Who calls the shots? Rethinking Few-Shot Learning for Audio

Yu Wang[1], Nicholas J. Bryan[2], Justin Salamon[2], Mark Cartwright[3], Juan Pablo Bello[1]

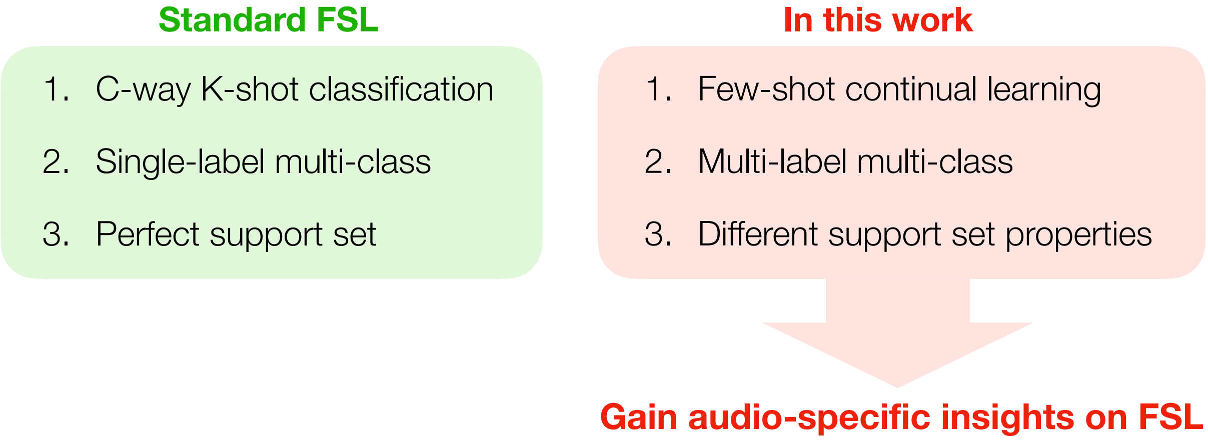
[1]Music and Audio Research Laboratory, New York University
[2]Adobe Research
[3]Department of Informatics, New Jersey Institute of Technology

MARL NYU Adobe NJIT

⌂ y-wang.weebly.com
✉ wangyu@nyu.edu

Goal: Audio-Specific Insights on FSL

- Learn to recognize a **new class** based on only **few examples** (the support set)



Standard FSL

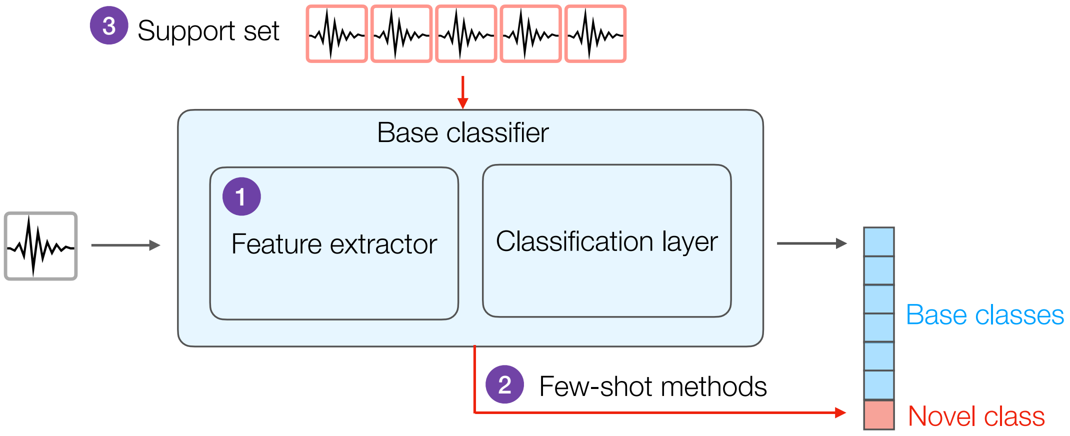
1. C-way K-shot classification
2. Single-label multi-class
3. Perfect support set

In this work

1. Few-shot continual learning
2. Multi-label multi-class
3. Different support set properties

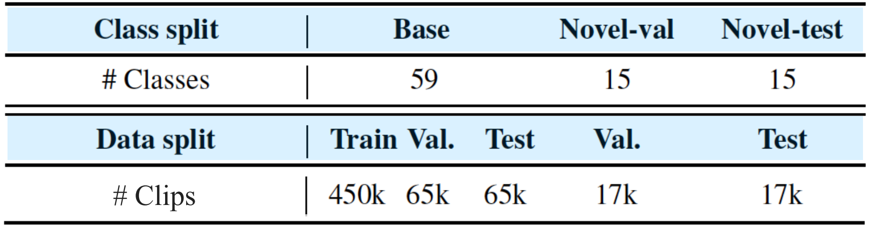
Gain audio-specific insights on FSL

Experiments



Dataset: FSD-MIX-CLIPS

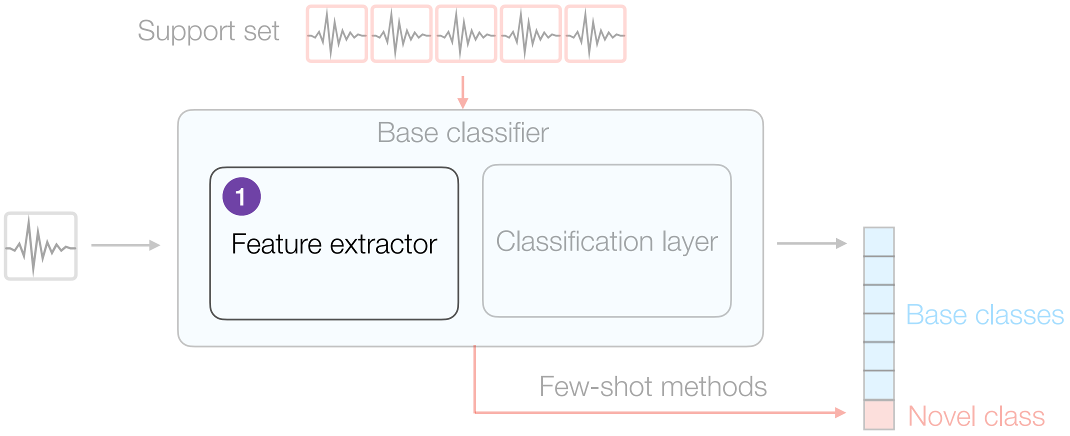
1. **Large in size**: ~615k 1s soundscapes
2. **Perfect annotation**: Programmatically-mixed using *Scaper*
   - Foreground: Single-labeled short clips in *FSD50k*
   - Background: Brownian noise
3. **Controlled acoustic properties**: Polyphony, SNR
4. **Freely available** on *Zenodo* (see github.com/wangyu/rethink-audio-fsl)



Class split	Base			Novel-val	Novel-test
# Classes	59			15	15
Data split	**Train**	**Val.**	**Test**	**Val.**	**Test**
# Clips	450k	65k	65k	17k	17k

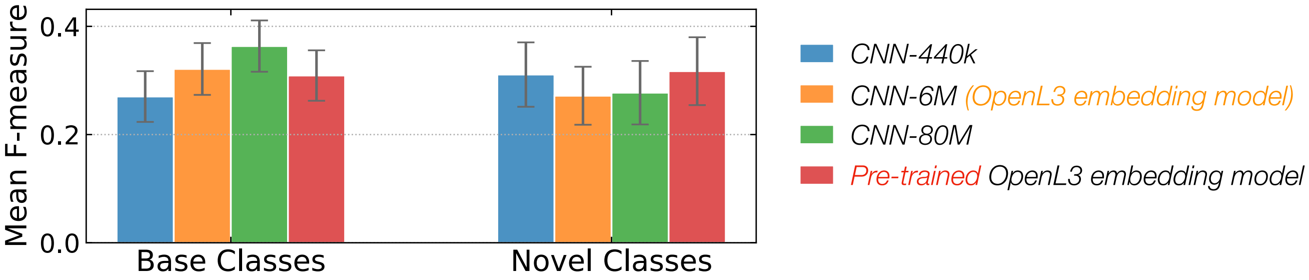
- Disjoint sets of *base* and *novel* classes
- Train on base data only
- Evaluate on base & novel data

Experiment: Feature Extractor



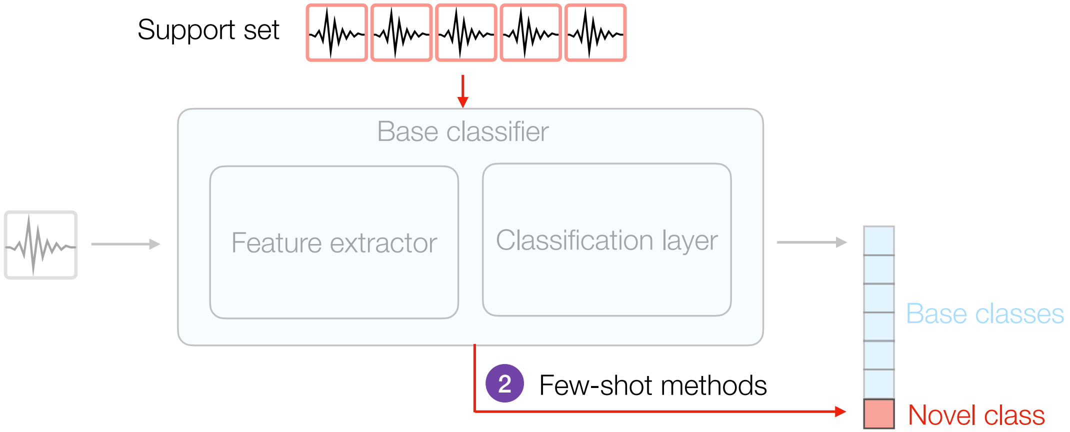
- 3 different models: *CNN-440k, CNN-6M, CNN-80M*
- Pre-training

Experiment: Feature Extractor



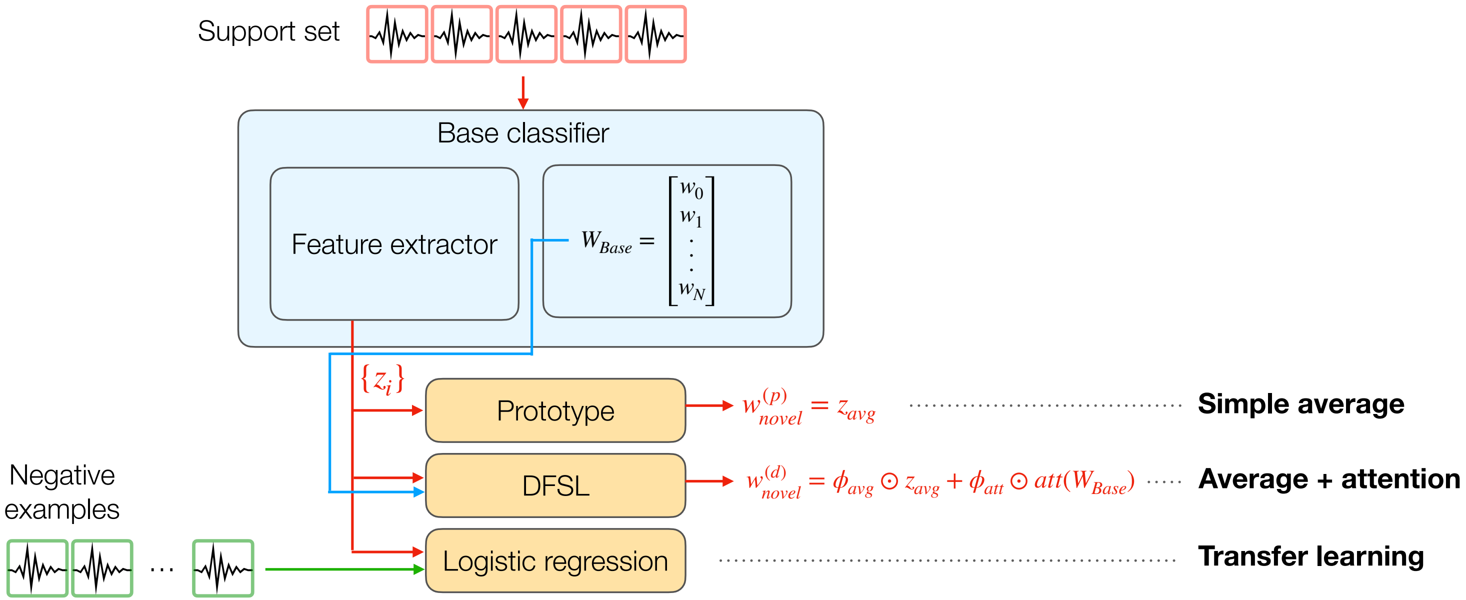
1. Trade-off between overfitting base classes and generalizing to novel classes.
2. Pre-trained OpenL3 model: best novel performance
   balance between base & novel classes

Experiment: Few-Shot Methods



- 3 few-shot methods
- Different number of support examples

Experiment: Few-Shot Methods



Support set

Base classifier

Feature extractor

$W_{Base} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_N \end{bmatrix}$

$\{z_i\}$

Prototype

DFSL

Logistic regression

Negative examples

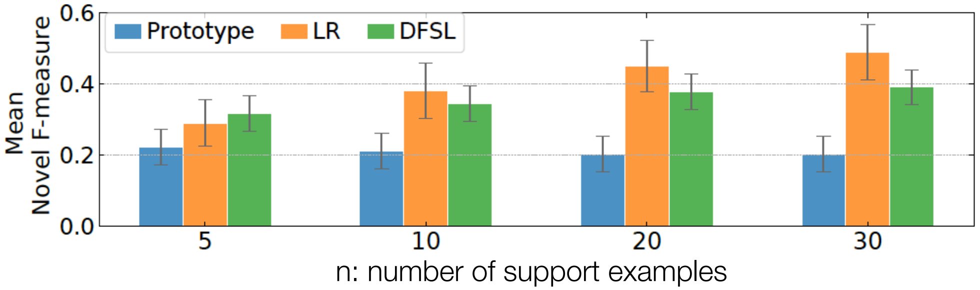
…

$w_{novel}^{(p)} = z_{avg}$ ………………………………… **Simple average**

$w_{novel}^{(d)} = \phi_{avg} \odot z_{avg} + \phi_{att} \odot att(W_{Base})$ ….. **Average + attention**

……………………………………………………… **Transfer learning**

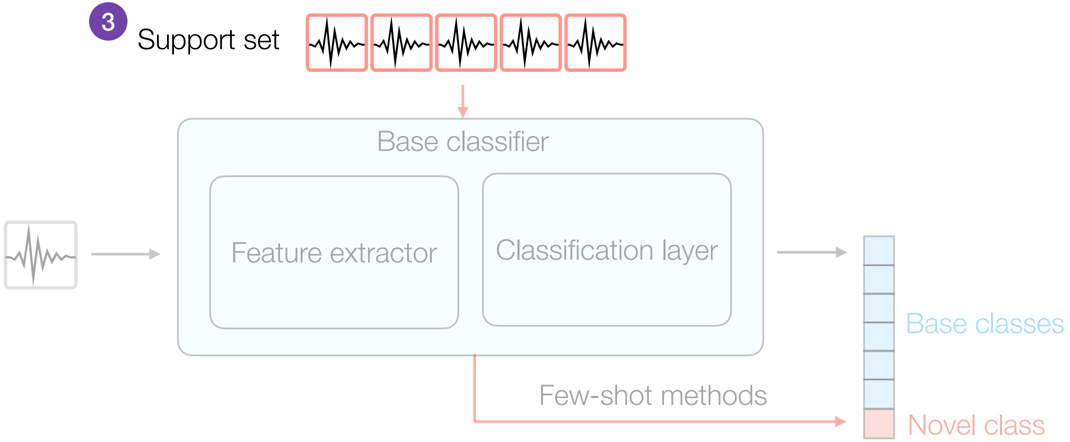
Experiment: Few-Shot Methods



1. n = 5: DFSL
2. n >= 10: LR
3. To minimize user labeling effort and runtime resources ⟶ DFSL + small n

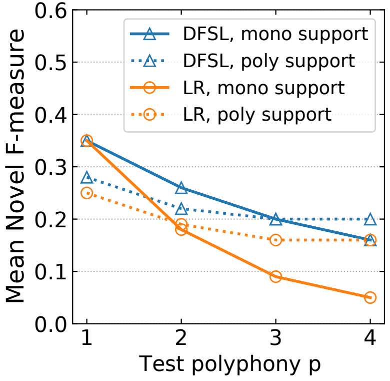
   otherwise ⟶ LR + large n

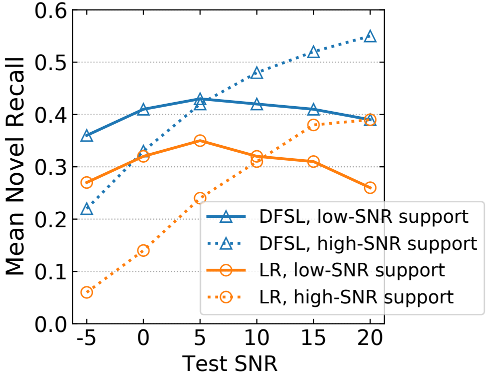
Experiment: Support Set Selection



- *Monophonic* vs. *Polyphonic*
- *High-SNR* vs. *Low-SNR*
- n = 5

Experiment: Support Set Selection

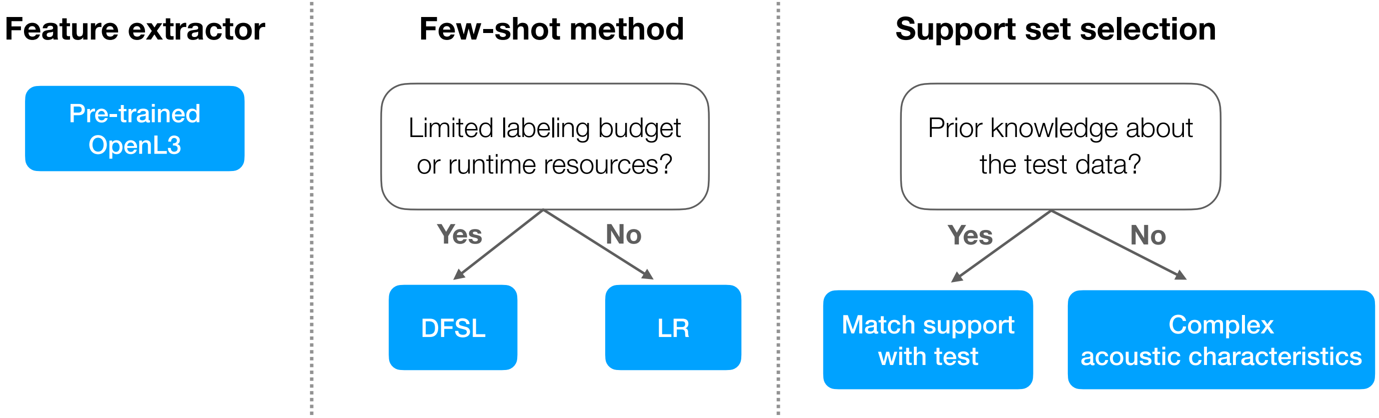




1. If we know test sample characteristics: matching those in the support set
2. If not: select support examples with more complex acoustics characteristics

Recap

- **Audio-specific recipe** for few-shot continual leaning on multi-label classification



Feature extractor

Pre-trained OpenL3

Few-shot method

Limited labeling budget or runtime resources?

Yes → DFSL

No → LR

Support set selection

Prior knowledge about the test data?

Yes → Match support with test

No → Complex acoustic characteristics

- Code, FSD-MIX-CLIPS dataset, FSD-MIX-SED dataset: `github.com/wangyu/rethink-audio-fsl`

Thanks! :)