sources?
- Would it be a good idea to introduce hierarchical instrument classes to the network such as strings, percussion, woodwinds, and so on?
- Would it be sensible to integrate a piano transcription model into the network?

References

1 Romain Hennequin, Anis Khlif, Fálix Voituret, Manuel Moussallam. Spleeter: A fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5(56): 2154 (2020)

2 Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, Yuki Mitsufuji. Open-Unmix – A Reference Implementation for Music Source Separation. Journal of Open Source Software, 4(41): 1667 (2019)

3.14 Designing Audio-Specific Deep Learning Front-End Transforms

Geoffroy Peeters (Telecom Paris, FR)

Joint work of Félix Mathieu, Thomas Courtat, Gaël Richard, Geoffroy Peeters

Deep learning front-ends (or encoders) refer to the first projections of a deep neural network (DNN) applied directly to the raw audio waveform. Therefore, the front-end is the DNN part that should be the most adapted to the specificities of the audio signal. Usually, the front-end transform consists of a set of 1D-convolutions applied to the waveform, corresponding to a set of temporal basis functions, kernels, or filters. Such basis functions can simply be the cosine and sine kernels of a short-time Fourier transform (STFT) or constant-Q transform (CQT), the modulus of which brings the nice phase-shift-invariance (PSI) property.

Such kernels can also be trained. For example, the authors of [1] propose to train an end-to-end DNN with a 1D-convolution front-end. In particular, the STFT is mimicked by using real-valued kernels with sizes and strides equal to the STFT window length and hop size. However, this strategy fails to reproduce the STFT's PSI property, requiring the learned kernels to account for every possible phase shift. To reduce the number of required phase shifts to be learned, [2] proposes to reduce the length of the kernels to three samples and adopt an approach similar to the VGG-net architecture (i.e., a pyramid of projections with small kernels).

To better understand what the kernels have learned, the modulus of their DFTs is often analyzed (with the hope that those will highlight band-pass filters). To make this explicit, in the SincNet as used in [3], the kernels are parameterized as band-pass filters (represented as the difference between two low-pass Sinc kernels) with learnable parameters. While interesting, SincNet kernels still do not provide PSI. To ensure PSI, [5] extended the SincNet to the Complex-Gabor, which, apart from bringing the PSI, also provides the best time–frequency trade-off. For the same purpose (getting PSI projections), [6] proposed to define the imaginary kernels as the Hilbert transform of the learned real-valued kernels. Recently, Ditter and Gerkmann [4] showed that it is possible to reach good results using an untrained multi-phase gammatone filter bank. As opposed to Fourier-like kernels, the projection is performed with amplitude-modulated kernels.

To gain the benefits of all the previous approaches, we recently proposed extended Hilbert–Bedrosian kernels, where we learn both a modulating and a modulated career basis kernel [7]. The proposed transform achieves the PSI property as well as an envelope-shift-invariance property. We show that we can use large temporal kernels (as the Fourier transform does), hence reducing the overall computational cost with state-of-the-art results for ConvTasNet-like architectures.

References

1 Sander Dieleman, Benjamin Schrauwen. End-to-end learning for music audio. ICASSP 2014: 6964–6968

2 Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, Juhan Nam. Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms. CoRR abs/1703.01789 (2017)

3 Mirco Ravanelli, Yoshua Bengio. Speaker Recognition from Raw Waveform with SincNet. SLT 2018: 1021–1028

4 David Ditter, Timo Gerkmann. A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet. ICASSP 2020: 36–40

5 Paul-Gauthier Noé, Titouan Parcollet, Mohamed Morchid. CGCNN: Complex Gabor Convolutional Neural Network on Raw Speech. ICASSP 2020: 7724–7728

6 Manuel Pariente, Samuele Cornell, Antoine Deleforge, Emmanuel Vincent. Filterbank Design for End-to-end Speech Separation. ICASSP 2020: 6364–6368

7 Félix Mathieu, Thomas Courtat, Gaël Richard and Geoffroy Peeters. Phase Shifted Bedrosian Filter Bank: An Interpretable Audio Front-End for Time-Domain Audio Source Separation. ICASSP 2022

3.15 Unsupervised Source Separation with Model-Based Deep Learning

Gaël Richard (Telecom Paris, FR)

Joint work of Kilian Schulze-Forster, Clément S. J. Doire, Gaël Richard, Roland Badeau

Like many other multimedia-related research areas, music audio analysis has rapidly moved towards pure data-driven deep learning models where domain knowledge is deduced from the processed data. Current state-of-the-art supervised deep learning methods for music source separation require an aligned dataset of mixtures with corresponding isolated source signals. Such datasets are difficult and costly to acquire. In a recent and preliminary study [1], we proposed a novel unsupervised model-based deep learning approach to the specific use case of choir singing separation. In this work, we represent each source by a differentiable parametric source-filter singing voice model. Then, we train the neural network to reconstruct the observed mixture as a sum of the sources by estimating the source models' parameters given their fundamental frequencies. We believe that integrating domain knowledge in the form of audio models into a data-driven method is key to developing efficient and more frugal machine listening systems. After a brief presentation of this work in the Dagstuhl Seminar, we discussed the different research avenues for exploring and extending this concept of model-based deep learning for music audio analysis.

References

1 Kilian Schulze-Forster, Clément S. J. Doire, Gaël Richard, Roland Badeau. Unsupervised Audio Source Separation Using Differentiable Parametric Source Models. CoRR abs/2201.09592 (2022)

3.16 Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms

Umut Şimşekli (INRIA – Paris, FR)

Joint work of Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gürbüzbalaban, Umut Şimşekli, Lingjiong Zhu
Main reference Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gürbüzbalaban, Umut Simsekli, Lingjiong Zhu: "Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms", in Proc. of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 18774–18788, 2021.
URL https://openreview.net/pdf?id=WlkzLjxpYe

Understanding generalization in deep learning has been one of the major challenges in statistical learning theory over the last decade. While recent work has illustrated that the dataset and the training algorithm must be taken into account to obtain meaningful generalization bounds, it is still theoretically not clear which properties of the data and the algorithm determine the generalization performance. In this talk, we approached this problem from a dynamical systems theory perspective where stochastic optimization algorithms are represented as random iterated function systems (IFS). Well studied in the dynamical systems literature, such IFSs can be shown (under mild assumptions) to be ergodic with an invariant measure that is often supported on sets with a fractal structure. As our main contribution in [1], we prove that the generalization error of a stochastic optimization algorithm can be bounded based on the complexity of the fractal structure that underlies its invariant measure. Leveraging results from dynamical systems theory, we show that the generalization error can be explicitly linked to the choice of the algorithm (e.g., stochastic gradient descent), algorithm hyperparameters (e.g., step size, batch size), and the geometry of the problem (e.g., Hessian of the loss).

References

1 Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gurbuzbalaban, Umut Şimşekli, Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. Advances in Neural Information Processing Systems 34 (2021).

3.17 Thinking Deeply about Ethics

Bob Sturm (KTH Royal Institute of Technology – Stockholm, SE)

Main reference Music at the Frontiers of Artificial Creativity and Criticism
URL https://musaiclab.wordpress.com/

In my stimulus talk, I reflected on three deep questions:

- How does my work benefit the world?
- How does my work harm the world?
- How do I know?

These are questions of ethics, and reflecting on them provides reusable insights. I illustrated these with my own personal journey of research in applying artificial intelligence to Irish traditional music. Frictions caused by this research motivated me to think about whether such an application harms the world or even whether it provides a benefit – outside of my own professional success. So, I set about learning the tradition by reading, listening, and playing an instrument with a teacher. I have started participating in the tradition to understand the origin of these frictions. These reflections and personal practice have widened my perspectives and developed a new appreciation. I have enlarged my networks outside the ivory tower and am asking more enriching questions. I have deepened my interaction with the world and have become sensitive to responsible engineering – which is constantly under review.

3.18 Knowledge Exchange between Computer Vision and MIR fields

Gül Varol (ENPC – Marne-la-Vallée, FR)

In this stimulus talk, we focused on fostering knowledge exchange between computer vision (CV) and music information retrieval (MIR) research communities. First, there was a discussion on *whether* this could be beneficial by illustrating some example works on CV applications on sequential data, highlighting similar problems, and making analogies with MIR. Next, several interdisciplinary research directions were shown to provide examples for *how* to enable more interaction between the two communities.

Neural network architectures increasingly become general-purpose across data modalities, such as image, video, audio, and text. A recent successful example is the family of attention-based Transformer models [9], typically taking sequential data as input. In our works, we adapted this model for problems such as subtitle text alignment in sign language videos [3], text-to-video retrieval [4], and 3D human motion synthesis from textual descriptions [2, 1]. While these tasks are different, common tools can be applied since the data types for video, text, and 3D motion are all similar in their sequential nature. In computer vision, many open questions remain on how to perform temporal modeling best, on how to deal with long-term sequences, and which architecture to employ (e.g., RNNs, CNNs, Transformers, MLPs). Considering that music and audio data are also temporal, any solutions found in one field can inspire the other. Moreover, there exist shared tasks such as alignment, where, for example, subtitle alignment can borrow ideas from MIR problems such as music-lyrics alignment and vice versa.

This Dagstuhl Seminar specifically focused on knowledge integration in deep neural networks. This talk provided several examples of how our works in CV inject domain knowledge, such as physics constraints in 3D hand-object reconstruction from images [6] and differentiable human body model SMPL [10] as a network layer [2]. This discussion led to making analogies between the differentiable 3D models and Differentiable Digital Signal Processing (DDSP) tools used in MIR. Furthermore, instead of integrating, one can *extract* knowledge from end-to-end black-box approaches. Using the attention mechanism as a way to perform localization in long sequences implicitly is an example for such knowledge extraction [5].

Several research problems were identified to have the potential to encourage collaborations across CV and MIR communities, among which are music-conditioned dance generation [7] and visual piano transcription [8]. We expect that bringing together researchers from the two fields for these collaborative projects can facilitate general knowledge exchange as a side product.

References

1 Mathis Petrovich, Michael J. Black, Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. arXiv, 2022.

2 Mathis Petrovich, Michael J. Black, Gül Varol. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. ICCV 2021: 10965–10975

3 Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, Andrew Zisserman. Aligning Subtitles in Sign Language Videos. ICCV 2021: 11532–11541

4 Max Bain, Arsha Nagrani, Gül Varol, Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. ICCV 2021: 1708–1718

5 Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, Andrew Zisserman. Read and Attend: Temporal Localisation in Sign Language Videos. CVPR 2021: 16857–16866

6 Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, Cordelia Schmid. Learning Joint Reconstruction of Hands and Manipulated Objects. CVPR 2019: 11807–11816

7 Ruilong Li, Shan Yang, David A. Ross, Angjoo Kanazawa. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. ICCV 2021: 13381–13392

8 A. Sophia Koepke, Olivia Wiles, Yael Moses, Andrew Zisserman. Sight to Sound: An End-to-End Approach for Visual Piano Transcription. ICASSP 2020: 1838–1842

9 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is All you Need. NIPS 2017: 5998–6008

10 Matthew Loper and Naureen Mahmood and Javier Romero and Gerard Pons-Moll, Michael J. Black. SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, 34(6): 248:1–248:16 (2015)

3.19 Towards Adaptive and Interactive Machine Listening with Minimal Supervision

Yu Wang (New York University – Brooklyn, US)

Machine listening aims at endowing a machine with the ability to perceive and understand audio signals as humans do. It can be applied to different audio domains such as music, speech, and environmental sounds. Nowadays, approaches based on deep learning have become mainstream tools and achieved state-of-the-art performance in multiple research areas, including machine listening. A deep learning model that generalizes well needs to be trained on a large amount of labeled data. While it is easy to collect a large amount of audio data, labeling them is very costly and often requires expert knowledge. Therefore, current studies in machine listening often suffer from small datasets, which lead to poor generalizability and small vocabulary. Existing strategies tackling this labeled data scarcity issue often still require a significant amount of human effort [1] or have generalizability issues when applied to real audio [2]. To address this, we propose a new perspective that instead of focusing on collecting more labeled data to train a giant universal model, we aim for a flexible and customizable system that can adapt to different tasks quickly with the help of minimal supervision from human input. We envision a paradigm where the model can learn to recognize a new target sound at inference time in a real-time, on-the-fly fashion, based on just a handful of examples (e.g., five) provided by a human user. To do so, we leverage metric-based few-shot learning techniques [3, 4] to learn a discriminative embedding space that can generate a robust representation for an unseen novel class based on a few examples. We applied the trained few-shot learning models to various tasks in different audio domains including keyword spotting [5], drum transcription [6], large vocabulary audio tagging [7], and musical source separation [8].

References

1 Mark Levy. Improving perceptual tempo estimation with crowd-sourced annotations. ISMIR 2011: 317–322

2 Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. WASPAA 2017: 344–348

3 Jake Snell, Kevin Swersky, Richard S. Zemel. Prototypical Networks for Few-shot Learning. NIPS 2017: 4077–4087

4 Spyros Gidaris, Nikos Komodakis. Dynamic Few-Shot Visual Learning Without Forgetting. CVPR 2018: 4367–4375

5 Yu Wang, Justin Salamon, Nicholas J. Bryan, Juan Pablo Bello. Few-Shot Sound Event Detection. ICASSP 2020: 81–85

6 Yu Wang, Justin Salamon, Mark Cartwright, Nicholas J. Bryan, Juan Pablo Bello. Few-shot Drum Transcription in Polyphonic Music. ISMIR 2020: 117–124

7 Yu Wang, Nicholas J. Bryan, Mark Cartwright, Juan Pablo Bello, Justin Salamon. Few-Shot Continual Learning for Audio Classification. ICASSP 2021: 321–325

8 Yu Wang, Daniel Stoller, Rachel M. Bittner, and Juan P. Bello. Few-shot musical source separation. ICASSP 2022

3.20 Is Deeper Better? Towards Reliable Evaluation in Music Transcription

Christof Weiß (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE) and Geoffroy Peeters (Telecom Paris, FR)

Main reference Christof Weiß, Geoffroy Peeters: "Deep-Learning Architectures for Multi-Pitch Estimation: Towards Reliable Evaluation", CoRR, Vol. abs/2202.09198, 2022.
URL https://arxiv.org/abs/2202.09198

Extracting pitch information from music recordings is a challenging and essential problem in music signal processing. Frame-wise transcription or multi-pitch estimation aims for detecting the simultaneous activity of pitches in polyphonic music recordings and has recently seen significant improvements thanks to deep-learning techniques, with a variety of proposed network architectures. In a recent study [1], we tested different architectures based on convolutional neural networks, the U-net structure, and self-attention components, also proposing several modifications to those. When comparing variants of these architectures with varying capacity using the MusicNet dataset [2], we observed the following:

- Most architectures yield competitive results, and larger model variants seem to be beneficial.
- These results substantially depend on randomness in parameter initialization, data augmentation, order of training examples, and dropout.
- In our cross-version evaluation, where we exploit that our dataset contains several performances (or versions) of each musical work, the particular choice of the training–test splits has a crucial influence on the results.

Concluding from these observations, we question the claim of superiority ("state-of-the-art") for particular architectures given only small improvements in many published results. To approach these problems, we suggest to shift the focus from technical design choices (such as network architectures) to the more general aspects of experimental design, validity, and generalization with a particular consideration of music-specific properties in the datasets.

References

1 Christof Weiß and Geoffroy Peeters. Deep-Learning Architectures for Multi-Pitch Estimation: Towards Reliable Evaluation. CoRR abs/2202.09198 (2022)

2 John Thickstun, Zaïd Harchaoui, Sham M. Kakade. Learning Features of Music from Scratch. CoRR abs/1611.09827 (2016)

4 Working Groups

4.1 Teaching MIR

Participants of Dagstuhl Seminar 22082

In this working group, we exchanged ideas about how we teach music processing and what conditions and requirements we find at our home institutes. Working in different departments (e.g., engineering, computer science, music sciences) and countries (e.g., France, Germany, South Korea, Netherlands, Sweden), we reported very different experiences. However, we agreed that music is a beautiful and instructive application domain, yielding an intuitive entry point to support education in various fields and on various levels [1]. We summarize some of our thoughts and discussion points in more detail below.

- From our practice as teachers, a shared experience is that our students often have diverse backgrounds and needs. Some students come from the music field with a good background in music theory and performance, while others come from more technical disciplines such as computer science or signal processing with solid programming skills. The heterogeneity of the audience within a lecture or course often makes it challenging to reconcile the needs of all participants. On the positive side, new possibilities for interdisciplinary and interactive collaborations (e.g., within the framework of project groups) are created.

- Especially when teaching a mixed audience, it is essential as a teacher to ensure that all students can follow the course. Most of us agreed that a "task-driven" teaching approach might be helpful, where one uses a concrete music processing task (e.g., beat tracking) as a starting point to motivate abstract concepts (e.g., periodicity analysis for time series). Such an approach also makes it possible to discuss music-theoretical concepts and properties of music signals before the task is mathematically modeled and tackled with algorithmic methods.
- We also discussed intensively whether and to what extent current deep learning (DL) techniques should be part of a lecture on music information retrieval (MIR) and music processing. On the one hand, DL-based methods are state-of-the-art for many MIR tasks, which speaks in favor of introducing students to such techniques at an early stage. On the other hand, these techniques are often like black boxes and may not provide profound insights into the MIR task and the underlying music data. We agreed that it is essential for students to gain the ability to question things and develop skills that go beyond "pushing buttons" and applying "off-the-shelf"-methods. As teachers, it is our responsibility to find a good balance to let our students acquire technical skills and gain a deep understanding on what they do.
- An essential aspect of MIR education is understanding the actual music data and its specific properties. Especially in the age of deep learning, a dataset's composition and data quality are crucial. Knowing possible data biases and annotation ambiguities is essential for understanding the behavior of a DL-based approach. Therefore, to sensitize our students to such aspects, we believe they should create and annotate a dataset themselves during their scientific training.

In conclusion, we agreed that music is an instructive, challenging, and multi-faceted domain of application, which may serve as a motivating and intuitive entry point for teaching and learning a wide range of topics beyond MIR, including signal processing, machine learning, and information retrieval. In addition, teaching a music processing or MIR class also provides an excellent opportunity to address sociological and ethical issues – aspects that are often neglected in the thicket of technological details.

References

1 Meinard Müller, Brian McFee, Katherine M. Kinnaird. Interactive Learning of Signal Processing Through Music: Making Fourier Analysis Concrete for Students. IEEE Signal Processing Magazine, 38(3): 73-84 (2021)

4.2 Integrating Musical Knowledge I

Participants of Dagstuhl Seminar 22082

In our group discussion, we identified a multitude of ways how musical knowledge could be incorporated into deep learning (DL) models. Almost all system design choices are, or can be informed by understanding the problem domain, including training data, augmentations, invariances, conditioning, representations, features, model architectures, loss functions and evaluation metrics. We then debated whether knowledge integration is a useful or necessary strategy, and how its influence and success could be measured. We noted the difficulty of discussing the topic without reference to one or more specific tasks. Unlike in other fields, where domain knowledge is grounded on physical objects, music is an abstract art form where musical concepts are often hard to grasp. In our discussion, going beyond Western music theory, we attempted to define musical knowledge broadly, covering general (perceptually motivated) concepts such as expectation and surprise. We agreed that musical knowledge can be used as an inductive bias to favour musically likely outcomes, but it remained unclear how one could demonstrate that this bias is more useful (or fair) than the natural biases of the datasets used. Furthermore, we identified example cases where musical knowledge has proven beneficial, and we mentioned HCQT as an input representation, music language models in transcription, and prior knowledge of instrumentation of the input data when performing source separation. We also discussed the potential reduction in training data requirements as an advantage of using domain knowledge, and thus in computational cost. Energy metrics would be one way to demonstrate this advantage, as a contribution to Green AI.

4.3 Integrating Musical Knowledge II

Participants of Dagstuhl Seminar 22082

In this working group, we discussed the question of knowledge integration for deep-learning systems in music information retrieval. First, we asked ourselves *where* and *how* to integrate knowledge: typical points of leverage can be the dataset creation, the input representation, the network architecture, the loss, or data augmentation strategies. Moreover, knowledge integration at the front-end side (signal processing knowledge) can have different effects than that at the back-end side (musicological or user knowledge). Having an interpretable intermediate representation can help to understand the behavior for downstream tasks as shown for music auto-tagging [1]. Whether knowledge integration is beneficial depends on the desired application, which can be categorized according to various dimensions such as level of expertise or type of task (generation or analysis). For instance, unexpected behavior of an end-to-end system might be considered valuable ("creative") for music generation. We then turned to an in-depth discussion about the benefits of knowledge integration instead of end-to-end learning. On the one hand, end-to-end systems have shown to be superior to systems integrating knowledge in fields such as computer vision. Moreover, the question arises whether integrating knowledge always leads to a bias. For instance, an input representation relying on the harmonic series will bias results towards music with harmonic instruments and may not work well with inharmonic instruments such as bells, gongs, or low piano notes. On the other hand, such a bias or characteristic behavior may be desired when targeting specific application scenarios. In this sense, knowledge integration can help make a generation system more controllable or an analysis system more objective, interpretable, or plausible. In particular, an analysis system with a human in the loop can greatly use musical knowledge for the reasons mentioned above. Furthermore, all participants agreed that knowledge integration has a high potential to account for smaller and more efficient networks, which helps to move forward towards "Green AI" without losing performance. Finally, touching upon questions raised in the earlier stimulus talks and panel discussions, we agreed that evaluation of model performance is tricky for the case of music information retrieval. Dataset size, annotation quality, annotator bias, and problematic evaluation metrics often limit the validity of experiments, especially given the small-scale improvements usually reported for novel systems. If we consider today's results in various MIR tasks to be close to a certain upper bound or "glass ceiling," the main value of knowledge integration may be getting to this upper bound in a more resource-efficient and interpretable way.

References

1 Minz Won, Sanghyuk Chun, Xavier Serra. Toward Interpretable Music Tagging with Self-Attention. CoRR abs/1906.04972 (2019)

Participants

- Stefan Balke
pmOne – Paderborn, DE
- Rachel Bittner
Spotify – Paris, FR
- Alice Cohen-Hadria
IRCAM – Paris, FR
- Simon Dixon
Queen Mary University of London, GB
- Simon Durand
Spotify – Paris, FR
- Sebastian Ewert
Spotify GmbH – Berlin, DE
- Magdalena Fuentes
NYU – Brooklyn, US
- Dasaem Jeong
Sogang University – Seoul, KR
- Michael Krause
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE
- Cynthia Liem
TU Delft, NL
- Gabriel Meseguer Brocal
Deezer – Paris, FR
- Meinard Müller
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE
- Juhan Nam
KAIST – Daejeon, KR
- Yigitcan Özer
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE
- Geoffroy Peeters
Telecom Paris, FR
- Gaël Richard
Telecom Paris, FR
- Umut Simsekli
INRIA – Paris, FR
- Daniel Stoller
Spotify GmbH – Berlin, DE
- Bob Sturm
KTH Royal Institute of Technology – Stockholm, SE
- Gül Varol
ENPC – Marne-la-Vallée, FR
- Yu Wang
New York University – Brooklyn, US
- Christof Weiß
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE