FMA: A DATASET FOR MUSIC ANALYSIS (ISMIR 2017)

张展鹏

2021 年 6 月 2 日

为什么音乐分析数据库是必要的?

MIR == Music Information Retrieval

- 特征和端到端学习需要大型音乐数据集。
- 大型音乐数据集需要实现浏览、搜索和组织等功能。
- Free，源代码和数据集均开放；Archive，数据集规模很大。

FMA 的特点

FMA == Free Music Archive

- Free，源代码和数据集均开放；Archive，数据集规模很大。
- FMA 提供了 917GB 的数据集和 343 天的知识共享许可。
- 这些音频来自 16341 位艺术家的 106574 首曲目和 14854 张专辑，按 161 种流派分层分类。
- FMA 提供了完整长度和高质量的音频及特征，以及级元数据、标签和简介文本等数据。

论文工作简述

- 对比现有工作，实现 FMA 相关系统设计。
- 为 MIR 任务提供训练、验证和测试集并评估效果。
- 对音乐进行流派识别。

FMA 跟同类音乐数据库比较

dataset[1]	#clips	#artists	year	audio
RWC [12]	465	-	2001	yes
CAL500 [45]	500	500	2007	yes
Ballroom [13]	698	-	2004	yes
GTZAN [46]	1,000	∼ 300	2002	yes
MusiClef [36]	1,355	218	2012	yes
Artist20 [7]	1,413	20	2007	yes
ISMIR2004	1,458	-	2004	yes
Homburg [15]	1,886	1,463	2005	yes
103-Artists [30]	2,445	103	2005	yes
Unique [41]	3,115	3,115	2010	yes
1517-Artists [40]	3,180	1,517	2008	yes
LMD [42]	3,227	-	2007	no
EBallroom [23]	4,180	-	2016	no[2]
USPOP [1]	8,752	400	2003	no
CAL10k [44]	10,271	4,597	2010	no
MagnaTagATune [20]	25,863[3]	230	2009	yes[4]
Codaich [28]	26,420	1,941	2006	no
FMA	**106,574**	**16,341**	**2017**	**yes**
OMRAS2 [24]	152,410	6,938	2009	no
MSD [3]	1,000,000	44,745	2011	no[2]
AudioSet [10]	2,084,320	-	2017	no[2]
AcousticBrainz [32]	2,524,739[5]	-	2017	no

[1] Names are clickable links to datasets' homepage.
[2] Audio not directly available, can be downloaded from ballroomdancers.com, 7digital.com, youtube.com.
[3] The 25,863 clips are cut from 5,405 songs.
[4] Low quality 16 kHz, 32 kbit/s, mono mp3.
[5] As of 2017-07-14, of which a subset has been linked to genre labels for the MediaEval 2017 genre task.

Table 1: Comparison between FMA and alternative datasets.

音乐分析数据库的需求 (1)——数据规模大

音乐分析数据库需要大规模数据集来避免过拟合。

- 一般地，样本容量越大，拟合结果越好。
- 大规模数据更接近真实，数据更平衡。数据的缺点包含噪声以及其他可能被模型利用的、跟基本事实混淆的特征。

FMA 的数据规模足够大。

音乐分析数据库的需求 (2)——版权

一般地，唱片公司会对音乐加以严格的版权限制。

- CC == Creative Commons (license)，知识共享许可协议。
- BY == Attribution，署名；您（用户）可以复制、发行、展览、表演、放映、广播或通过信息网络传播本作品；您必须按照作者或者许可人指定的方式对作品进行署名。
- 4.0，发布日期是 2013 年 11 月 25 日；它不需要移植就可以适用于各地的法律，4.0 版并不鼓励移植，而是希望能作为一个全球通用的许可方式。
- MIT，作者是麻省理工学院；与其他常见的软件许可协议（如 GPL、LGPL、BSD）相比，它是相对宽松的软件许可协议。

FMA 收集版权许可的音轨并重新分发。数据在 CC BY 4.0 下获得许可；源代码在 MIT 下获得许可。

音乐分析数据库的需求 (3)——音频可下载

我们希望音乐分析数据库足够大，同时有原始音频可供自由分析。

- 较小的数据集通常会分发音频，大多数较大的数据集不会。
- 某些数据集仅分发部分从音频提取的特征，它们的用途受到了限制。
- 某些数据集仅提供下载链接，这样的下载资源不受控制，可能被移除。

FMA 较新，直接提供了音频。

音乐分析数据库的需求 (4)——音频质量好

音频质量有剪辑时长和采样率这 2 个主要指标。

- 其他音乐分析数据库通常提供 10-30s 的剪辑，这会导致同一首曲目的不同剪辑可能导致不一样的预测结果；另外，研究人员难以自由截取音频片段。
- 某些音乐分析数据库的比特率低至 32kbit/s；主流数据库，例如 MSD，比特率为 104kbit/s。

FMA 的文件名称是音轨 ID，以 mp3 格式编码。大多数采样率为 44100Hz，比特率为 320kbit/s（平均为 263kbit/s），并且是立体声的。

音乐分析数据库的需求 (5)——元数据丰富

FMA 在元数据覆盖率上具有显著优势。

100% track_id	100% title	93% number
2% information	14% language_code	100% license
4% composer	1% publisher	1% lyricist
98% genres	98% genres_all	47% genre_top
100% duration	100% bit_rate	100% interest
100% #listens	2% #comments	61% #favorites
100% date_created	6% date_recorded	22% tags
100% album_id	100% title	
94% type	96% #tracks	
76% information	16% engineer	18% producer
97% #listens	12% #comments	38% #favorites
97% date_created	64% date_released	18% tags
100% artist_id	100% name	25% members
38% bio	5% associated_labels	
43% website	2% wikipedia_page	
	5% related_projects	
37% location	23% longitude	23% latitude
11% #comments	48% #favorites	10% tags[1]
99% date_created	8% active_year_begin	
	2% active_year_end	

[1] One of the tags is often the artist name. It has been subtracted.

Table 2: List of available per-track, per-album and per-artist metadata, i.e. the columns of `tracks.csv`. Percentages indicate coverage over all tracks, albums, and artists.

音乐分析数据库的需求 (6)——容易获取

FMA 易于访问。它的数据集只包含 CSV 格式记录的元数据和和 MP3 格式的音频，可以直接下载。

音乐分析数据库的需求 (7)——数据存档、易配置

FMA 的全体文件和档案都经过校验并存档托管；公开了用于

- (i) 收集数据，
- (ii) 分析数据，
- (iii) 生成子集和拆分，
- (iv) 计算特征，
- (v) 测试基线

的所有源代码。这些源代码容易被修改以被研究人员用于计算自己的特征和评估自己的方法。另外，依赖于公共曲库和 API，任何人都可以重新创建或扩展曲库。

FMA 的规模

- 截止 2017 年 4 月 1 日，最大音轨 ID 高达 155320，其中 109727 首是有效的，丢失的 45594 首可能对应已删除的曲目。
- 在 109727 首有效的音轨中，180 首无法下载，286 首无法被 ffmpeg 剪辑，71 首无法用以提取特征，2616 首的许可证禁止重新分发，剩下 106574 首可以自由使用。
- FMA 未人为筛选音轨，以去除流派过多、过长、和稀有流派等曲目。这样做是为了使音轨分布接近真实情况。

FMA 的成长

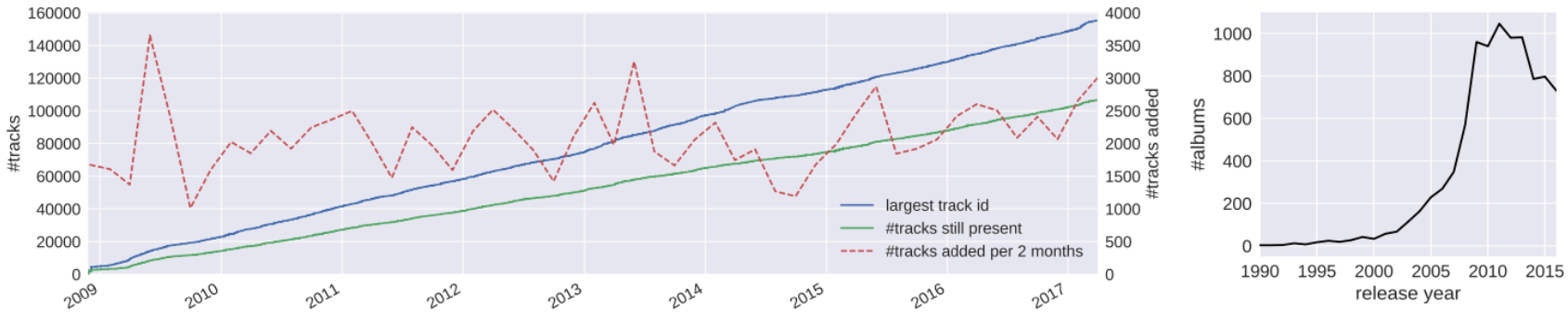

Figure 1: (left) Growth of the archive, created in 11/2008. (right) Number of albums released per year (min 1902, max 2017).

依赖元数据构建关系型数据库

元数据被清理和格式化，构建关系型数据库。

track						album			artist	
track_id	title	genres_all	genre_top	dur.	listens	title	listens	tags	name	location
150073	Welcome to Asia	[2, 79]	International	81	683	Reprise	4091	[world music, dubtronica, fusion]	DubRaJah	Russia
140943	Sleepless Nights	[322, 5]	Classical	246	1777	Creative Commons Vol. 7	28900	[classical, alternate, soundtrack, piano, ...	Dexter Britain	United Kingdom
64604	i dont want to die alone	[32, 38, 456]	Experimental	138	830	Summer Gut String	7408	[improvised, minimalist, noise, ...	Buildings and Mountains	Oneonta, NY
23500	A Life In A Day	[236, 286, 15]	Electronic	264	1149	A Life in a Day	6691	[idm, candlestick, romanian, candle, ...	Candlestickmaker	Romania
131150	Yeti-Bo-Betty	[25, 12, 85]	Rock	124	183	No Life After Crypts	3594	[richmond, fredericksburg, trash rock, ...	The Crypts!	Fredericksburg

Table 3: Some rows and columns of the metadata table, stored in `tracks.csv`.

音轨时长和播放量分布

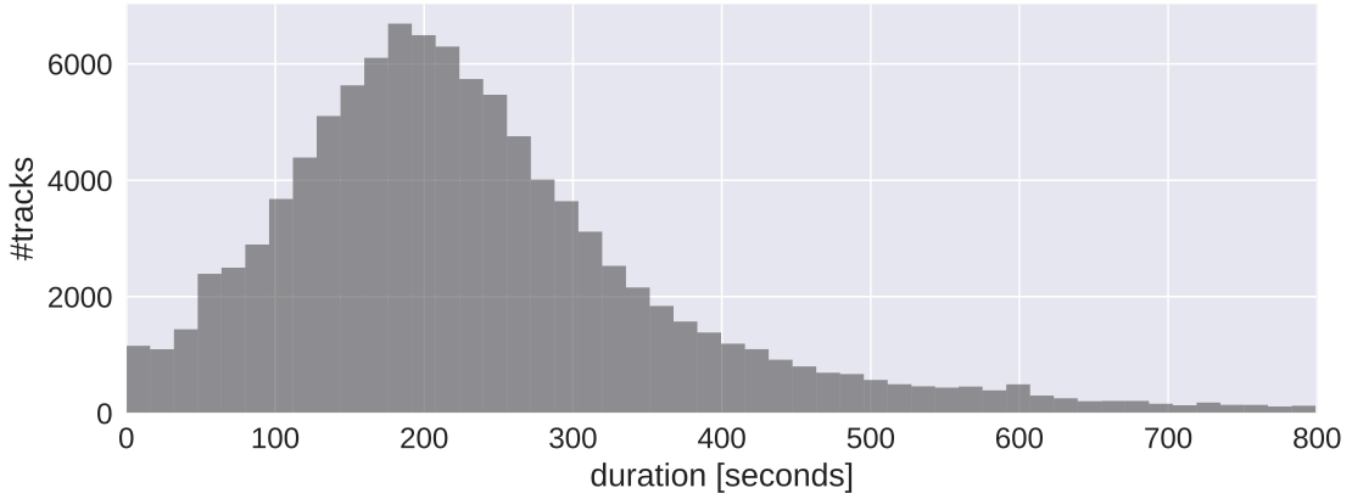

Figure 2: Track duration (min 0, max 3 hours).

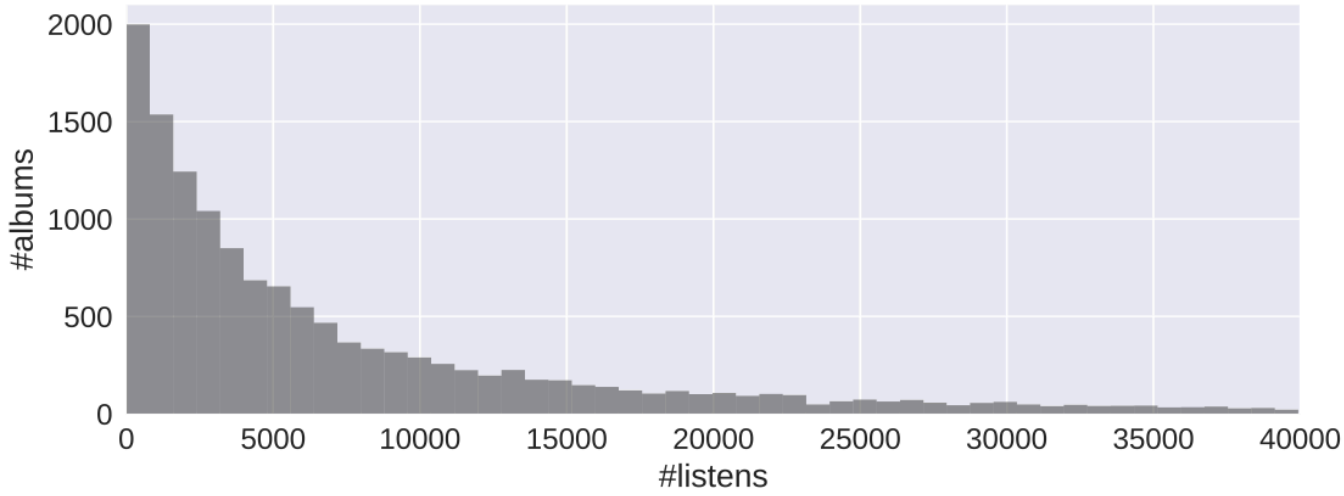

Figure 3: Album listens (min 0, max 3.6 millions).

流派分类

FMA 具有精细的流派信息——内置的层次结构，并且由艺术家自行注释；也可以由众包和专家来补充。

- 标签噪声是不可避免的。
- 流派标签的层次结构由 161 个流派组成，其中 16 个是根流派，其他是子流派。

流派分层分类举例 (1)

id	parent	top_level	title	#tracks
38	None	38	Experimental	38,154
15	None	15	Electronic	34,413
12	None	12	Rock	32,923
1235	None	1235	Instrumental	14,938
25	12	12	Punk	9,261
89	25	12	Post-Punk	1,858
1	38	38	Avant-Garde	8,693

Table 4: An excerpt of the genre hierarchy, stored in `genres.csv`. Some of the 16 top-level genres appear in the top part, while some second- and third-level genres appear in the bottom part.

流派分层分类举例 (2)

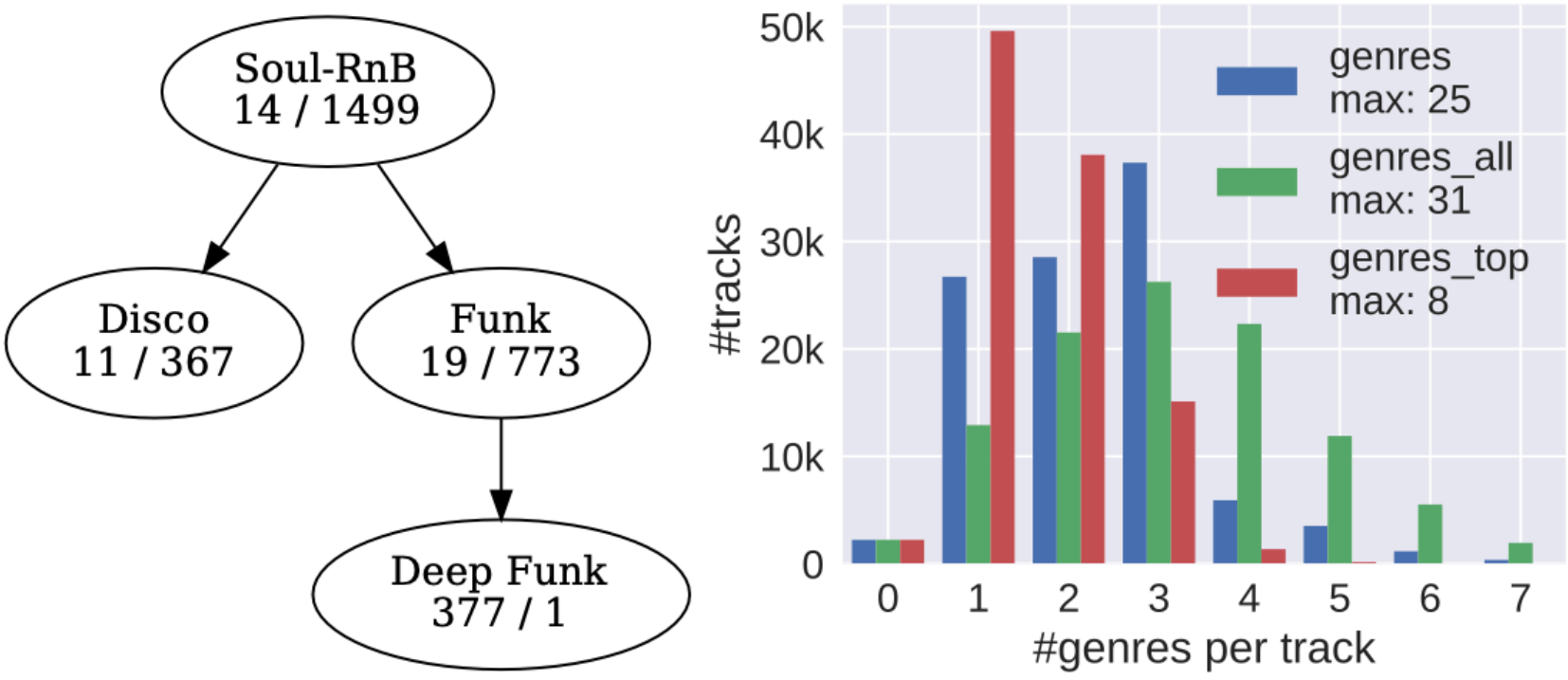

Figure 5: (left) Example of genre hierarchy for the top-level Soul-RnB genre. Left number is the `genre_id`, right is the number of tracks per genre. (right) Number of genres per track. A 3 genres limit has been introduced early on by the administrators.

特征提取——流派分类统计

流派分类任务是预先进行的。

Each feature set (except zero-crossing rate) is computed on windows of 2048 samples spaced by hops of 512 samples. Seven statistics were then computed over all windows: the mean, standard deviation, skew, kurtosis, median, minimum and maximum.

平均值、标准偏差、偏斜、峰度、中值、最小值和最大值

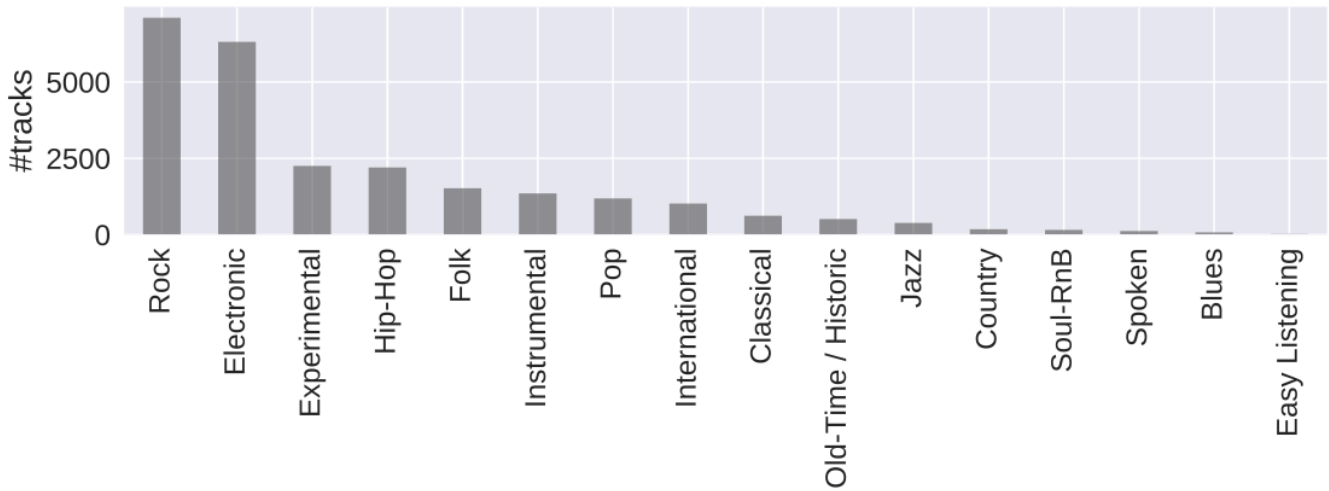

Figure 6: (top) Tracks per (sub-)genre on the full set (min 1, max 38,154). (bottom) Tracks per all 16 root genres on the medium subset (min 21, max 7,103). Note how experimental music is much less represented in the curated medium subset.

子数据集——Small, medium, large and full

dataset	clips	genres	length	size	
			[s]	[GiB]	#days
small	8,000	8	30	7.4	2.8
medium	25,000	16	30	23	8.7
large	106,574	161	30	98	37
full	106,574	161	278	917	343

Table 5: Proposed subsets of the FMA.

数据集划分与抽样方法

经典分割：训练集: 验证集: 测试集 = 8:1:1。为了使使用 FMA 的结果可重现，如果使用交叉验证，则应合并训练集和验证集。数据集划分应该满足下列约束条件：

- 每个根流派均保证在所有分组（split）中表示。
- 较小的流派子集会满足 8:1:1 的划分比例；但不保证最小的 7 个子流派出现在所有分组中。
- 因为同一个艺术家的歌曲同时出现在训练和测试集会导致分类准确性比实际表现得乐观，所以同一个艺术家的歌曲只会出现在单个分组中。
- 没有流派标签的 2231 首歌曲会被完全分配给训练集用于搬家度学习，以及作为额外的训练样本。

FMA 的用途举例

- 音乐分类和注释。包括流派识别、艺术家识别、年份预测和特点自动标记等。
- 流派识别。音乐流派是通过文化、艺术家和市场力量的复杂相互作用而产生的类别，用于表征作品之间的相似性。
- 数据分析。这一步主要指分析音频。FMA 的完整曲目可用性允许对音乐属性进行适当的研究，例如音乐结构分析；元数据是对现有数据集的有价值补充，用以元数据分析。

FMA 设计与实现的结论

- FMA 是一个可以在研究人员之间轻松共享的数据集。
- FMA 允许用户对数据进行自由的定义和拆分。
- FMA 的样本容量足够大，这意味着 FMA 提供的样本几乎不会出现不平衡的问题，足够接近真实情况。
- FMA 适合用以音乐流派识别，即 MGR。
- 曲库也是知识库，共享是第一位的，这会方便教育和研究。

谢谢！