Sound event detection with neural networks

Moritz Augustin

NI Group Meeting, 13th July 2018

Introduction

- Two!Ears EU project
  - sound event database: *NIGENS* anechoic earsignals (Neural Information Processing Group General Sounds Database Earsignals)
  - binaural simulator software: scene mixtures varying noise & sources
  - system expertise: Ivo
- Methods relied on
  - linear feedforward models: logistic regression (Trowitzsch, Mohr, Kashef, Obermayer 2017, IEEE Audio Speech Language Process.)
  - engineered features that contained all temporal information
- Extension in progress: nonlinear & temporal models
  - deep neural networks
    - recurrent networks: long short-term memory (Heiner Spieß, NI project; Changbin Lu, Master thesis)
    - feedforward (convolutional) neural networks (Alessandro Schneider, Bachelor thesis)
  - can be directly applied to features with fine temporal resolution
  - representations learnt via supervised training
  - expectation: improved generalization performance over baseline model

CURRENTLY EDITED BY HEINER

Data: Isolated Sounds

NIGENS: database with **human-labeled (on/offsets)** stereo recordings (3m head-source distance)

Examples (choose 3 example sounds compatible to mixture below (start with mixture first) and playback them either via vlc/link/embed here s.t. pdf works?)

Add waveform and labels for the 3 examples: should be one labelthrough (e.g. Piano/Caravan.wav) and one with breaks (e.g. Footsteps/*)
examples should match to the chosen example mixture on the next slide

Class	Waves (count)	min-max (sec)	Total time (min)
Alarm	>=30	fehlt-fehlt	fehlt
Baby	>=30	fehlt-fehlt	fehlt
Crash	>=30	fehlt-fehlt	Fehlt
Dog	>=30	fehlt-fehlt	Fehlt
Engine	>=30	fehlt-fehlt	Fehlt
FemaleScream	>=30	fehlt-fehlt	Fehlt
FemaleSpeech	>=30	fehlt-fehlt	Fehlt
Fire	>=30	fehlt-fehlt	Fehlt
Footsteps	>=30	fehlt-fehlt	Fehlt
Knock	>=30	fehlt-fehlt	Fehlt
MaleScream	>=30	fehlt-fehlt	Fehlt
MaleSpeech	>=30	fehlt-fehlt	Fehlt
Phone	>=30	fehlt-fehlt	Fehlt
Piano	>=30	fehlt-fehlt	Fehlt
Total	**758?**	fehlt-fehlt	Fehlt
General class	>=30	fehlt-fehlt	Fehlt

Trowitzsch, Taghia, Kashef, Obermayer 2016, Zenodo: https://zenodo.org/record/168042

CURRENTLY EDITED BY HEINER

Data: Scenes

- Sound mixtures via binaural simulator
- Master source with **one** (of 758) waves (e.g. piano77 [replace with correct no according to chosen mixtsure])
- 0-3 distractor sources: random waves from any class (incl. general & master)
- Min 30s (repeating short master waves)
- Scene: Fixed scene params (#src, azimuth, SNR)
- Scene instance: Scene with fixed waves (create the wave of the chosen mixture mathcing the example sounds from the previous slide)

Fig 1 here

CURRENTLY EDITED BY HEINER

Data: Features

Figs 1-5 here

CURRENTLY EDITED BY HEINER

Data: Validation & Testing

Figs 5 & 6 here

CURRENTLY EDITED BY HEINER

Data: Implementation

- Should be used also vom convnet and MLP => comparability
- Data standardization

Figs 7-9 here

Model 1: Multilayer Perceptron

Model 2: Convolutional neural net

Model 3: Long short-term memory

- Motivate verbally: simple RNN cannot learn long term relationships due to vanishing gradient
- Inclusive: state transitions, no peepholes (name it)
- i.e. Vanilla LSTM
- Alternative: GRU

(Recurrent) Regularization

- Early stopping
- Inseatd of weight decay/L1 we use dropout... Dropout & CuDNN
  - Give weight matrix based recurrent variational dropout idea

Hyperparameter Optimization

- Mention scene subsampling & partial CV here, use fig but not show

Preliminary Results

Issues

- Too limited Computational resources (despite Youssef‘s effort / NVIDIA‘s donations / math cluster‘s some GPUs)

Outlook