Training neural audio classifiers with few data

Jordi Pons

jordipons.me – @jordiponsdotme

Internship at Telefónica Research
Summer 2018

Supervised by Joan Serrà

Training neural audio classifiers with few data

HOW?

- **Strong regularization**
  - Will show the limitations of the standard deep learning pipeline
- **Prototypical networks**
  - A distance-based classifier that operates over a learn latent space
- **Transfer learning**
  - Enables to leverage external sources of audio data

Methodology

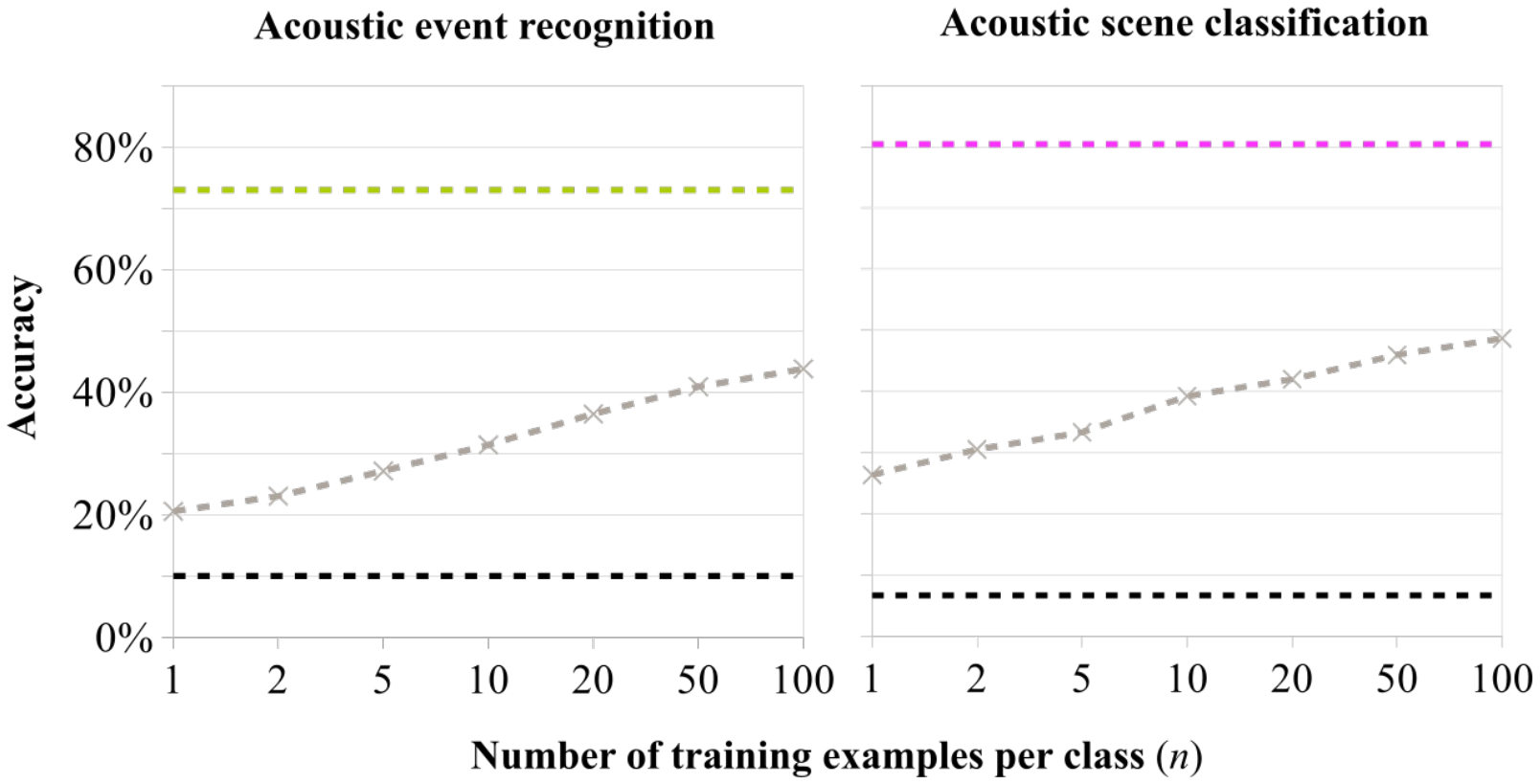
Targeted tasks and our data

- **Acoustic Event Recognition** (US8K dataset)
  - 8,732 urban sounds
  - **10 classes**: *car horn, children playing, dog bark, gun shot, siren, ...*
  - 10 folds

- **Acoustic Scene Classification** (ASC-TUT dataset)
  - 4,680 training audio segments
  - 1,620 evaluation audio segments
  - **15 classes**: *park, home, office, train, bus, ...*

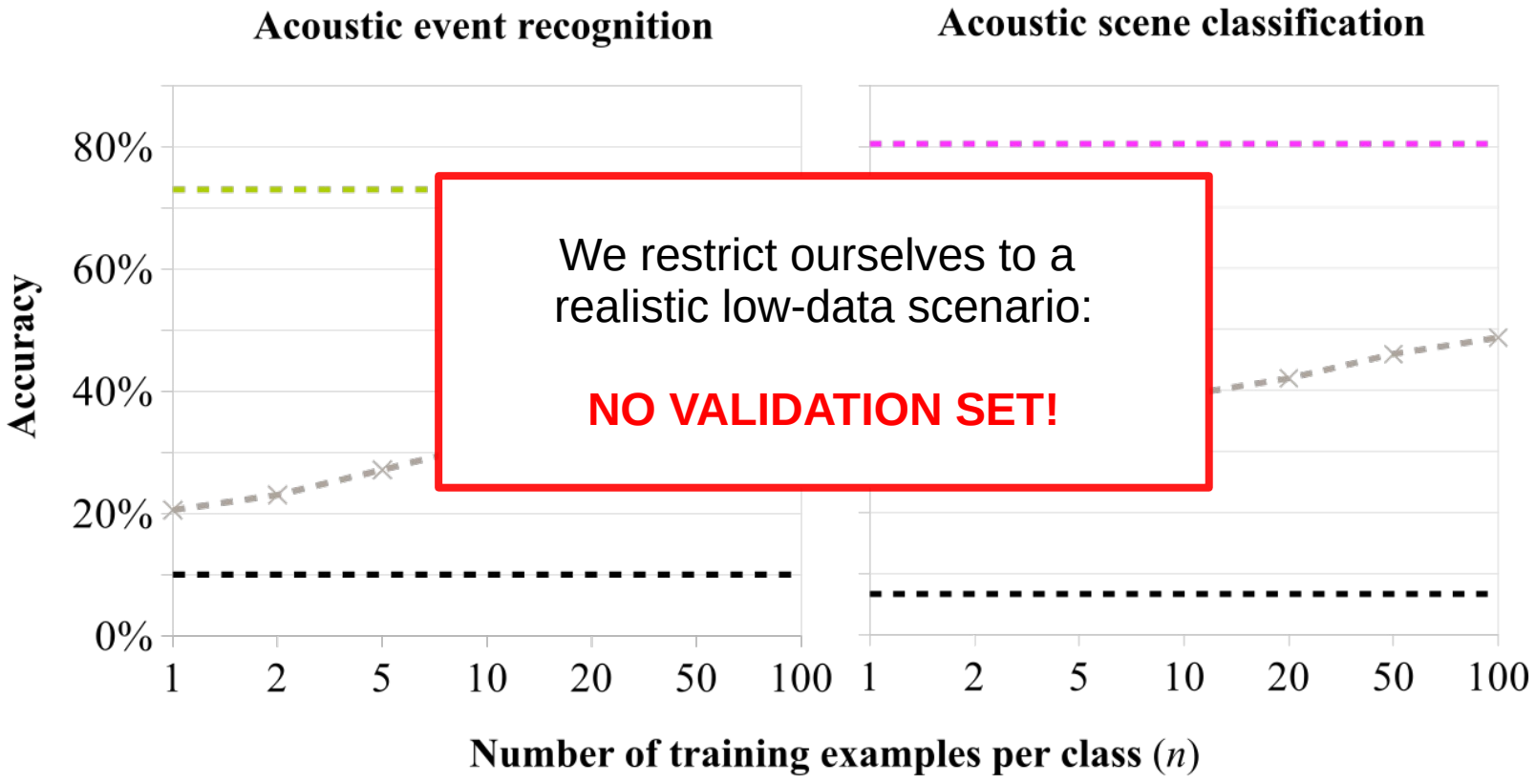
Evaluation

The MFCC's + nearest neighbor baseline case



Evaluation

The MFCC's + nearest neighbor baseline case



Regularized models

Prototypical networks

Transfer learning

Regularized models

Regularized models

Input: log-mel spectrogram of 128 bins x 3 sec (128 frames)

- **SB-CNN**: **250k** parameters
  - Inspired by AlexNet's computer vision architecture
  - *3 CNN layers (5x5) with max-pool + dense layer + softmax*
- **VGG**: **50k** parameters
  - yet another computer vision architecture
  - *5 CNN layers (3x3) with max-pool (2x2) + softmax*
- **TIMBRE**: **10k** parameters
  - The smallest CNN one can imagine for learning timbral traces
  - *1 CNN layer (vertical filters 108x7) with maxpool + softmax*

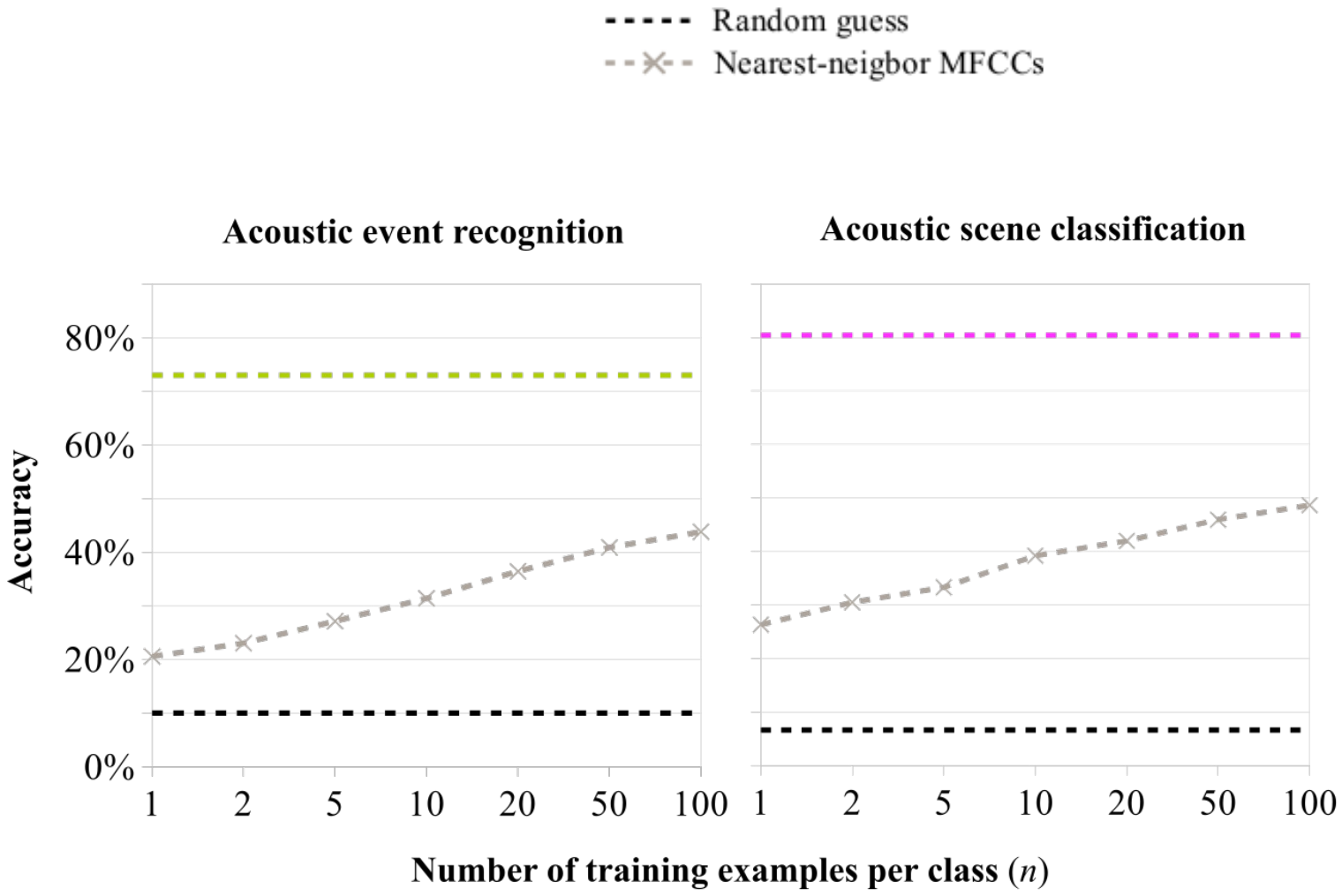
Regularized models

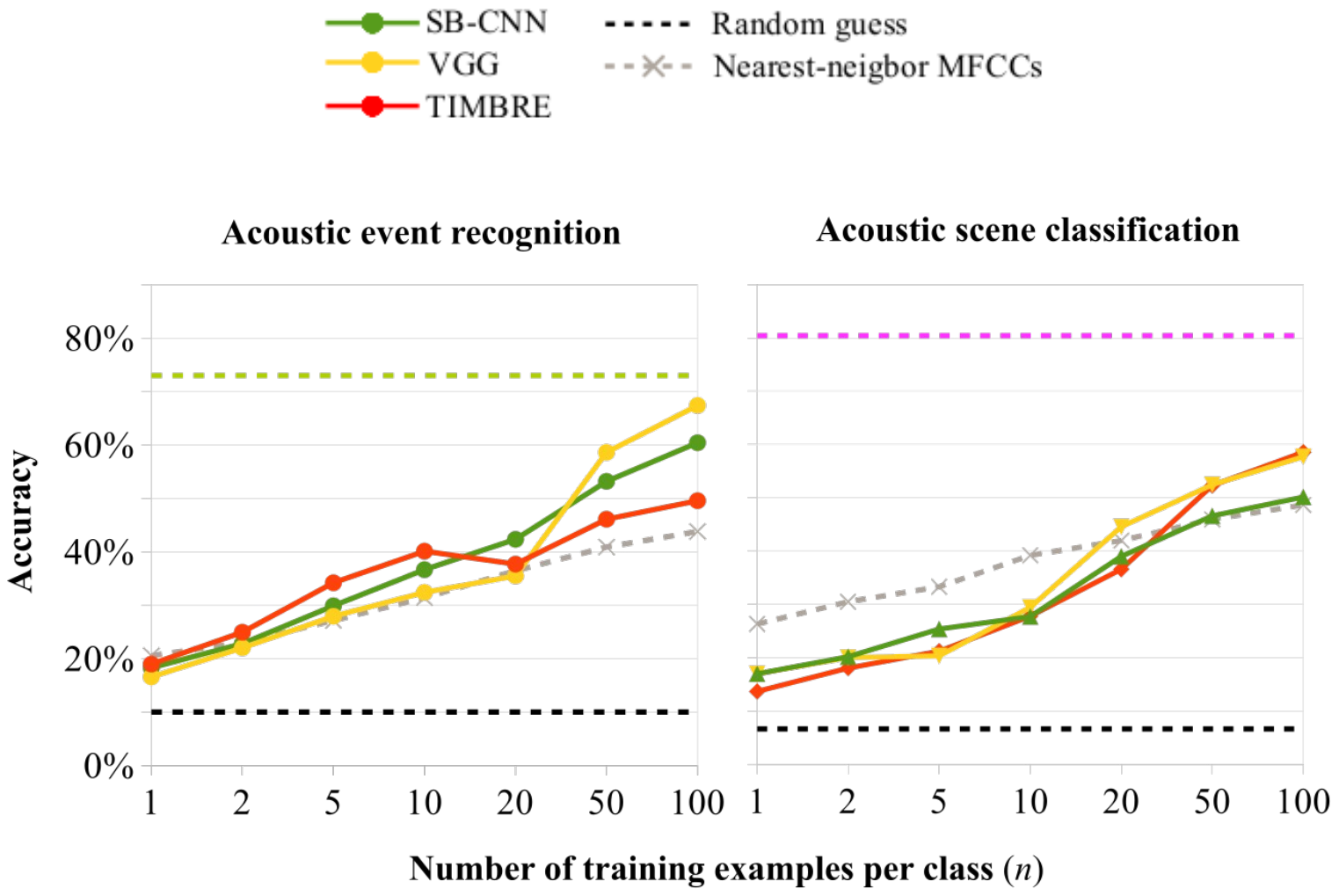
Input: log-mel spectrogram of 128 bins x 3 sec (128 frames)

- **SB-CNN**: **250k** parameters
  - Inspired by AlexNet's computer vision architecture
  - *3 CNN layers (5x5 ... ax*
- **VGG**: **50k** paramet
  - yet another compu
  - *5 CNN layers (3x3*
- **TIMBRE**: **10k** parameters
  - The smallest CNN one can imagine for learning timbral traces
  - *1 CNN layer (vertical filters 108x7) with maxpool + softmax*

Remember:
NO VALIDATION SET!

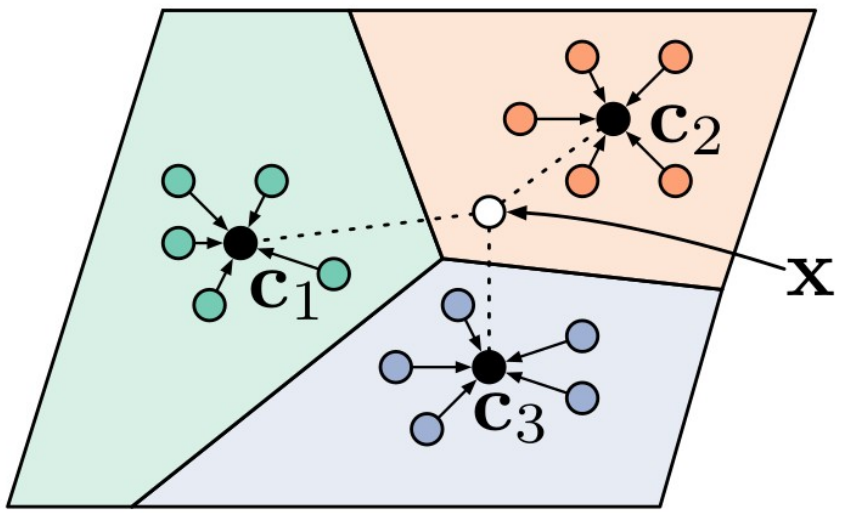
We only train for 200 epochs!


Random guess
Nearest-neigbor MFCCs
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)


SB-CNN
VGG
TIMBRE
Random guess
Nearest-neigbor MFCCs
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)

Prototypical networks

Prototypical networks



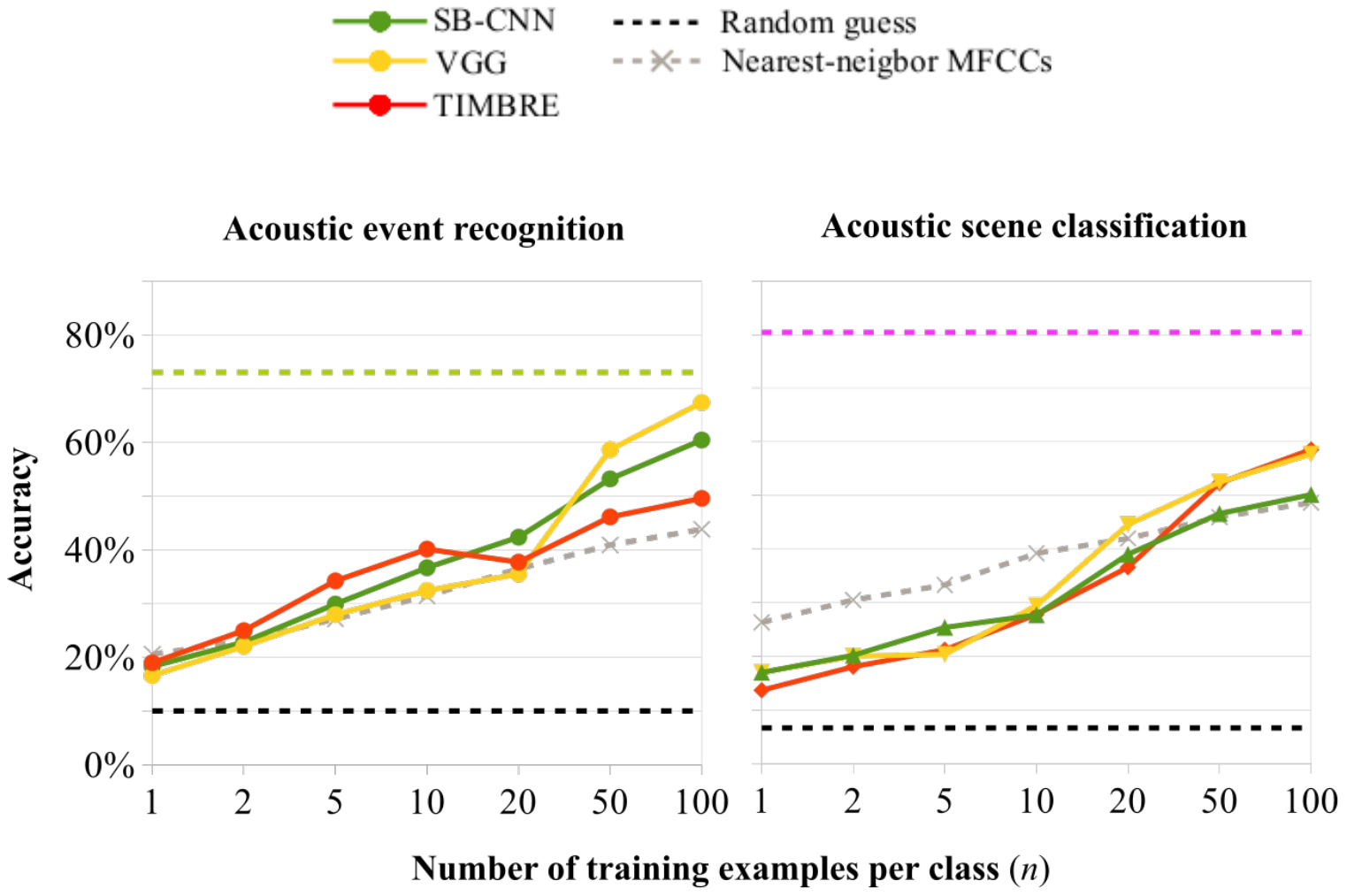
In our experiments:
a VGG parametrizes $f_\phi(\cdot)$

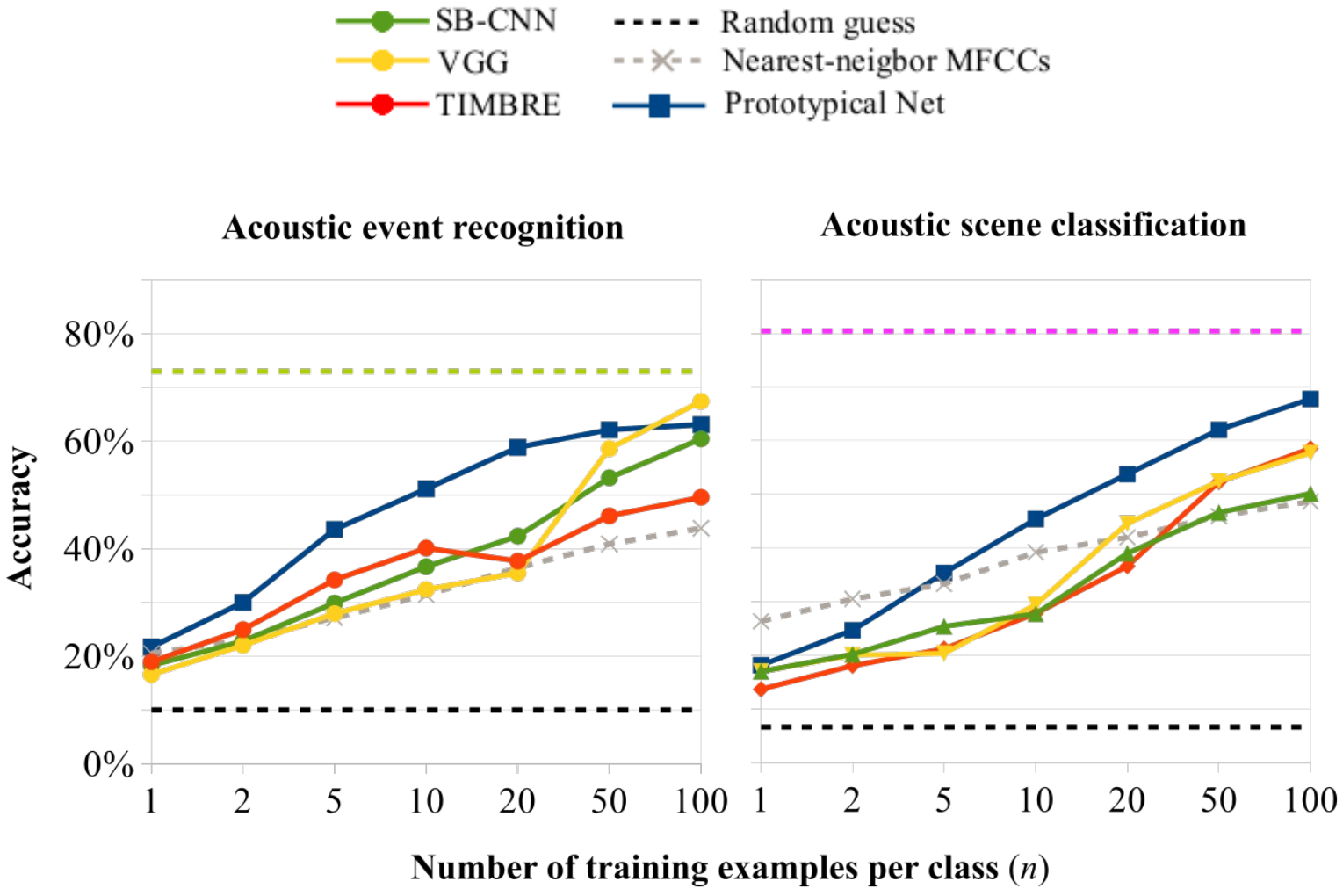
0. Compute a prototype per class (k):

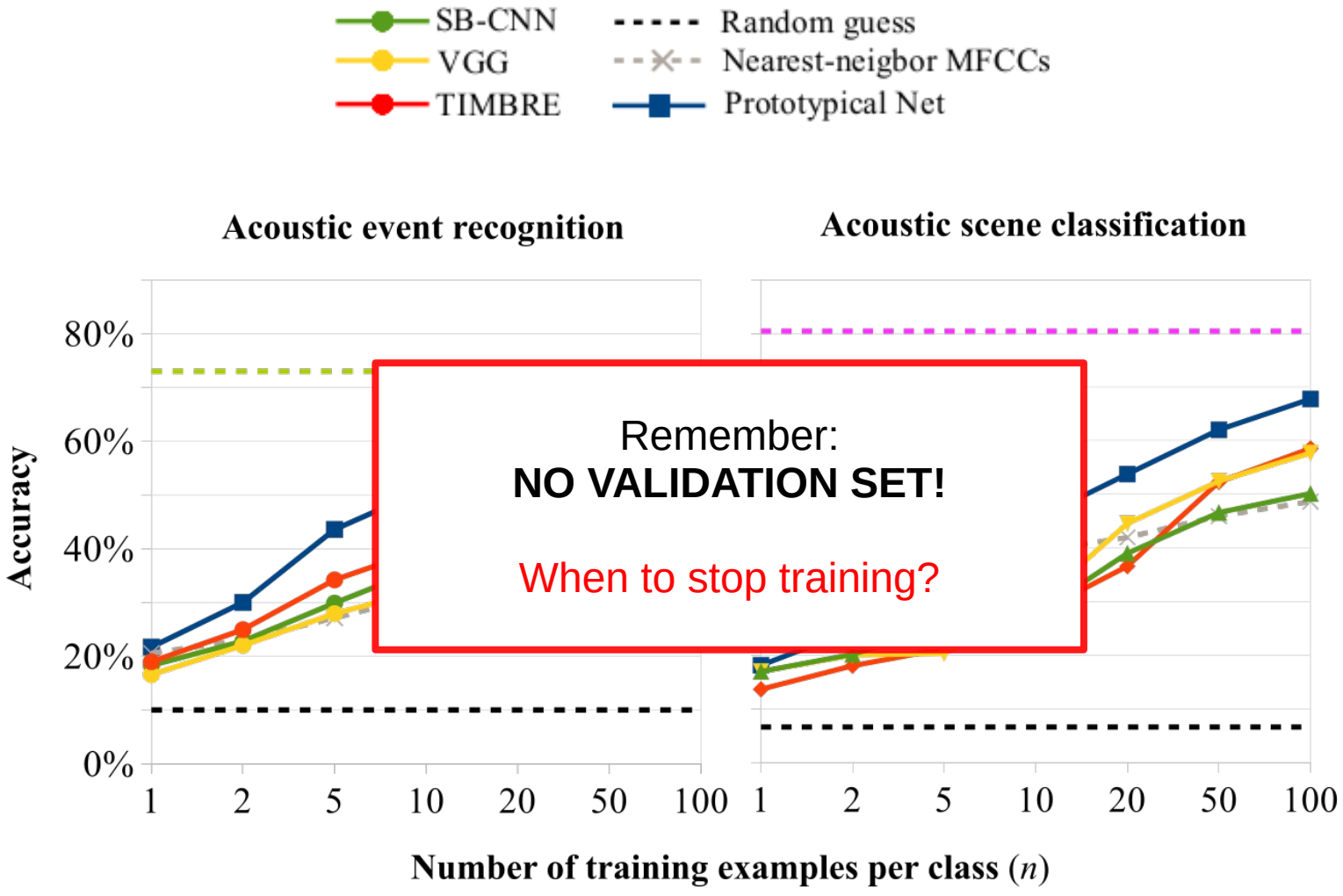
$$c_k = \mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} f_\phi(x_i)$$

1. Learning $f_\phi(\cdot)$: to separate classes in the embedding space of size 10.

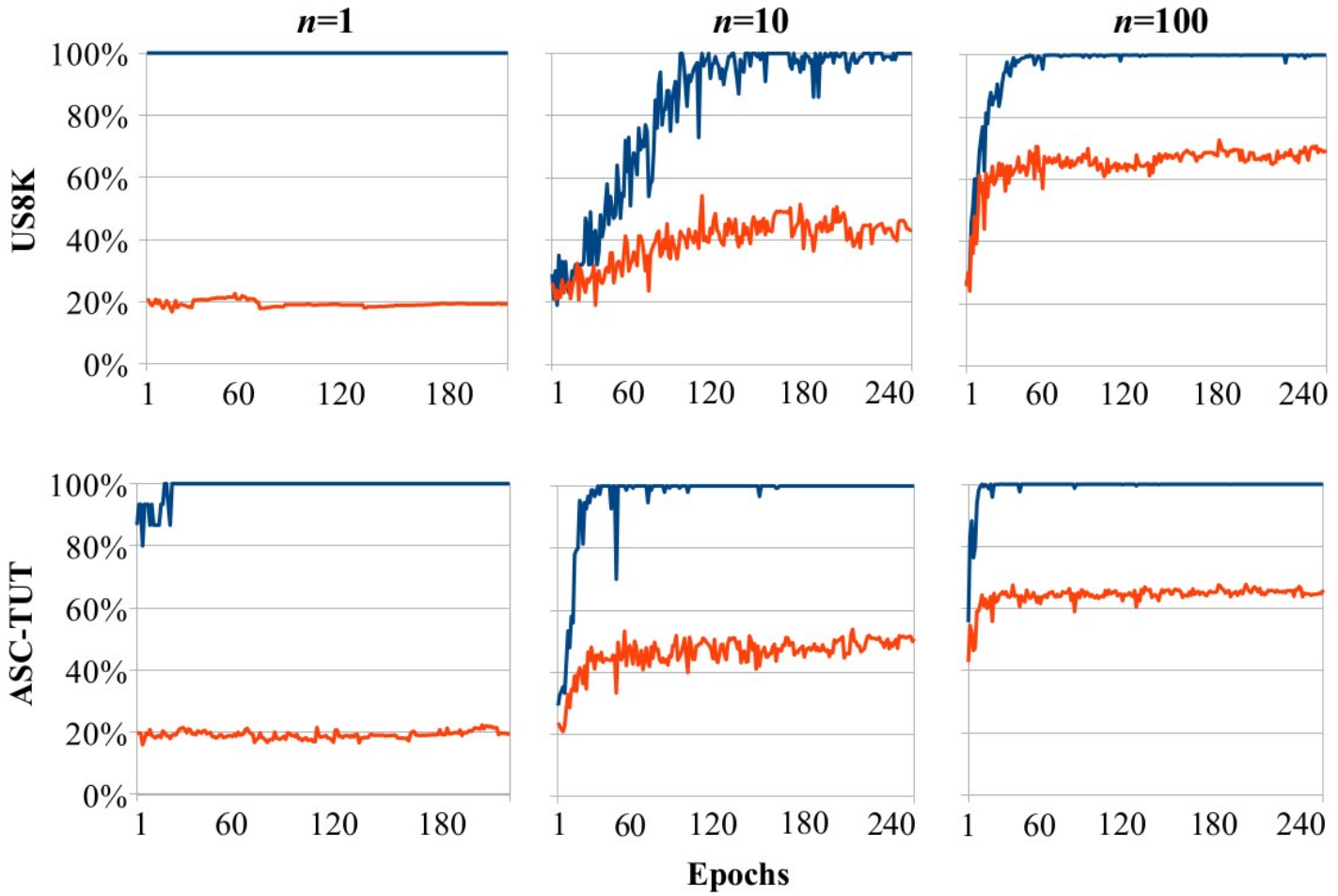
2. Classification: distribution based on a softmax over distances to the prototypes in the embedding space.


SB-CNN
VGG
TIMBRE
Random guess
Nearest-neigbor MFCCs
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)


SB-CNN
VGG
TIMBRE
Random guess
Nearest-neigbor MFCCs
Prototypical Net
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)


SB-CNN
VGG
TIMBRE
Random guess
Nearest-neigbor MFCCs
Prototypical Net
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)
Remember:
NO VALIDATION SET!
When to stop training?

The “just overfit” criteria for prototypical networks



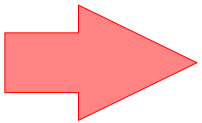
Prototypical networks do **generalize** although they **overfit**

Stop training: **after train-set accuracy does not improve**

Transfer learning

Transfer learning

pretrain with source task



finetune with target task(s)

AudioSet dataset
(acoustic event recognition)
2M Youtube audios

Pre-trained **VGGish** on AudioSet:
6 CNN layers (3×3)
with max-pool layers (2×2) +
3 dense layers (4096, 4096, 128)

US8K dataset
(acoustic event recognition)

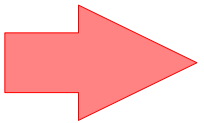
ASC-TUT dataset
(acoustic scene classification)

Finetuning of classifier:
dense softmax layer

Finetuning of prototypical network's **embedding**

Transfer learning

pretrain with source task



finetune with target task(s)

AudioSet data
(acoustic event rec
2M Youtube au

US8K dataset
tic event recognition)

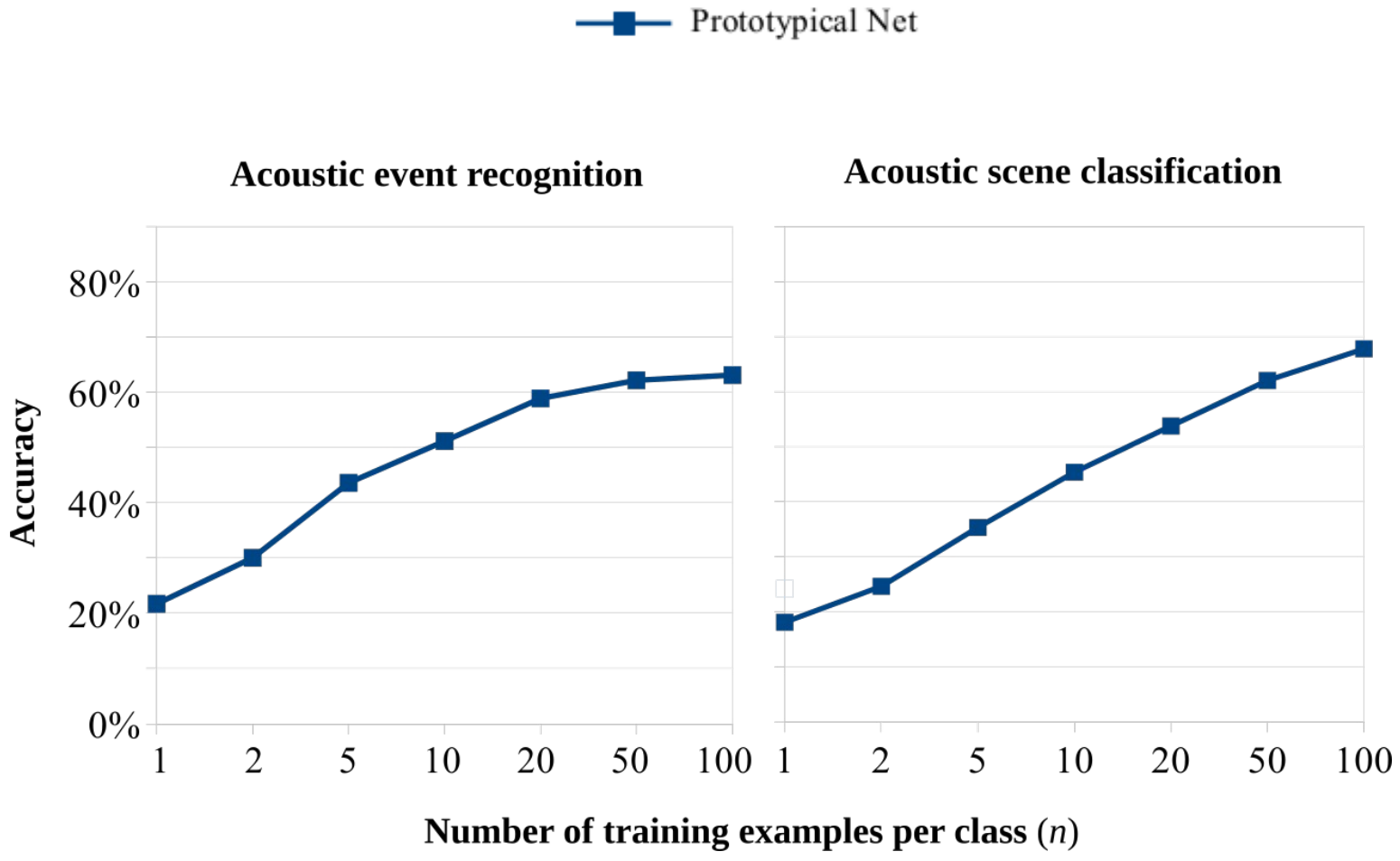
SC-TUT dataset
c scene classification)

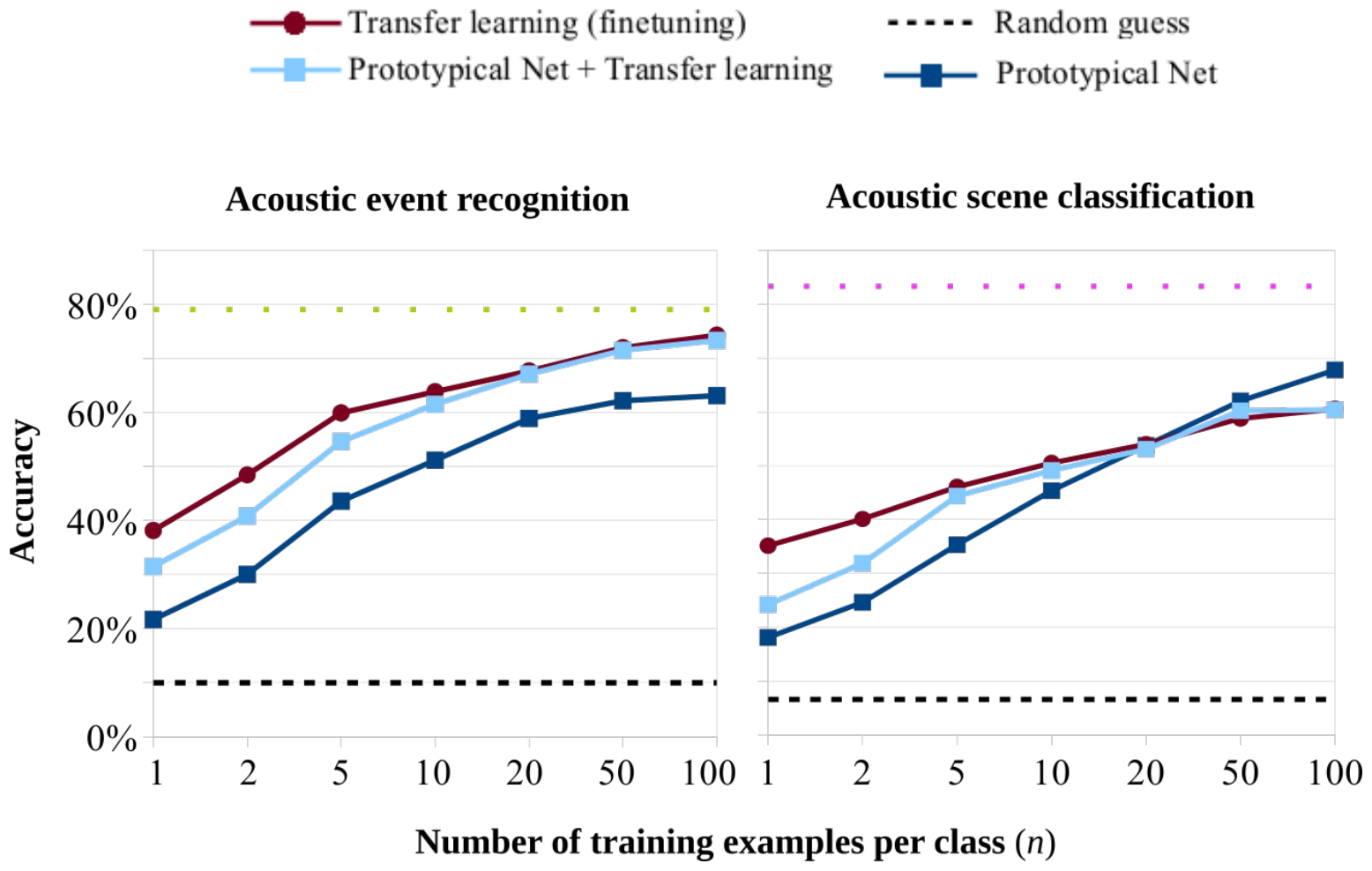
Remember:
NO VALIDATION SET!

Pre-trained **VGGish** on AudioSet:
6 CNN layers (3×3)
with max-pool layers (2×2) +
3 dense layers (4096, 4096, 128)

Finetuning of classifier:
dense softmax layer

Finetuning of prototypical network's **embedding**


Prototypical Net
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)


Transfer learning (finetuning)
Prototypical Net + Transfer learning
Random guess
Prototypical Net
Acoustic event recognition
Acoustic scene classification
Accuracy
80%
60%
40%
20%
0%
1
2
5
10
20
50
100
Number of training examples per class (n)

Conclusions

Training neural audio classifiers with few data

- **Strong regularization**
  - To realize the limitations of the standard deep learning pipeline
- **Prototypical networks**
  - A distance-based classifier that operates over a learn latent space
  - Particularly useful when:
    - No validation set is available
    - No additional “similar” data is accessible
- **Transfer learning**
  - Enables to leverage external sources of audio data

Remember:
NO VALIDATION SET!

Training neural audio classifiers with few data

Jordi Pons

jordipons.me – @jordiponsdotme

Code is available!
https://github.com/jordipons/neural-classifiers-with-few-audio

+ info in our paper:
https://arxiv.org/abs/1810.10274