klustr: dimensionality reduction and visualization of large audio datasets

|  |  |
| --- | --- |
|  | Lamtharn Hantrakul, Avneesh Sarwate |
|  | MT Hauz  1065 Terrell St.  Georgia Tech Center for Music Technology  Atlanta, GA 30318 |

ABSTRACT

We present *klustr*: a tool for automatic dimensionality reduction and 2D visualization of large audio datasets. The tool facilitates visual navigation and discovery of similar sounding samples in a collection, allowing musicians and researchers to quickly find sounds that are similar to one another based on perceptual features such as timbre. Normally, the process of navigating a large collection of samples involves painstakingly traversing line by line through a list of filenames that may be tagged with appropriate titles such as “kick\_1.wav”, or may not be tagged at all. At best, a taxonomy and ontology of labels may be provided as part of the library. *Klustr* forms relationships between samples based on the audio file and not high level descriptors. In this paper, we present a version of *Klustr* that is optimized for navigating drum samples typically found in pop, hip-hop and electronic music.

1. INTRODUCTION

From speech, images and sensor input, data from the natural world is often high dimensional. Instead of interacting directly with this high dimensional data however, a collection of audio samples for example, is often instead organized using simple high level descriptors, such as the type of sound e.g “vocal\_shout”. However, these labels are often not available, and when available, do not capture the nuances of relationships between sounds. How different is “vocal\_shout\_1” from “vocal\_shout\_2”? What about “vocal\_male\_3”? In order to build relationships between samples at the timbral level, we must form representations drawn from the high dimensional audio data. These are typically in the form of STFT or MFCC features, but even these are in dimensions in the order of 100.

Dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) are often used to extract key components that capture the distribution of the data using a smaller number of dimensions. This reduced data often yields better performance in classic machine learning algorithms such as Support Vector Machines (SVM) and Naïve Bayes [].

Alternatively, dimensionality reduction can be used to visualize the representation learned by non-linear models such as neural networks. A highly popular stochastic dimensionality reduction algorithm called t-distributed neighbor embedding (TSNE) [] has become the standard tool for visualizing the embeddings learnt by deep convolutional neural networks []. TSNE is particularly adept at preserving spatial relationships in higher dimensions but in 2 or 3 dimensions. This enables researchers to get a better intuition of the inner workings of their model and tune hyper-parameters accordingly.

In this paper, we can use these dimensionality reduction techniques to directly encode features into a 2 dimensional plane. Our purpose is not to find the best set of features for a classifier, nor to understand the internal representation of neural network. Instead, we are attempting to take advantage of dimensionality reduction to generate a navigable 2D map where neighbouring samples are similar in sounds e.g a raspy snare is located next to another gritty snare, but is further away from a kick drum.

1. Related work

Kernalized Sorting

AudioQuilt

1. PAGE size

The proceedings will be printed on portrait A4-size paper (21.0cm x 29.7cm). All material on each page should fit within a rectangle of 17.2cm x 25.2cm, centered on the page, beginning 2.0cm from the top of the page and ending with 2.5cm from the bottom. The left and right margins should be 1.9cm. The text should be in two 8.2cm columns with a 0.8cm gutter. All textmust be in a two-column format. Text must be fully justified.

1. Algorithm Overview and Description
   1. Normal or Body Text

Please use a 10pt (point) Times font. Sans-serif or non-proportional fonts can be used only for special purposes, such as distinguishing source code text.

The first paragraph in each section should not be indented, but all other paragraphs should be.

* 1. Title and Authors

The title is 14pt Times, bold, caps, upper case, centered. Authors’ names are omitted when submitting for double-blind reviewing. The following is for making a camera-ready version. Authors’ names are centered. The lead author’s name is to be listed first (left-most), and the co-authors’ names after. If the addresses for all authors are the same, include the address only once, centered. If the authors have different addresses, put the addresses, evenly spaced, under each authors’ name.

http://i.creativecommons.org/l/by/3.0/88x31.png © First author, Second author, Third author.

Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** First author, Second author, Third author. “Paper Template For ISMIR 2017”, 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

* 1. First Page Copyright Notice

Please include the copyright notice exactly as it appears here in the lower left-hand corner of the page. It is set in 8pt Times. After your paper is accepted, you will need to insert the appropriate author names and paper title in the copyright notice when submitting the camera-ready version. For LATEX users, this will be handled by the template automatically. For Word users, this has to be done manually.

* 1. Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

1. Evaluation

First level headings are in Times 10pt bold, centered with 1 line of space above the section head, and 1/2 space below it. For a section header immediately followed by a subsection header, the space should be merged.

* 1. Second Level Headings

Second level headings are in Times 10pt bold, flush left, with 1 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

* + 1. Third Level Headings

Third level headings are in Times 10pt italic, flush left, with 1/2 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

Using more than three levels of headings is highly discouraged.

1. COnclusion
   1. Footnotes

Indicate footnotes with a number in the text.[[1]](#footnote-1) Use 8pt type for footnotes. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a 0.5pt horizontal rule.

|  |  |
| --- | --- |
| String value | Numeric value |
| Hello ISMIR | 2017 |

**Table 1.** Table captions should be placed below the table.

figure

**Figure 1**. Figure captions should be placed below the figure.

* 1. Figures, Tables, and Captions

All artwork must be centered, neat, clean, and legible. All lines should be very dark for purposes of reproduction and art work should not be hand-drawn. The proceedings are not in color, and therefore all figures must make sense in black-and-white form. Figure and table numbers and captions always appear below the figure. Leave 1 line space between the figure or table and the caption. Each figure or table is numbered consecutively. Captions should be Times 10pt. Place tables/figures in text as close to the reference as possible. References to tables and figures should be capitalized, for example: see Figure 1 and Table 1. Figures and tables may extend across both columns to a maximum width of 17.2cm.

1. Equations

Equations should be placed on separate lines and numbered. The number should be on the right side, in parentheses, as in Eqn (1).

 (1)

1. Citations

All bibliographical references should be listed at the end, inside a section named “REFERENCES,” numbered and in alphabetical order according to IEEE formatting. All references listed should be cited in the text. When referring to a document, type the number in square brackets [1], or for a range [1-4].

When the following words appear in the conference publication titles, please abbreviate them: Proceedings -> Proc.; Record -> Rec.; Symposium -> Symp.; Technical Digest -> Tech. Dig.; Technical Paper -> Tech. Paper; First -> 1st; Second -> 2nd; Third -> 3rd; Fourth/nth -> 4th/nth.

As submission is double blind, refer to your own published work in the third person. That is, use “In the previous work of [2],” not “In our previous work [2].” If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form “A. Anonymous.”

1. REFERENCES
2. E. Author: “The Title of the Conference Paper,” *Proceedings of the International Symposium on Music Information Retrieval*, pp. 000–111, 2000.
3. A. Someone, B. Someone, and C. Someone: “The Title of the Journal Paper,” *Journal of New Music Research*, Vol. A, No. B, pp. 111–222, 2010.
4. X. Someone and Y. Someone: Title of the book, Editorial Acme, Porto, 2012.

1. This is a footnote. [↑](#footnote-ref-1)