Toward interpretable polyphonic SED with attention maps based on local prototypes

Pablo Zinemanas, Martín Rocamora, Eduardo Fonseca, Frederic Font, and Xavier Serra

upf. — **M**usic **T**echnology **G**roup · DCASE 2021 Workshop · UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Problems of black-box models

- Reinforce inequality and bias
- Vulnerable to adversarial attacks
- Hard to debug

Interpretable machine learning

- Post-hoc explanations — Black box / Proxy model
- Inherently interpretable models
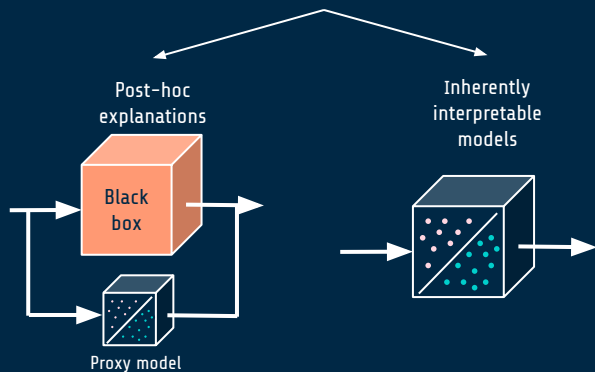
Proposed interpretable model

Mel-spectrogram X — Encoder f(·) — Latent space Z F C T — Decoder g(·) — $\tilde{X}$

Encoder Attention s(·)

Prototypes P_1 ... P_M C C

Attention maps S F M T — * — Reconstructed latent space $\hat{Z}$ F C T — Decoder g(·) — $\hat{X}$

Mean time dimension $\bar{S}$ F M — Fully-connected layer — Output K
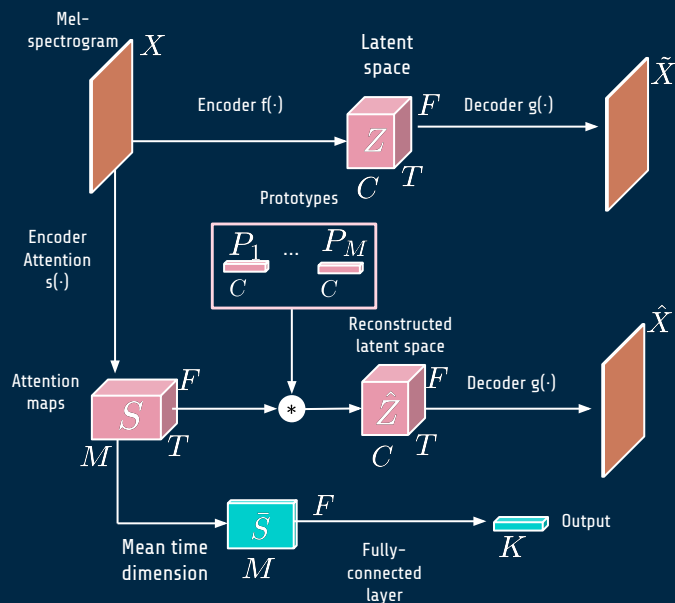
Explanations in form of attention maps
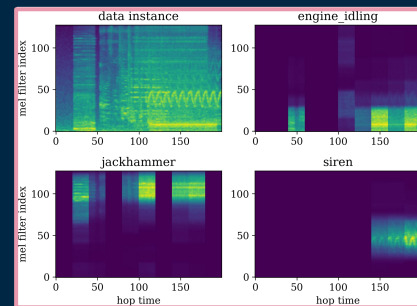
Results comparable to black-box baselines

F1 (%)

- CNN: 57,3
- Openl3+MLP: 58,2
- Ours: 58,8