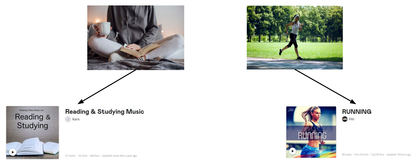
Should we consider the users in contextual music auto-tagging models?

Karim M. Ibrahim, Elena Epure, Geoffroy Peeters, Gaël Richard

karim.ibrahim@telecom-paris.fr

https://github.com/KarimMibrahim/user-aware-music-autotagging

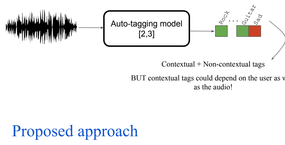
The listening context strongly
influence the listening preferences [1]



The Question: Would considering the user when auto-tagging tracks with "contextual" tags improve the performance?

Contextual tags ⟶ Tags that describe the listening situation of the users (e.g. location, activity, time).

Classic approach



Contextual + Non-contextual tags

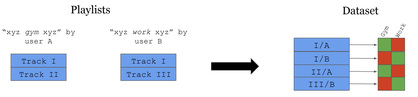
BUT contextual tags could depend on the user as well as the audio!

Proposed approach

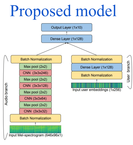
1. Collect a dataset of tracks tagged with different contexts by different users.
2. Train two auto-taggers, one using only audio and one using audio+user information.
3. Evaluate the two model using different user-focused evaluation protocols.

How to collect the dataset?

"Through the user-created playlists"

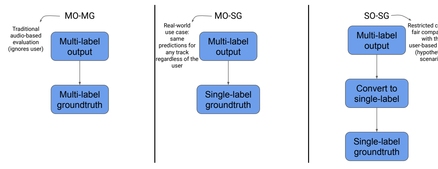


Proposed model



Evaluation Scenarios

We need to compare the **multi-label** audio model to the **single-label** user model.
Hence, we convert the audio model output using different scenarios to compare.



User-satisfaction evaluation

We also propose a user-satisfaction protocol to evaluate on each user independently

We define it as:

$$S_{\mathbb{U}} = \frac{1}{N} \sum_{u \in \mathbb{U}} S_u, \text{ where } N = |\mathbb{U}|$$

Evaluation Metric

$$S_u = f(G_u, P_u)$$

User's Groundtruth matrix

$$G_u = \{0,1\}^{n_u \times m_u}$$

User's Predictions matrix

$$P_u = \{0,1\}^{n_u \times m_u}$$

n_u Present tracks for user u

m_u Present contexts for user u

Evaluation results

MO-MG

	AUC	Recall	Precision	f1-score
car	0.56	0.97	0.48	0.64
gym	0.73	0.91	0.57	0.7
happy	0.58	0.97	0.37	0.53
night	0.6	0.99	0.49	0.65
relax	0.79	0.88	0.59	0.7
running	0.66	0.95	0.54	0.69
sad	0.78	0.85	0.48	0.61
summer	0.6	1	0.61	0.76
work	0.54	1	0.47	0.64
workout	0.76	0.89	0.49	0.63
average	0.66	0.94	0.53	0.66

MO-SG

	AUC	Recall	Precision	f1-score
car	0.545	0.973	0.088	0.162
gym	0.677	0.934	0.137	0.238
happy	0.563	0.975	0.066	0.124
night	0.575	0.993	0.095	0.173
relax	0.740	0.926	0.163	0.277
running	0.612	0.957	0.118	0.210
sad	0.741	0.894	0.137	0.237
summer	0.577	1.000	0.156	0.270
work	0.526	1.000	0.085	0.156
workout	0.717	0.913	0.107	0.192
average	0.627	0.957	0.115	0.204

SO-SG

	AUC	Recall	Precision	f1-score
car	0.545	0.000	0.000	0.000
gym	0.677	0.378	0.191	0.245
happy	0.563	0.000	0.000	0.000
night	0.575	0.000	0.051	0.001
relax	0.740	0.639	0.241	0.350
running	0.612	0.039	0.178	0.064
sad	0.741	0.001	0.417	0.002
summer	0.577	0.414	0.192	0.262
work	0.526	0.000	0.105	0.001
workout	0.717	0.186	0.193	0.189
average	0.627	0.166	0.156	0.111

User+Audio

	AUC	Recall	Precision	f1-score
car	0.62	0.11	0.15	0.13
gym	0.73	0.15	0.24	0.18
happy	0.64	0.22	0.12	0.15
night	0.62	0.03	0.15	0.05
relax	0.77	0.43	0.31	0.36
running	0.7	0.28	0.22	0.24
sad	0.84	0.54	0.35	0.42
summer	0.66	0.2	0.28	0.23
work	0.59	0.02	0.13	0.04
workout	0.75	0.4	0.2	0.26
average	0.69	0.24	0.22	0.21

Evaluation results

User-satisfaction evaluation

	Accuracy	Recall	Precision	f1-score
Audio	0.21	0.204	0.243	0.216
Audio+User	0.254	0.246	0.295	0.26

Conclusion: We **do** need to consider the users in contextual auto-tagging

References

[1] North, Adrian C., and David J. Hargreaves. "Situational influences on reported musical preference." *Psychomusicology: A Journal of Research in Music Cognition* 15.1-2 (1996): 30.

[2] Keunwoo Choi, George Fazekas, and Mark Sandler,"Automatic tagging using deep convolutional neural networks,"arXiv preprint arXiv:1606.00298, 2016.

[3] Pons, Jordi, et al. "End-to-end learning for music audio tagging at scale." *arXiv preprint arXiv:1711.02520* (2017).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068