Fairness and underspecification in acoustic scene classification
The case for disaggregated evaluations

Poster ID: 29

A. Triantafyllopoulos[12], M. Milling[2], K. Drossos[3], B.W. Schuller[124]
[1]audEERING, [2]EIHW, University of Augsburg, [3]Audio Research Group, Tampere University, [4]GLAM, Imperial College

Introduction

- ASC models being used in critical applications, e.g.:
  - Security
  - Autonomous driving
- ASC models should perform equally across devices/locations
  - Locations correspond to neighbourhoods of diverse social groups
  - Devices targeted at different users
- Underspecification can pose a problem to generalisation
  - Different models behave different across subpopulations
  - Operationalisation becomes harder

Methodology

- Train 5 different architectures
  - 3-layer FFNN
  - TDNN (x-vector system)
  - CNN6, CNN10, CNN14 (VGGish CNNs)
- Investigate 2 different datasets
  - TUT-Urban (different cities and locations)
  - TUT-Mobile (different cities, locations, and devices)
- Utilise disaggregated evaluations
  - *Unitary*: examine different factors in isolation
  - *Intersectional:* examine multiple factors jointly

Discussion

- Disaggregated performance showing large variance
  - Should be reported along with aggregated performance
- Models trained on identical settings (data/hparams) show different behaviour on specific subpopulations
  - Model selection for practitioners becomes harder
- Fairness will be an issue for real-life ASC applications
  - Should be addressed with (differential) fairness algorithms

Results

- F_1 ratio (F_1 score at location over F_1 for class) showing a high variance
  - Fairness: different locations will get a different performance (Fig. 1)
  - Underspecification: disaggregation per city reveals different trends for different models (Fig. 2)
    - TDNN has high variance for Stockholm (2nd last) and low variance for Barcelona (1st)
    - CNNs have low variance for Stockholm (2nd last) and high variance for Barcelona (1st)
- Accuracy per city (Fig. 3) showing both fairness and underspecification issues
  - Fairness: accuracy range across different cities ~10%
  - Underspecification: different models showing diverse performance on different subpopulations
    - CNN6 (green) and CNN14 (purple) have lower performance on Vienna vs Stockholm
    - CNN10 (red) has higher performance on Vienna vs Stockholm
  - Disaggregation per device (Fig. 4) showing different behaviour for device/city combinations

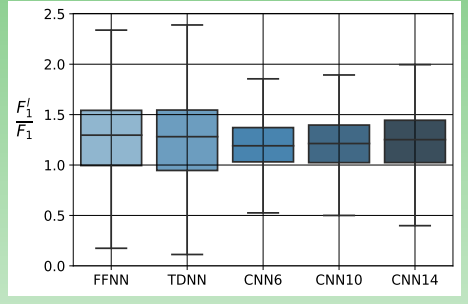

Fig. 1: F_1 ratio per location

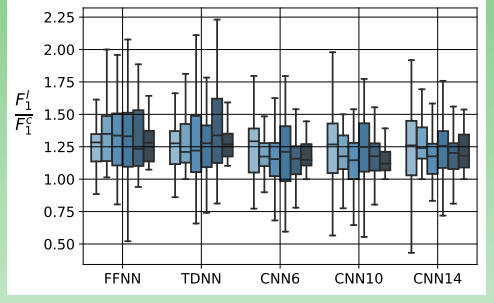

Fig. 2: F_1 ratio per location and city

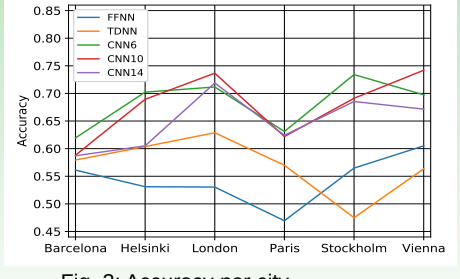

Fig. 3: Accuracy per city

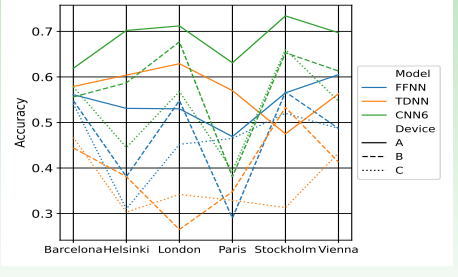

Fig. 4: Accuracy per city and device