Shift-Invariant Dictionary Learning using TCN-WTA Autoencoders for Discovering Musical Relations

Anonymous ACL submission

Abstract

Music hierarchical temporal structure is full of shift invariant patters. The standard methods to encode a generic sequence is usually achieved by recurrent architectures or more recently with transformer that adopts the mechanism of attention. However, RNNs and transformers models do not take advantage of this prior information, or attempt to find repetitive building blocks. Temporal Convolutional Nets can be used to extract shift invariant features of a specific length defined by the kernel size. Using a fully convolutional temporal autoencoder we can find a shift invariant dictionary that can recreate multivariate musical signals. This architecture can strided with no overlap, and be combined to K-WTA activation function to obtain a sparse dictionary promote a sparse representation. In addition to gaining insight into this shift invariant patters, some results indicate that CNN architectures can outperform recurrent networks on specific task and provide server other advantages across a diverse range of tasks and datasets, while demonstrating longer effective memory. We show a few applications of this sparse representation on task to find key signatures, time signatures, artist detection, and music generation. To assist related work, we have made code available.

1 Introduction

What are the benefits of having sparse models? First, as we will show, they can be used to encode prior knowledge in the sparsity patterns. Second, they are lightweight— requiring less memory to store and allowing faster inference and easier interpretability. Nowadays, we often start with models with hundreds of millions to billions of parameters. Sparsity provides a way to completely discard some of these parameters in an informed and principled manner, resulting in smaller model size. For example, mobile applications (e.g., Google Now, Siri, etc.) stand to benefit from smaller models since mo- bile phones typically have less storage and computing power than standard computers. Sparse models come with their own challenges. New varieties of sparse models often require a specialized optimization method, as we will see throughout this thesis. Last, some of the state-of-the-art methods for benchmarks tasks in various application areas such as computational biology (Kim and Xing, 2008) and computer vision are sparse models (Ranzato et al., 2006; Lin and Kung, 2014), empirically demonstrating that they can also lead to statistical improvements if the prior knowledge is correct (Stojnic et al., 2009).

2 Related Work

2.1 Dictionary learning

Given the data: $X = [x_1, \ldots, x_K], x_i \in \mathbb{R}^d$ We want a dictionary $\mathbf{D} \in \mathbb{R}^{d \times n} : D = [d_1, \ldots, d_n]$ And a representation $R = [r_1, \ldots, r_K], r_i \in \mathbb{R}^n$ such that the reconstruction $\|X - \mathbf{D}R\|_F^2$ is minimized and r_i are sparsed. The optimization problem can be formulated as:

$$\underset{\mathbf{D} \in \mathcal{C}, r_i \in \mathbb{R}^n, \lambda > 0}{\operatorname{argmin}} \sum_{i=1}^K \|x_i - \mathbf{D}r_i\|_2^2 + \lambda \|r_i\|_0$$

$$\mathcal{C} \equiv \left\{\mathbf{D} \in \mathbb{R}^{d \times n} : \|d_i\|_2 \leq 1 \forall i = 1, \ldots, n\right\}$$

However this formulation does not look for shift invariant features.

2.2 Shift-invariant dictionary learning (SIDL)

In previous works, various shift-invariant dictionary learning (SIDL) methods have been employed to discover local patterns that are embedded across a longer time series in sequential data such as audio signals. While [Shift-Invariant Sparse Coding for Audio Classification] employs shift- invariant sparse coding with a convolutional optimization and gradient descent method for an audio classification task, [Efficient Shift-Invariant Dictionary Learning] demonstrates an efficient algorithm with the ability to combine shift-invariant patterns in a sparse coding of the original data for audio reconstruction and classification tasks. Such unsupervised learning methods have shown to be powerful in discovering shift-invariant patterns and a handful of studies have implemented SIDL for the purpose of music. Although music transcription and classification tasks have seen a strong usage of sparse dictionary learning in the past [Shift-Invariant Sparse Coding for Audio Classification, NMF based Dictionary Learning for Automatic Transcription of Polyphonic Piano Music, Sparse and Shift-Invariant Representations of Music, Music Genre Classification using On-line Dictionary Learning, Learning Sparse Dictionaries for Music and Speech Classification, Context-Dependent Piano Music Transcription With Convolutional Sparse Coding], we have yet to see a study that harnesses the advantages of sparse representation for the purpose of music creation. Instead, the popular methods for discovering music relations and achieving music generation have been a transformer with some sort of attention mechanism or the recurrent architectures. [Discovering Music Relations with Sequential Attention], for instance, uses an attention module that is tailored to the discovery of sequence level relations in music, while studies like [A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music] uses the recurrent variational autoencoder and a hierarchical decoder in order to model long-term musical structures.

2.3 Temporal Convolutional Networks (TCN)

The NLP4MusA style defines a printed ruler which should be presented in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document without the provided style files, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\aclfinalcopy` command in the document preamble.)

Benifits of TCN over RNNs (?)

- Low dimensionality feature extraction
- Music Simplification (**?**).
- Music Decomposition (**?**).
- Compressing music (**?**).

2.4 TCN K-WTA Autoencoders

In our study, we use a fully convolutional temporal autoencoder to find shift-invariant dictionaries while ensuring sparsity via the K-WTA activation function [Winner-Take-All Autoencoders]. The use of K-WTA in conjunction with dictionary learning was inspired by [Towards Contrastive Learning for Time-Series] where the K-WTA's ability to achieve sparse representations is explored in the context of constructive learning for time-series data. The use of Temporal Convolutional Nets was encouraged by various advantages that convolutional architectures bring for sequence modeling over recurrent networks as illustrated in [An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling]. Some of the most notable benefits include longer effective memory and low memory requirement when training. We explore these benefits for the purpose of music, which inherently requires a longer history due to musical temporal structure. Moreover, the low memory requirement of the convolutional architecture combined with a sparse representation in dictionary learning presents a strong potential for lighter and faster modeling with a high prospect of being applied to a real-time and on-line dictionary learning in the future. In this paper, we propose SIDL using TCN WTA-Autoencoders for discovering music relations—salient features of a specific performer or music, and illustrate potential applications in music analysis and creation.

3 Experiments

We show a few applications of our TCN k-WTA model

- Cleaning musical sections
- Unsupervised feature extraction
- Generating new music

Dataset	Size	Instrument	Encoding
MAESTRO (**?**)	1020 (Hrs)	Piano	One-hot encoding over 388 different MIDI events. Every datapoint here has an arbitrary length
Groove **?**	3.6 (Hrs)	Drum	T timesteps (one per 16th note) and 27 MIDI events. We use fixed length 64 time step sections

Table 1: Datasets used to experiment with fully convolutional temporal autoencoder model. All datasets used are MIDI format

3.1 Datasets

We use two distinct datasets: MAESTRO (**?**), and grove Groove (**?**) . See table 2 for more details on the datasets used. We also use distinct MIDI representations for each dataset.

3.2 Model Implementation

Our model implementation differs slightly for the different datasets used:

MAESTRO
Our TCN-KWTA Autoencoder is designed with [1, 8, 16, 32, 1000, 1] layers. The sparse representation is the layer before the last. Our WTA activation function is in the layer before the last. We also use a decaying WTA activation fucntions

GROOVE
Our TCN-KWTA Autoencoder is designed with [1, 8, 16, 32, 1000, 1] layers. The sparse representation is the layer before the last. Our WTA activation function is in the layer before the last. We also use a decaying WTA activation fucntions

3.3 Music Reconstructions

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. Use the two-column format only when you begin the abstract.

The title, author names and addresses should be completely identical to those entered to the electronical paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

3.4 Keep Top N Features

Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side.

Example applications

- Low dimensionality feature extraction
- Music Simplification (**?**).
- Music Decomposition (**?**).
- Compressing music (**?**).

3.5 Genrating Structured Drums

Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number sub-subsections.

3.6 Genrating Piano Music

Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

3.7 Unsupervised Stylistic Segmentation

Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections

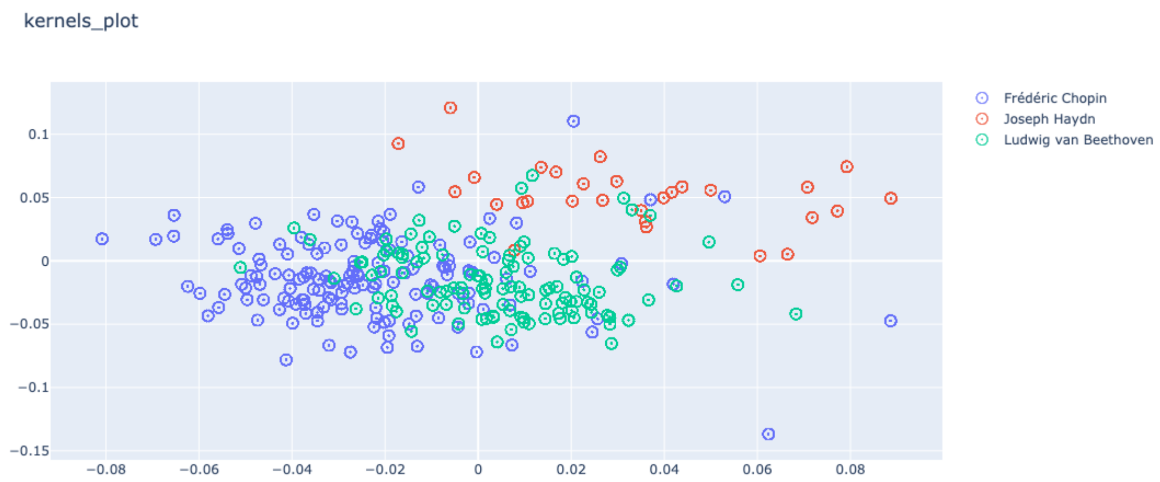

Figure 1: After training the model we can use it to encode datapoints of arbitrary length unsupervised stylistic segmentation. We use PCA on the average sparse code for each piece. We project into 2 dimensional sparse to visualize

with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number sub-subsections.

- Example citing an arxiv paper: (**?**).
- Example article in journal citation: (**?**).
- Example article in proceedings, with location: (**?**).
- Example article in proceedings, without location: (**?**).

See corresponding .bib file for further details.

Submissions should accurately reference prior and related work, including code and data. If a piece of prior work appeared in multiple venues, the version that appeared in a refereed, archival venue should be referenced. If multiple versions of a piece of prior work exist, the one used by the authors should be referenced. Authors should not rely on automated citation indices to provide accurate references for prior and related work.

4 Discussion

In an effort to accommodate people who are color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. A simple criterion: All curves and points in your figures should be clearly distinguishable without color.

5 Conclusion

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration “translation”.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.