Representation learning for acoustic scene analysis

Victor Bisot

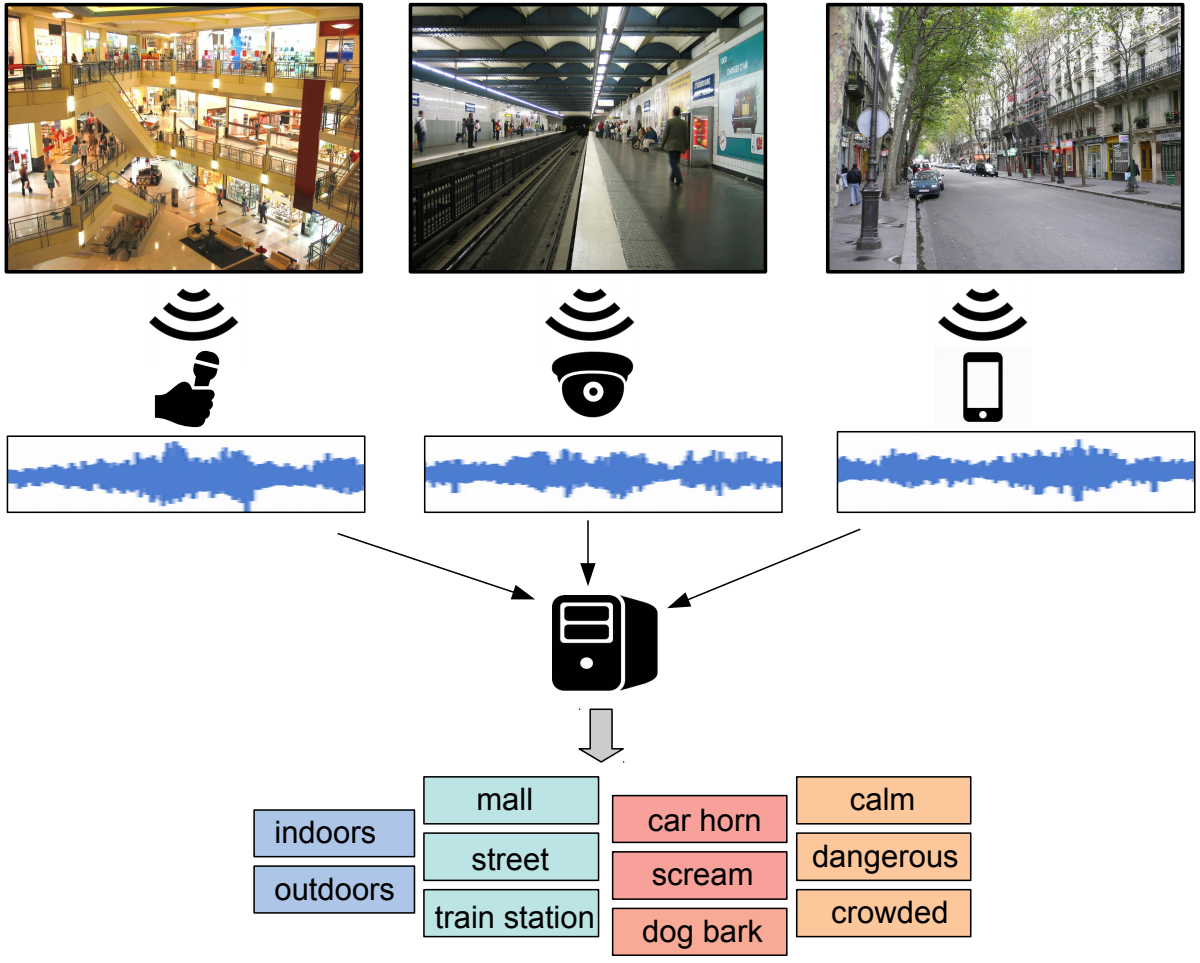
Ph.D. overview

Work with Romain Serizel, Slim Essid and Gaël Richard

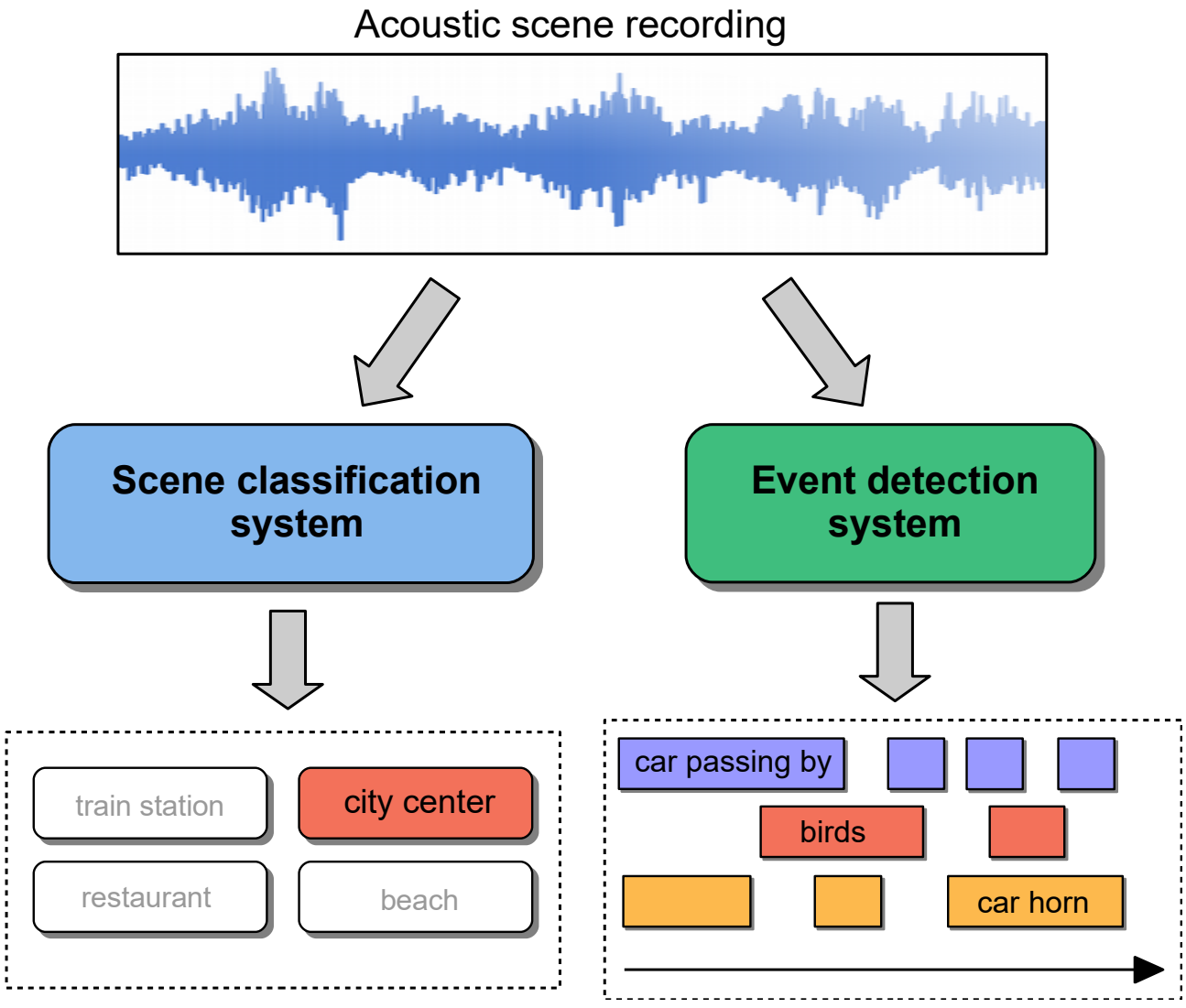
October 12, 2018

Acoustic scene and event analysis

- Make machines capable of analyzing and reacting to their surrounding acoustic environment



The different tasks



Applications and interest

- Context aware devices
  - Smart cars and homes
  - Noise monitoring in cities
  - Surveillance

Applications and interest

- Context aware devices
  - Smart cars and homes
  - Noise monitoring in cities
  - Surveillance
- Analysis and indexing
  - Multimedia indexing
  - Forensics
  - Bio-acoustics

Applications and interest

- Context aware devices
  - Smart cars and homes
  - Noise monitoring in cities
  - Surveillance
- Analysis and indexing
  - Multimedia indexing
  - Forensics
  - Bio-acoustics
- Machines can outperform humans
  - $\rightarrow$ below 60% accuracy for humans
  - $\rightarrow$ up to 90% for computational approaches

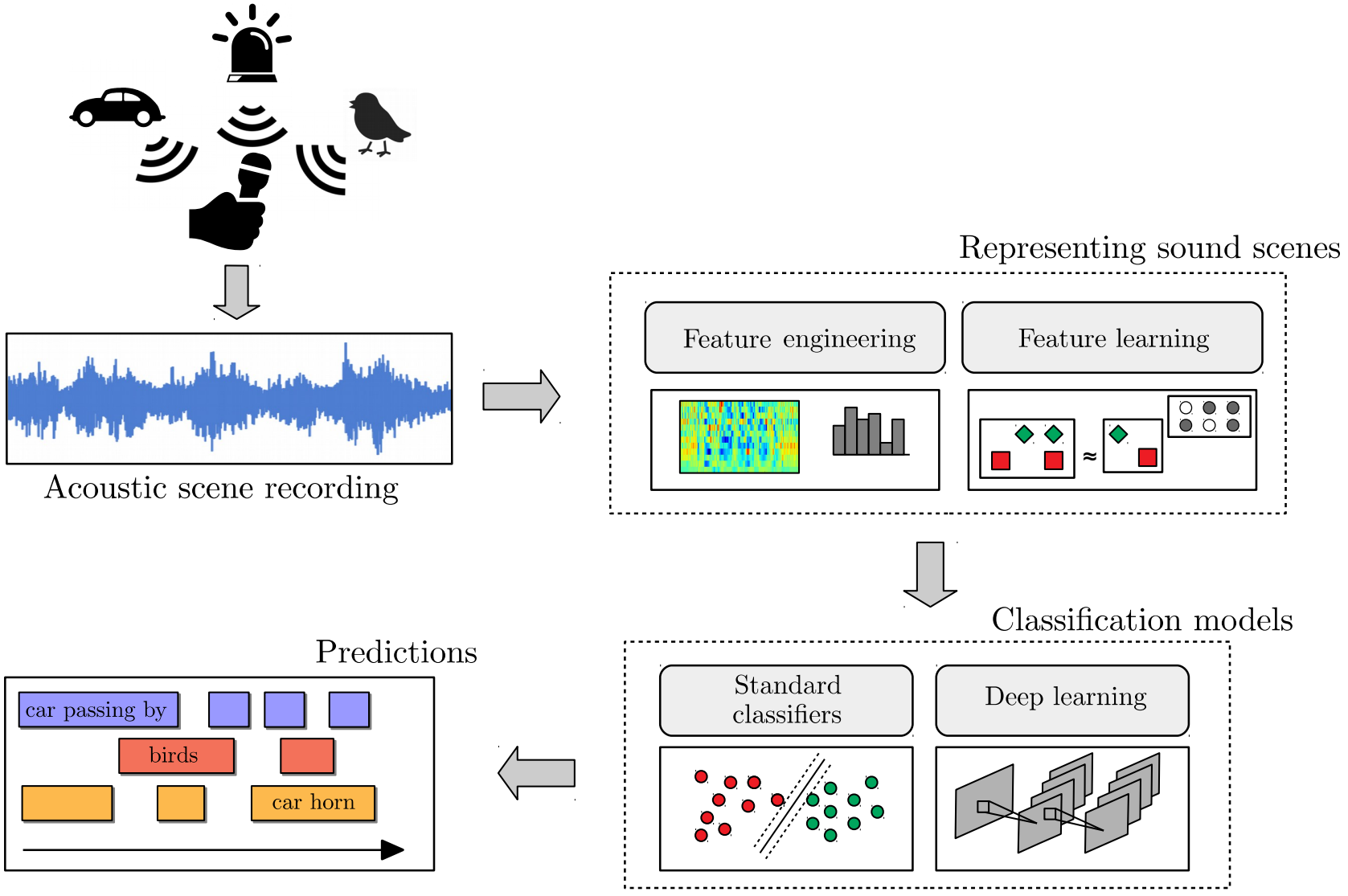
Applications and interest

- Context aware devices
  - Smart cars and homes
  - Noise monitoring in cities
  - Surveillance
- Analysis and indexing
  - Multimedia indexing
  - Forensics
  - Bio-acoustics

- Machines can outperform humans
  - $\rightarrow$ below 60% accuracy for humans
  - $\rightarrow$ up to 90% for computational approaches

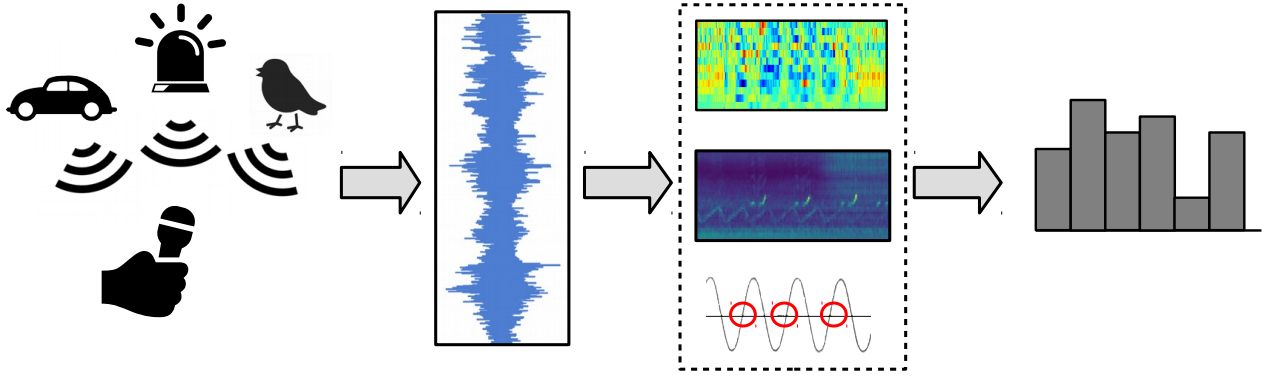
	Scene classification			**Event detection**		
Dcase challenge	**2013**	**2016**	**2017**	**2013**	**2016**	**2017**
Nbr teams	10	35	49	7	21	34

Acoustic scene analysis systems



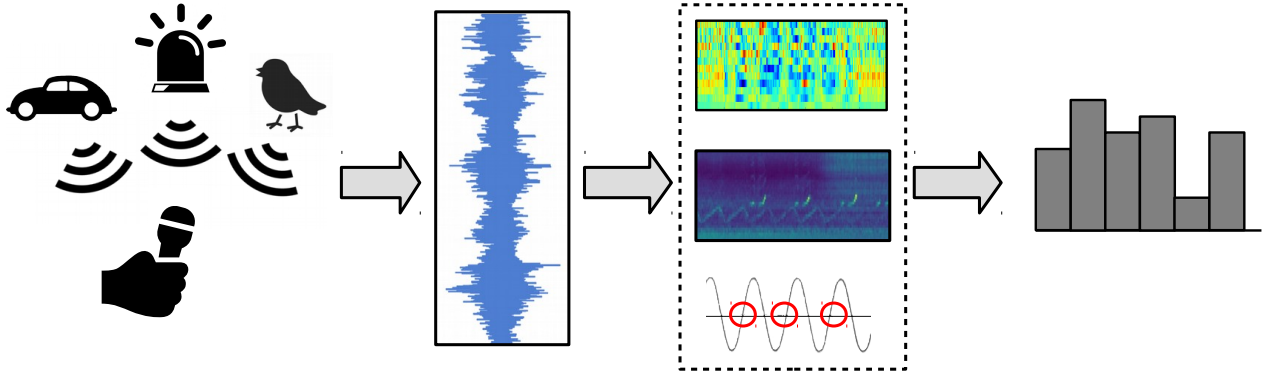
Early approaches

- Combine hand-crafted acoustic features inspired from music and speech processing tasks

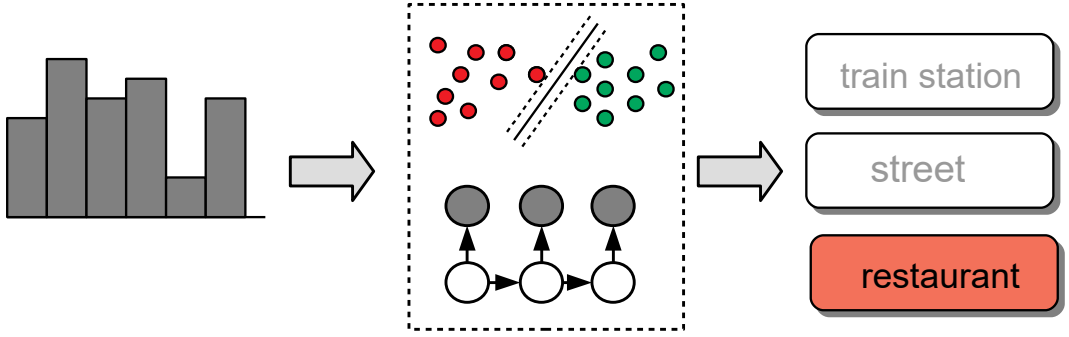


Early approaches

- Combine hand-crafted acoustic features inspired from music and speech processing tasks



- Rely on simple linear or probabilistic classifiers to interpret the features

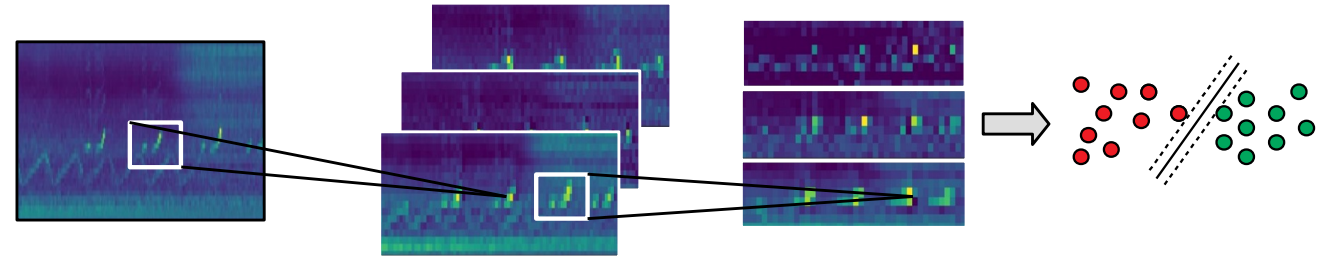


Modern approaches and remaining problems

- Recent rise of deep learning approaches
  - → Scene and event classification [Piczark et al., 2015; Valenti et al. 2016]
  - → Event detection and tagging [Cakir et al., 2015,2017; Xu et al. 2017]

Modern approaches and remaining problems

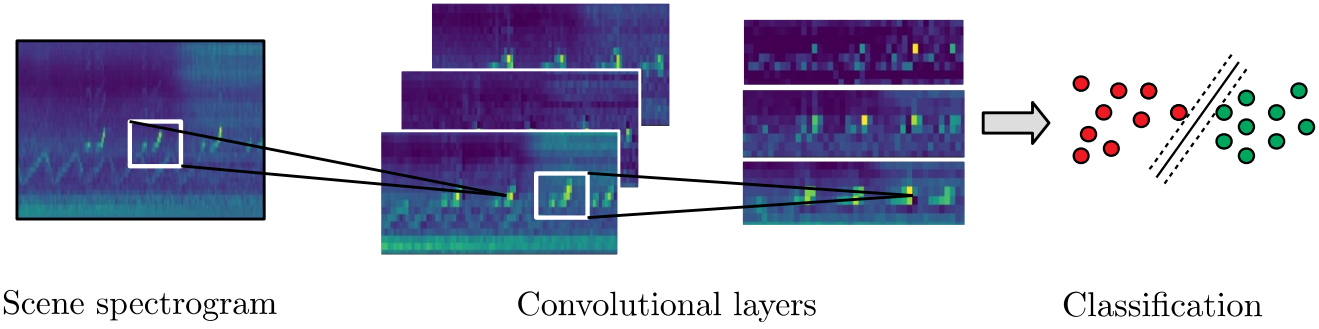
- Recent rise of deep learning approaches
  - → Scene and event classification [Piczark et al., 2015; Valenti et al. 2016]
  - → Event detection and tagging [Cakir et al., 2015,2017; Xu et al. 2017]
- Learning complex models from time-frequency representations
  - → Representation learning = convolutional layers



Scene spectrogram Convolutional layers Classification

Modern approaches and remaining problems

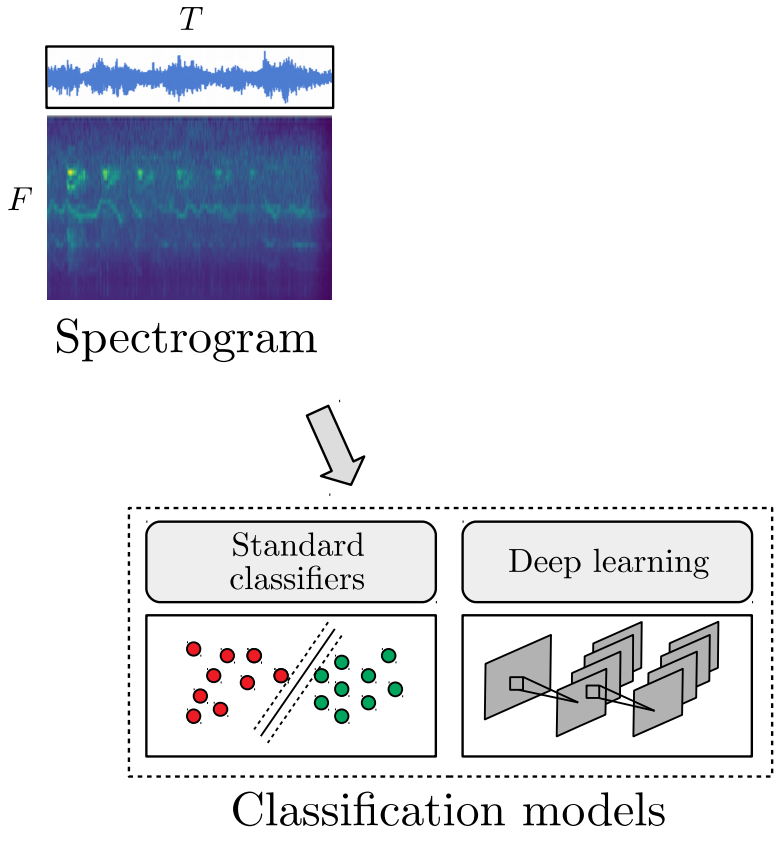
- Recent rise of deep learning approaches
  - → Scene and event classification [Piczark et al., 2015; Valenti et al. 2016]
  - → Event detection and tagging [Cakir et al., 2015,2017; Xu et al. 2017]
- Learning complex models from time-frequency representations
  - → Representation learning = convolutional layers



- Are simple spectrograms the best input?
- What happens in more challenging conditions (overlap, noise ...)?
- How to deal with limited training data?

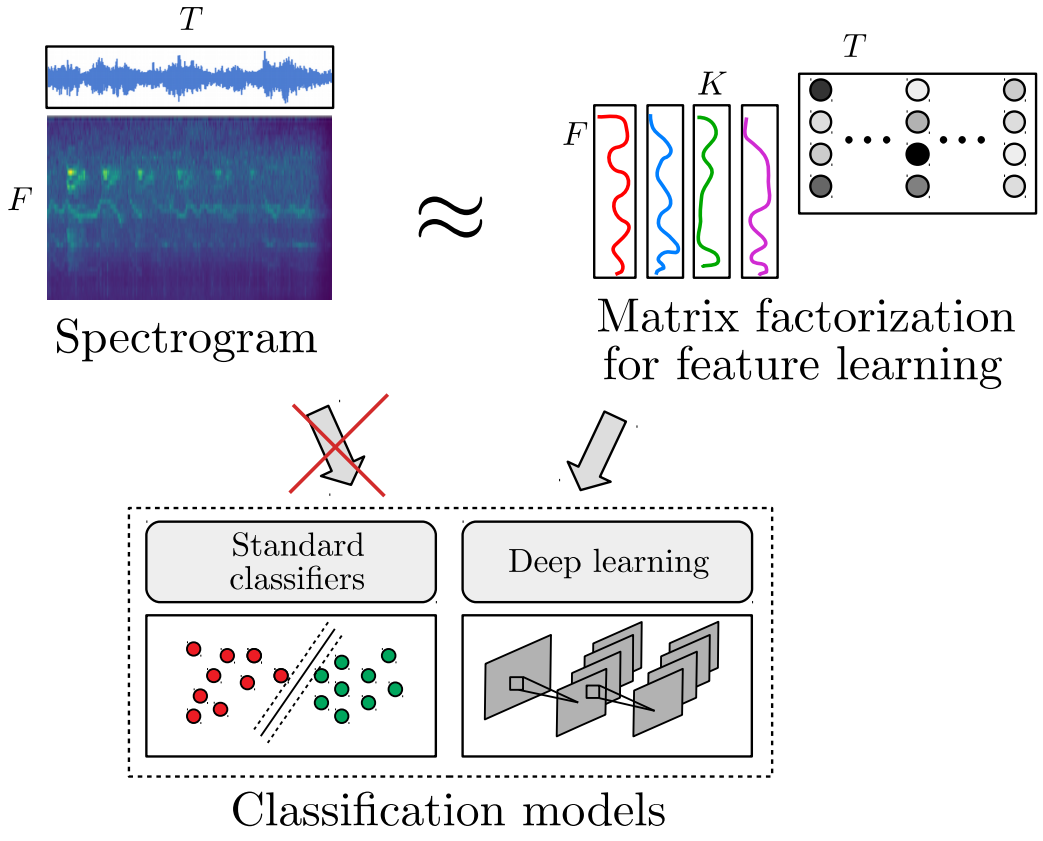
Our approach

- An extra feature learning step with nonnegative matrix factorization (NMF) techniques
  - → Learn decompositions suited for representing multi-source environments



Our approach

- An extra feature learning step with nonnegative matrix factorization (NMF) techniques
  - → Learn decompositions suited for representing multi-source environments



Outline and contributions

- Combination of spectrogram image features for improved acoustic scene classification

Outline and contributions

- Combination of spectrogram image features for improved acoustic scene classification
- Unsupervised feature learning with NMF
  - $\rightarrow$ Automatically learn relevant features in multi-source environments
  - $\rightarrow$ Promising improvements with a simple system design

Outline and contributions

- Combination of spectrogram image features for improved acoustic scene classification
- Unsupervised feature learning with NMF
  - → Automatically learn relevant features in multi-source environments
  - → Promising improvements with a simple system design
- Supervised NMF with TNMF
  - → Learn discriminative dictionaries for better representations
  - → Simple models to challenge complex neural networks

Outline and contributions

- Combination of spectrogram image features for improved acoustic scene classification
- Unsupervised feature learning with NMF
  - → Automatically learn relevant features in multi-source environments
  - → Promising improvements with a simple system design
- Supervised NMF with TNMF
  - → Learn discriminative dictionaries for better representations
  - → Simple models to challenge complex neural networks
- Deep learning and NMF
  - → NMF as suitable input to deep networks
  - → Event detection with NMF + recurrent and convolutional networks
  - → Jointly learn nonnegative dictionaries and neural networks

Outline and contributions

- Combination of spectrogram image features for improved acoustic scene classification

- Unsupervised feature learning with NMF
  - → Automatically learn relevant features in multi-source environments
  - → Promising improvements with a simple system design

- Supervised NMF with TNMF
  - → Learn discriminative dictionaries for better representations
  - → Simple models to challenge complex neural networks

- Deep learning and NMF
  - → NMF as suitable input to deep networks
  - → Event detection with NMF + recurrent and convolutional networks
  - → Jointly learn nonnegative dictionaries and neural networks

Outline

Unsupervised feature learning
Scene classification evaluation

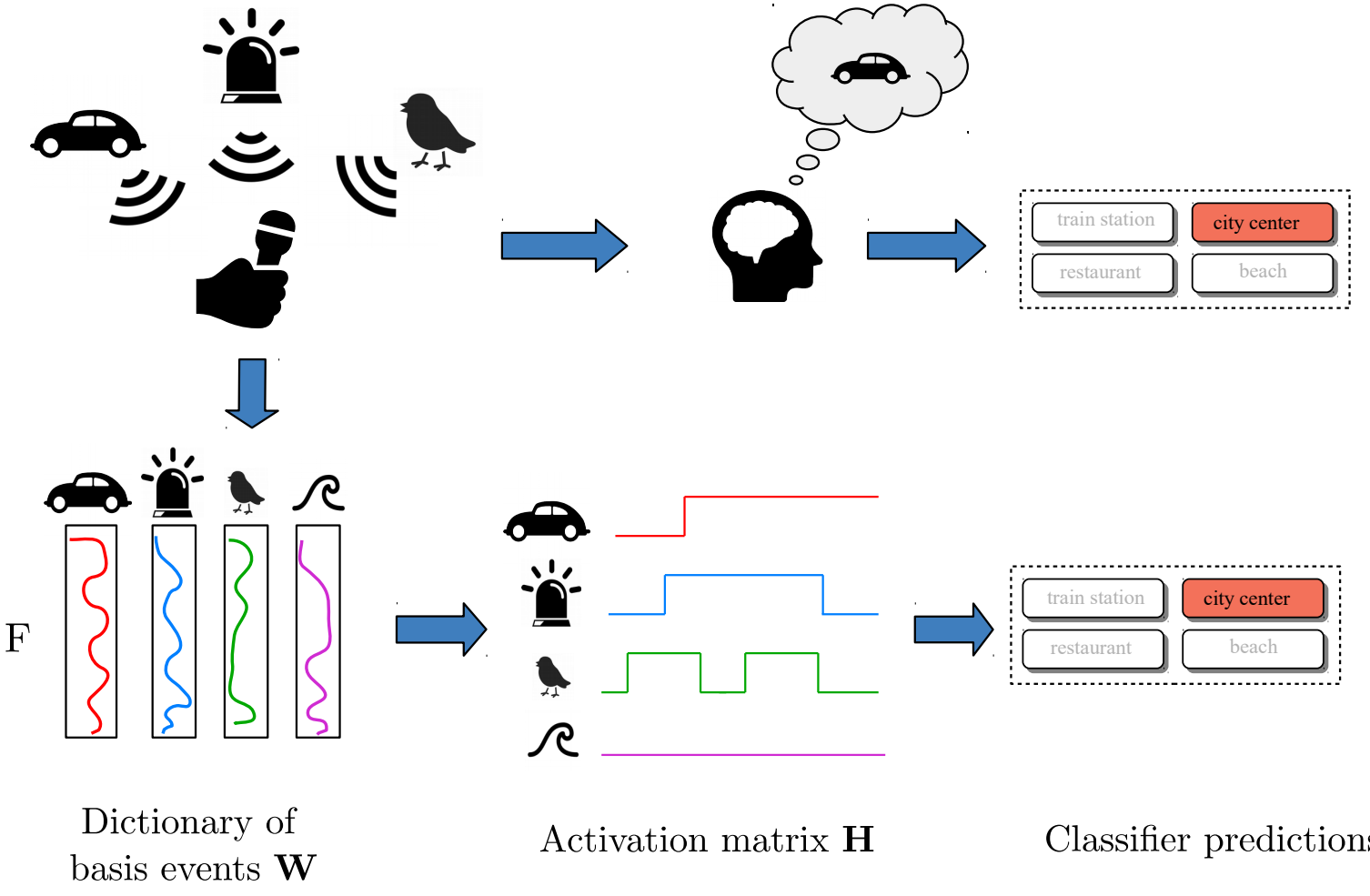
Task-driven NMF

Deep learning and NMF

Conclusion

Dictionary learning and scene understanding

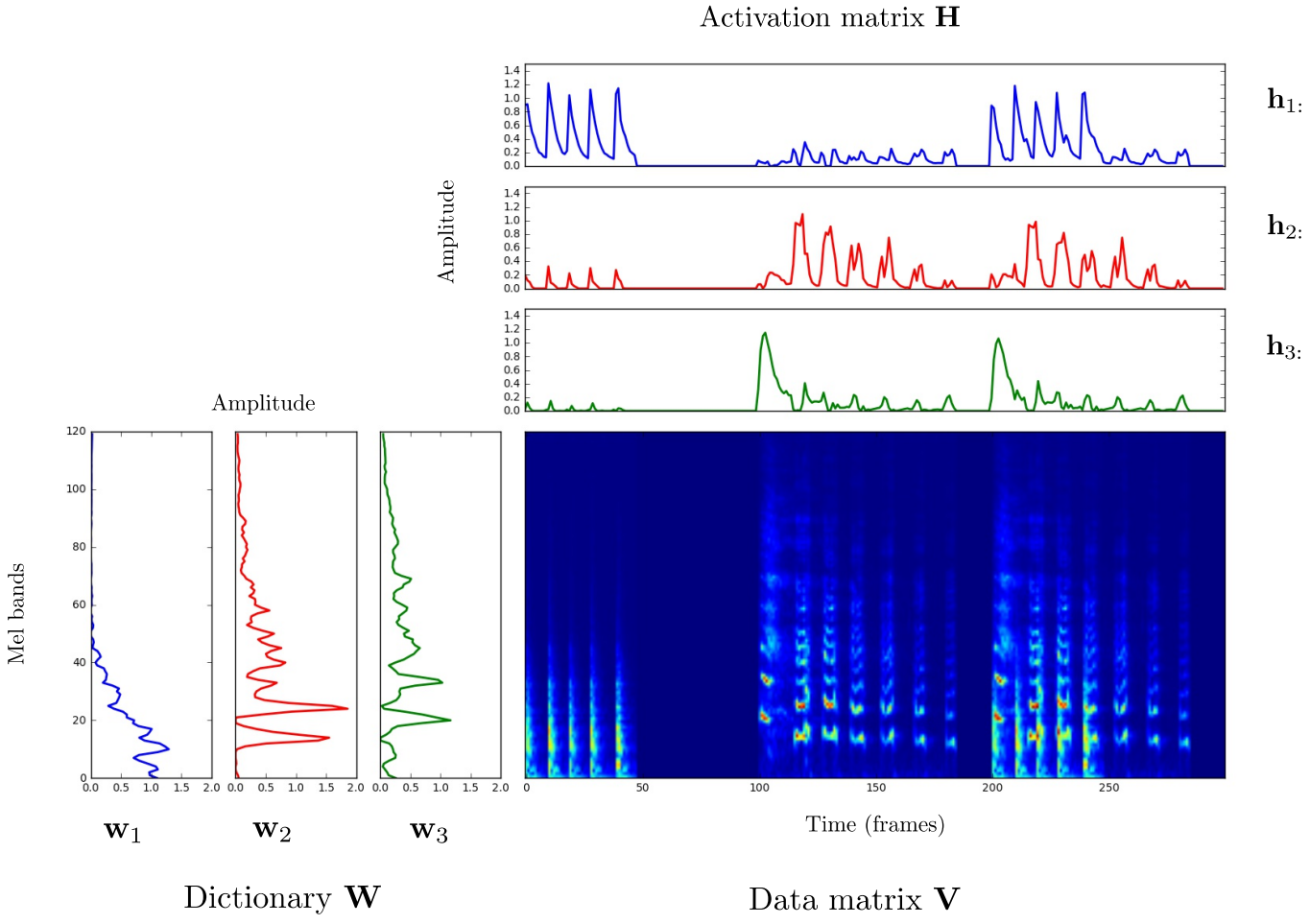
- Humans rely on specific event cues to recognize acoustic environments



Nonnegative matrix factorization

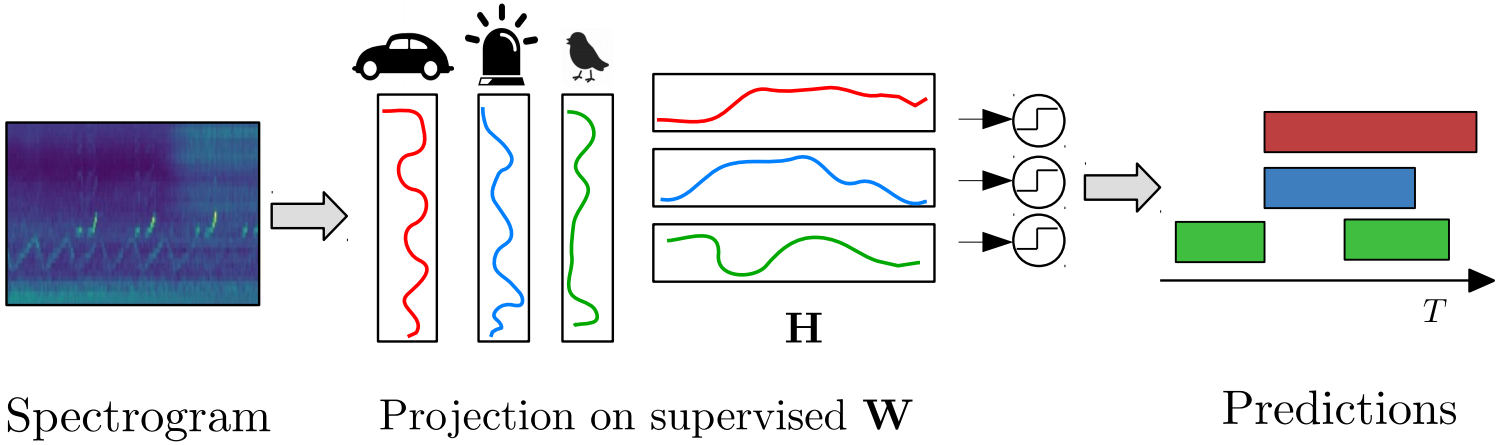
NMF model with K components

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}\|\mathbf{WH}) \text{ with } \mathbf{W} \in \mathbb{R}_+^{F\times K} \text{ and } \mathbf{H} \in \mathbb{R}_+^{K\times T}$$



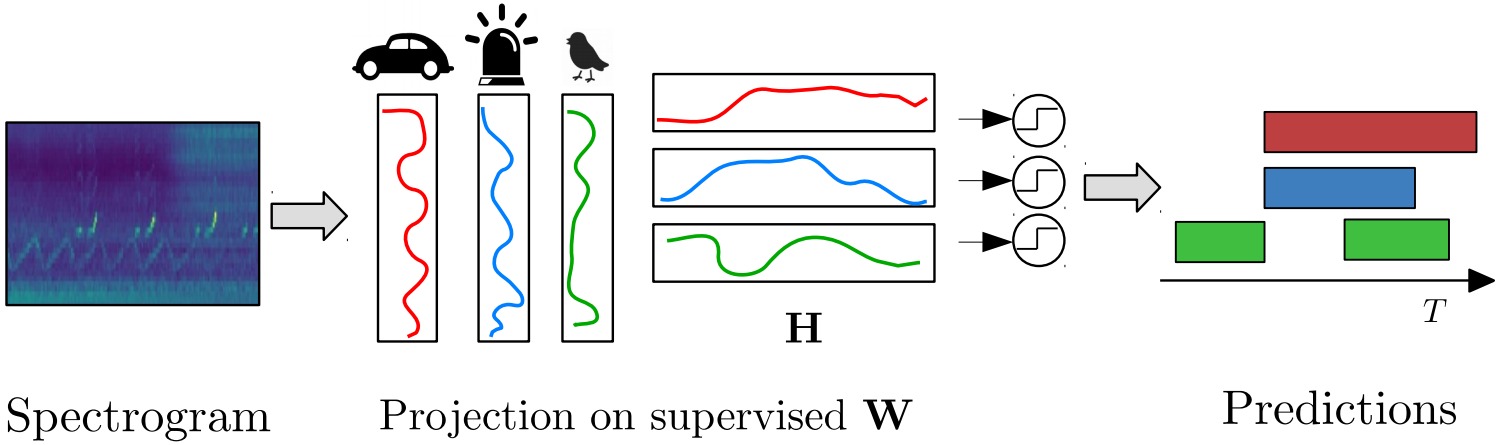
A different approach

- Distance between individual NMF dictionaries as classification criteria [Cauchi et al., 2011; Benetos et al., 2012]
- NMF as a direct event detection tool: event activity by thresholding activation matrix [Mesaros et al.,2015,2016; Benetos et al., 2016]



A different approach

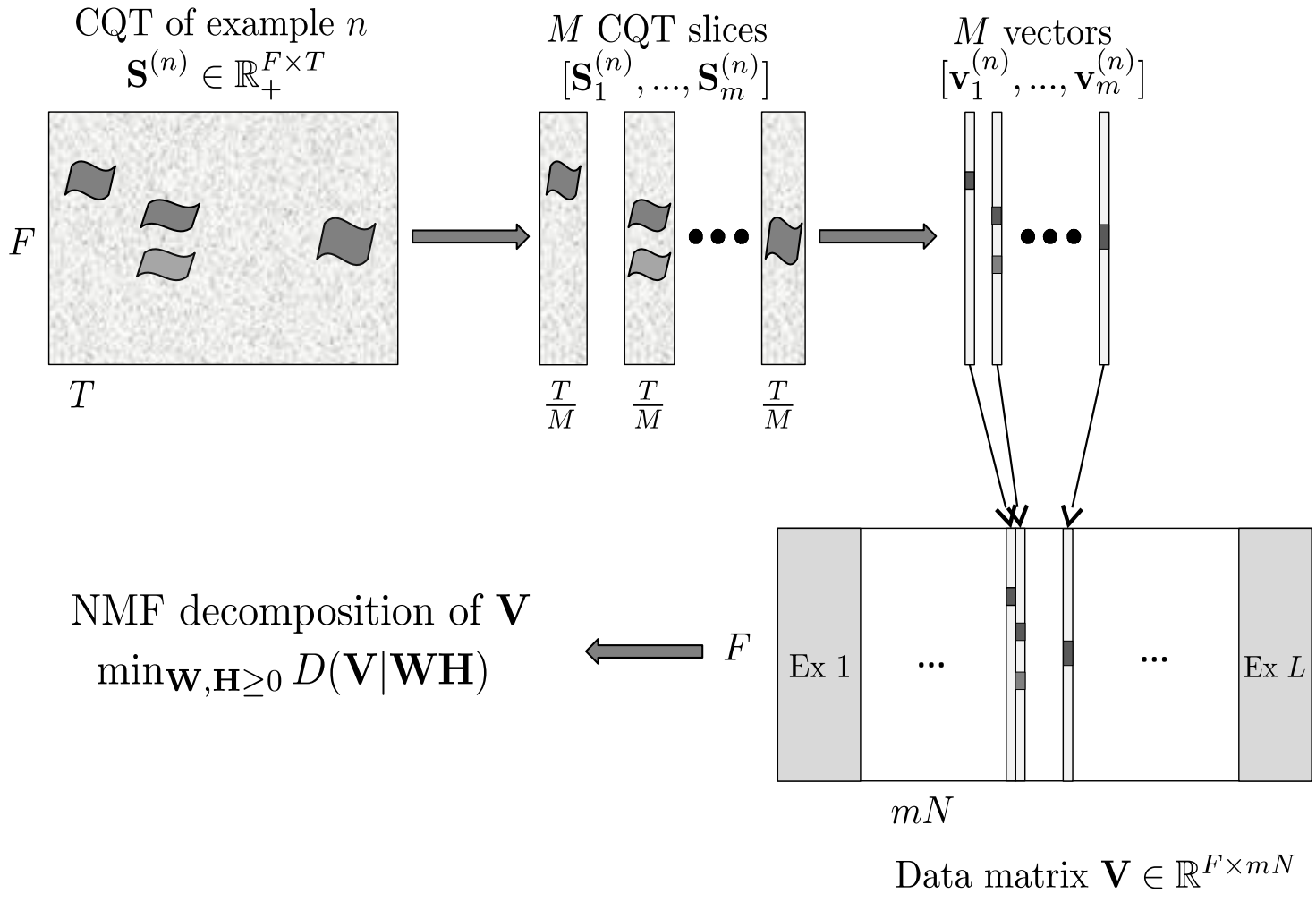
- Distance between individual NMF dictionaries as classification criteria [Cauchi et al., 2011; Benetos et al., 2012]
- NMF as a direct event detection tool: event activity by thresholding activation matrix [Mesaros et al.,2015,2016; Benetos et al., 2016]



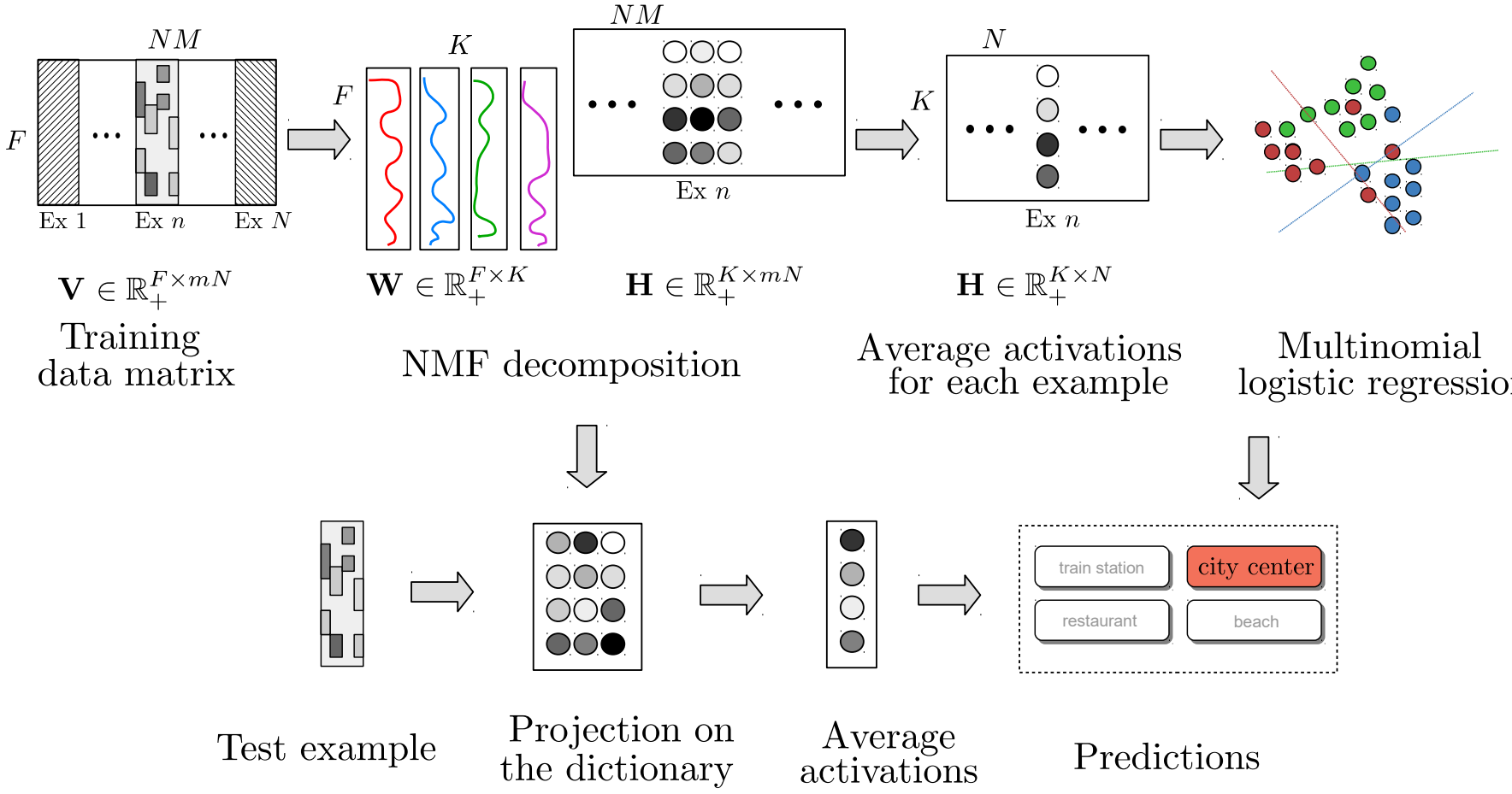
- Our approach: use NMF as a **feature learning** tool
  - → Learn one dictionary of basis events representing the entire data
  - → Train classifiers separately on NMF representations

From spectrograms to NMF features

- Group all scene constant-Q transform spectrograms in one data matrix



Scene classification with NMF



Outline

Unsupervised feature learning
    Scene classification evaluation

Task-driven NMF

Deep learning and NMF

Conclusion

Scene classification datasets

- Evaluation on 3 of the largest public scene classification datasets

	LITIS Rouen	**DCASE 2016**	**DCASE 2017**
Nbr of segments	3026	1170	4680
Length of segments	30 sec	30 sec	10 sec
Nbr of labels	19	15	15
Cross validation folds	20	4	4

bus	car	café/restaurant	city center
forest path	grocery store	home	beach
library	metro station	office	residential area
train	tram	park	

First scene classification results

- Accuracy scores on the 3 scene classification datasets

Feature	**LITIS Rouen**	**DCASE 2016**	**DCASE2017**
Kernel PCA	**96.0**	80.2	82.2
Sparse NMF	94.6	**82.7**	**84.4**.
Convolutive NMF	94.8	**82.5**	83.7
Baseline	91.7	72.5	74.8
HOG + SPD [1,2]	94.0	77.7	80.3
MFCC + RQA [1,3]	86.0	67.1	-

- Comparison of proposed techniques to best hand-crafted features from DCASE 2013 challenge
  - Histogram of oriented gradients (HOG) image features
  - MFCC + Recurrence quantification analysis (RQA)

[1] Rakotomamonjy and Gasso 2014, [2] Bisot et al. 2015, [3] Roma et al. 2013,

First scene classification results

- Accuracy scores on the 3 scene classification datasets

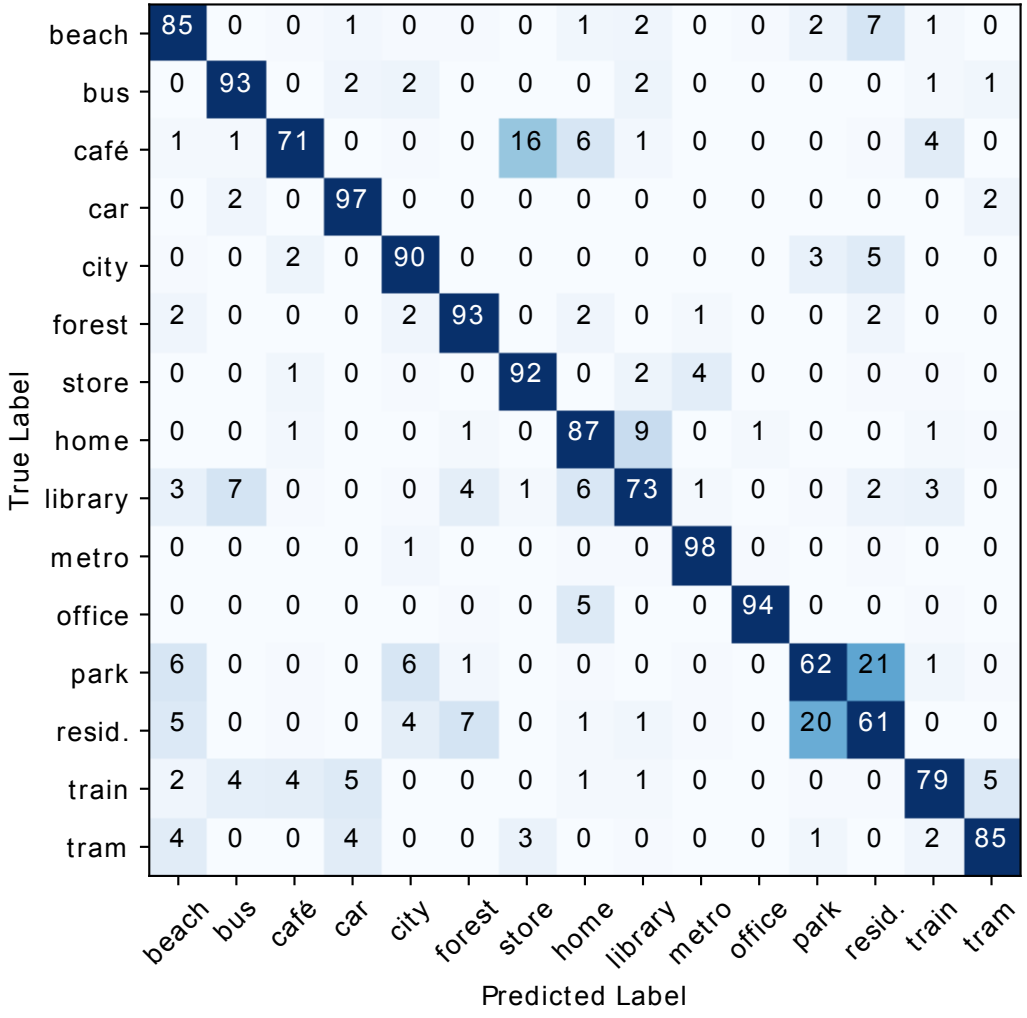
Feature	**LITIS Rouen**	**DCASE 2016**	**DCASE2017**
Kernel PCA	**96.0**	80.2	82.2
Sparse NMF	94.6	**82.7**	**84.4.**
Convolutive NMF	94.8	**82.5**	83.7
Baseline	91.7	72.5	74.8
HOG + SPD [1,2]	94.0	77.7	80.3
MFCC + RQA [1,3]	86.0	67.1	-

- Comparison of proposed techniques to best hand-crafted features from DCASE 2013 challenge
  - Histogram of oriented gradients (HOG) image features
  - MFCC + Recurrence quantification analysis (RQA)

[1] Rakotomamonjy and Gasso 2014, [2] Bisot et al. 2015, [3] Roma et al. 2013,

Potential for improvement

- Confusion matrix for Sparse NMF on DCASE 2017 dataset



- Confusions between similar acoustic environments
- Promising results with unsupervised feature learning and simple classifiers

- Can we adapt the decomposition to the task?

Outline

Unsupervised feature learning

Task-driven NMF
- TNMF model
- Event detection evaluation

Deep learning and NMF

Conclusion

Outline

Unsupervised feature learning

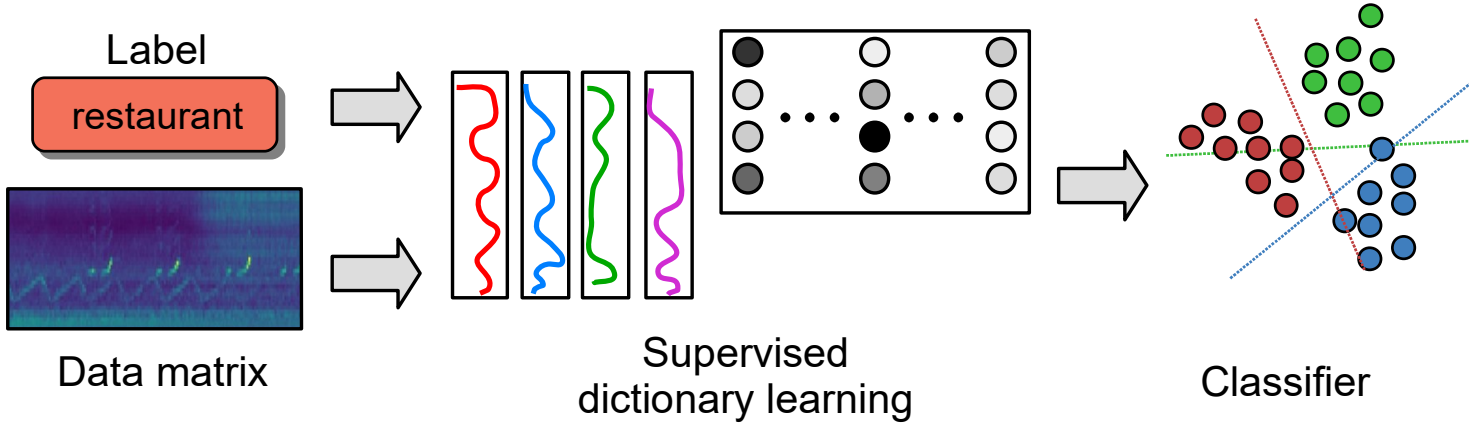
Task-driven NMF
- TNMF model
- Event detection evaluation

Deep learning and NMF

Conclusion

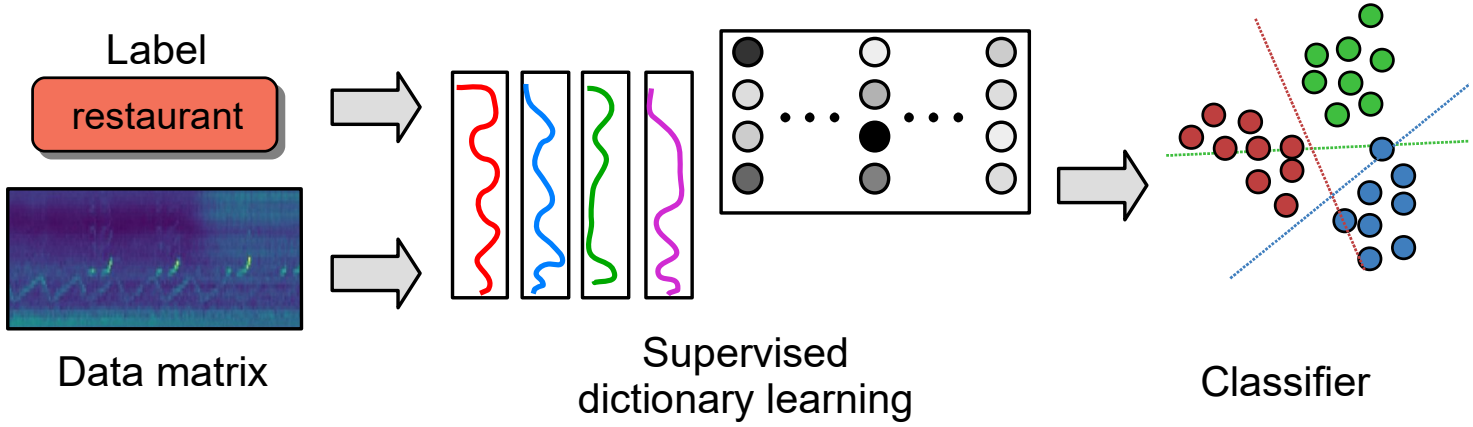
Supervised Matrix Factorization

- Make the decomposition better at dealing with the task at hand
  - → Leverage label information to learn more discriminative dictionaries



Supervised Matrix Factorization

- Make the decomposition better at dealing with the task at hand
  - → Leverage label information to learn more discriminative dictionaries



Task-driven dictionary learning (TDL) [Mairal et al. 2009]

- Jointly learn dictionaries and classifiers
- Allows for efficient nonnegative variants
  - → source separation [Sprechmann et al. 2014]
  - → speaker identification [Serizel et al. 2017]
  - → acoustic scene and event classification

Task-driven dictionary learning

TDL formulation

- Composition of optimal projection function and a supervised loss function

$$\begin{cases} \mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}^K} \|\mathbf{v} - \mathbf{W}\mathbf{h}\| + \lambda_1 \|\mathbf{h}\|_1 + \frac{\lambda_2}{2} \|\mathbf{h}\|_2^2 \\ \\ \min_{\mathbf{W}, \mathbf{A}} \mathbb{E}_{y, \mathbf{v}}[\ell_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))] + \frac{\nu}{2} \|\mathbf{A}\|_2^2 \end{cases}$$

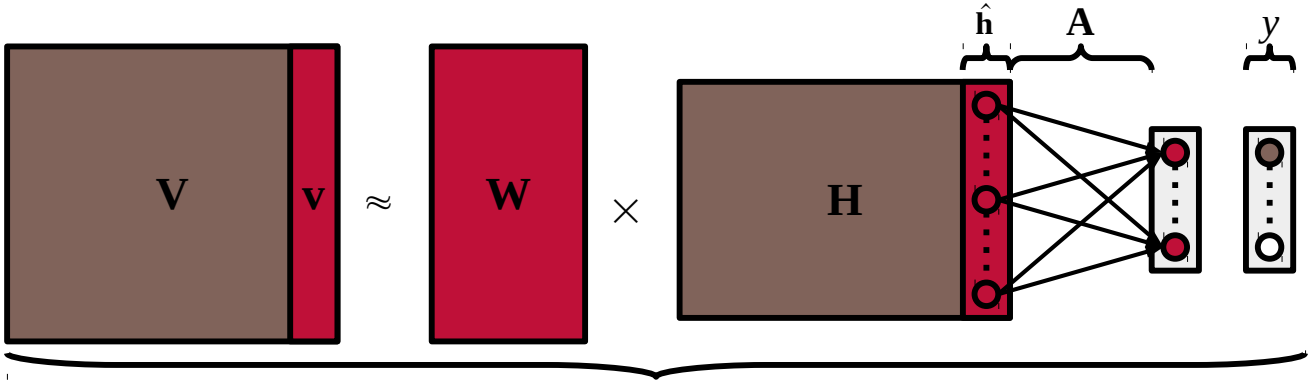
- $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$: optimal projection of a data point on the dictionary $\rightarrow$ learned features for classification
- ℓ_1 and ℓ_2 regularizations: λ_1 and λ_2
- ℓ_s: supervised classification loss
- $\mathbf{A}$ classifier parameters

TNMF problem formulation

Task-driven NMF

$$\begin{cases} \mathbf{h}^\star(\mathbf{v},\mathbf{W}) = \min_{\mathbf{h}\in\mathbb{R}_+^K} D_\beta(\mathbf{v}|\mathbf{W}\mathbf{h}) + \lambda_1\|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|_2^2 \\ \min_{\mathbf{W}\geq 0,\mathbf{A}} \mathbb{E}_{y,\mathbf{v}}[\ell_s(y,\mathbf{A},\mathbf{h}^\star(\mathbf{v},\mathbf{W}))] + \frac{\nu}{2}\|\mathbf{A}\|_2^2 \end{cases}$$

- Regroup the sparse NMF-based classification systems in one problem
  - $\mathbf{h}^\star(\mathbf{v},\mathbf{W}) \rightarrow$ sparse NMF projection with the β-divergence
  - $\ell_s(y,\mathbf{A},\mathbf{h}^\star(\mathbf{v},\mathbf{W})) = -\log(\mathrm{P}(y|\mathbf{h}^\star,\mathbf{A})) \rightarrow$ multinomial logistic regression



TNMF algorithm

Alternate update of $\mathbf{A}$ et $\mathbf{W}$

For a fixed number of iterations:

{Classifier update}

- Compute optimal projection on full training data $\mathbf{H}^{\star}(\mathbf{V}, \mathbf{W})$
- Update $\mathbf{A}$ with one iteration of L-BFGS for logistic regression

TNMF algorithm

Alternate update of **A** et **W**

For a fixed number of iterations:

{Classifier update}

- Compute optimal projection on full training data $\mathbf{H}^\star(\mathbf{V}, \mathbf{W})$
- Update $\mathbf{A}$ with one iteration of L-BFGS for logistic regression

{Dictionary update}

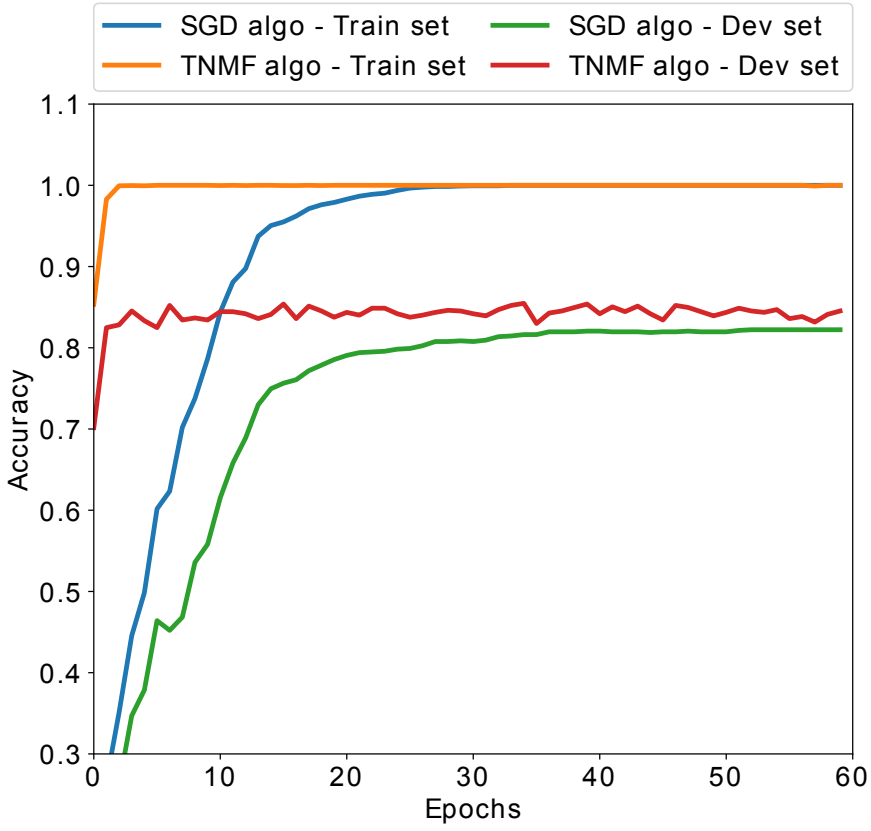
- On one epoch
  1. Draw a random data point $\mathbf{v}$ with label y
  2. Compute optimal projection $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ and gradient $\nabla_{\mathbf{W}} \ell_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))$
  3. Update $\mathbf{W}$ by projected gradient $\mathbf{W} \leftarrow \Pi_{\mathcal{W}_+}[\mathbf{W} - \rho_t \nabla_{\mathbf{W}} \ell_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))]$

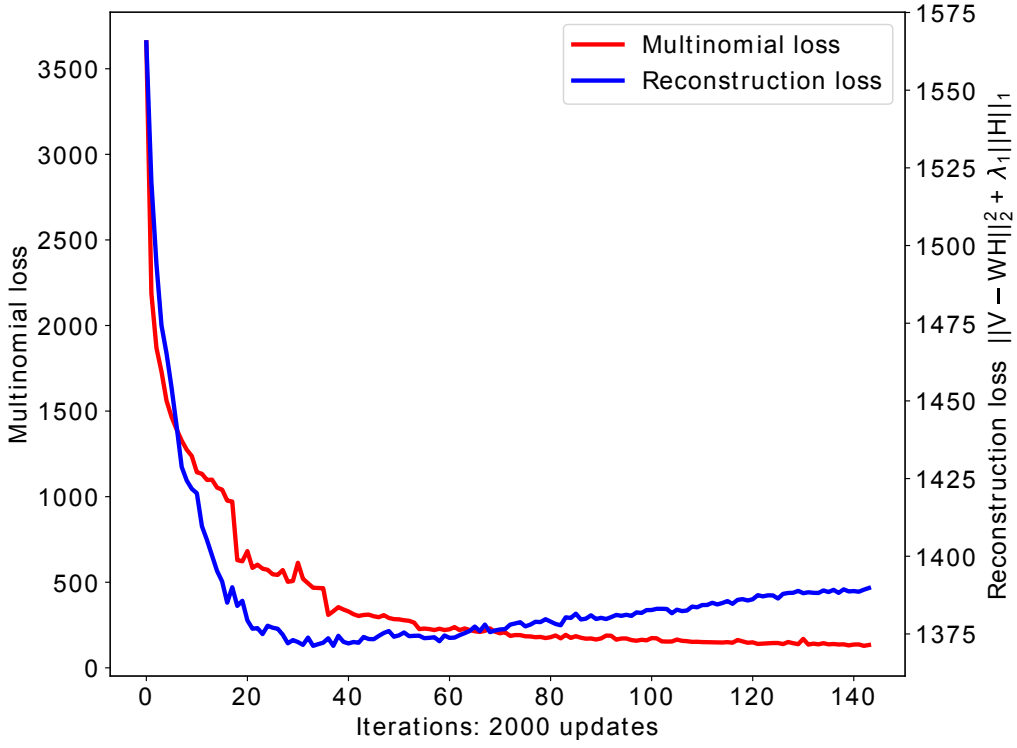
- $\Pi_{\mathcal{W}_+}$ projection operation on nonnegative dictionaries with unit ℓ_2 norm basis vectors

Code available on github.com/rserizel/TGNMF

Algorithm comparison

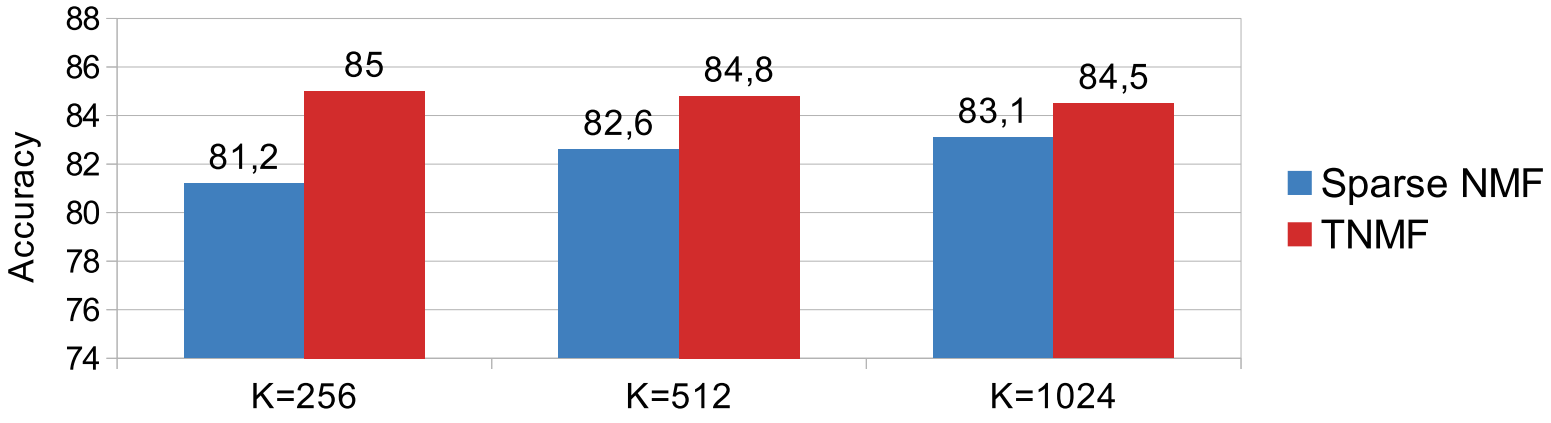
- Algorithm comparison on DCASE 2017 ASC dataset



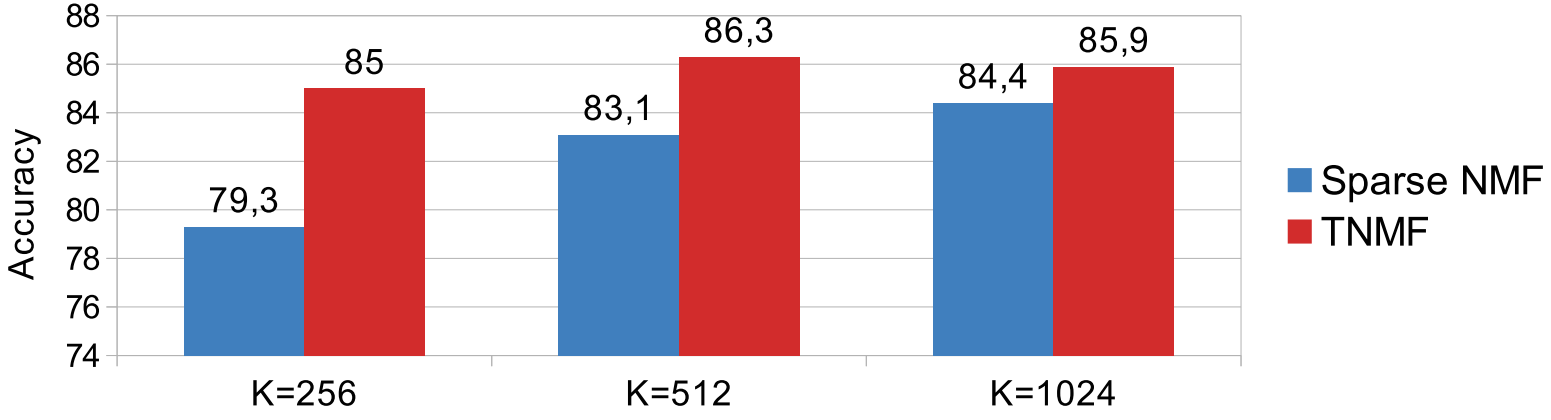


Scene classification results

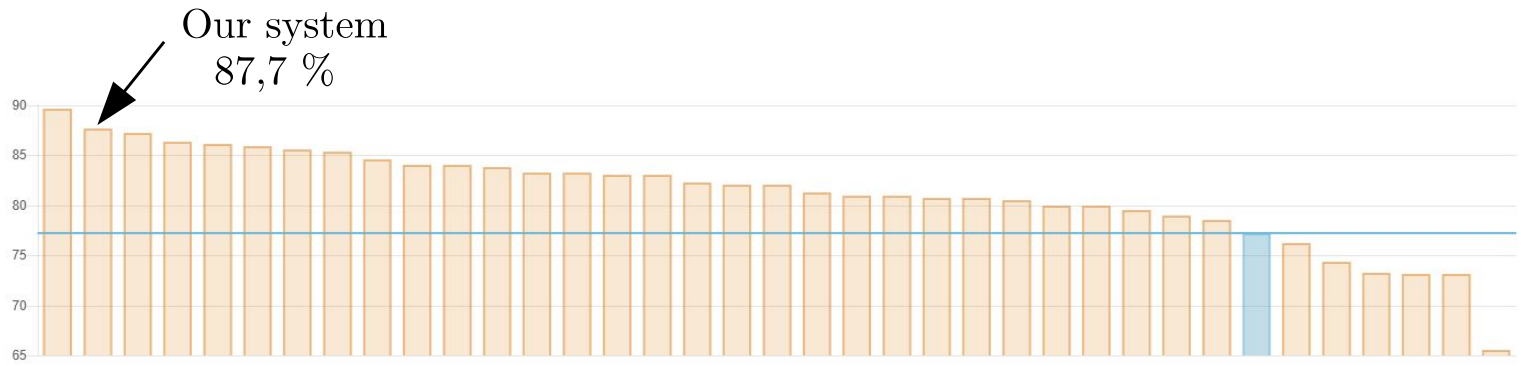
- DCASE 2016 development data set



- DCASE 2017 development data set



DCASE 2016 challenge



	Features	Classifier	Accuracy	Rank/49
[1]	Mel spectrum	CNN	86.2	7
[2]	Audio features	DNN fusion	86.4	5
[3]	Cepstral features	DNN + GMM	**87.2**	4
[4]	MFCC + spectrograms	CNN + I-vector	**89.7**	1
Our system	**CQT**	**TNMF fusion**	**87.7**	3

- TNMF system ranked 2nd out of 35 teams
- Above almost all deep learning-based systems

[1] Valenti et al. [2] Marchi et al. [3] Park et al. [4] Eghbal-Zadeh et al.

Outline

Unsupervised feature learning

Task-driven NMF

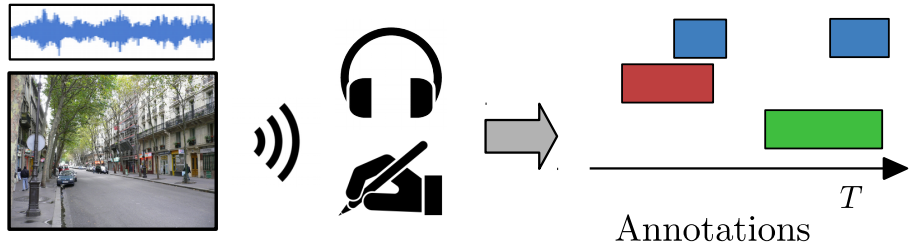
- TNMF model
- Event detection evaluation

Deep learning and NMF

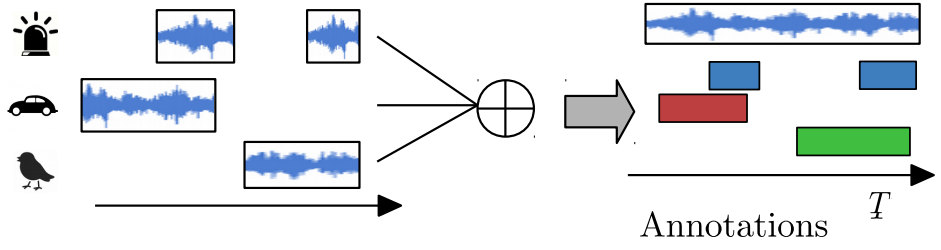
Conclusion

Sound event detection

- DCASE 2016 task3: event detection in real life conditions
  - → 12 recordings of 3 to 5 minutes per scene
  - → 2 environments: *home* and *residential area*
  - → 7 to 11 event categories per scene

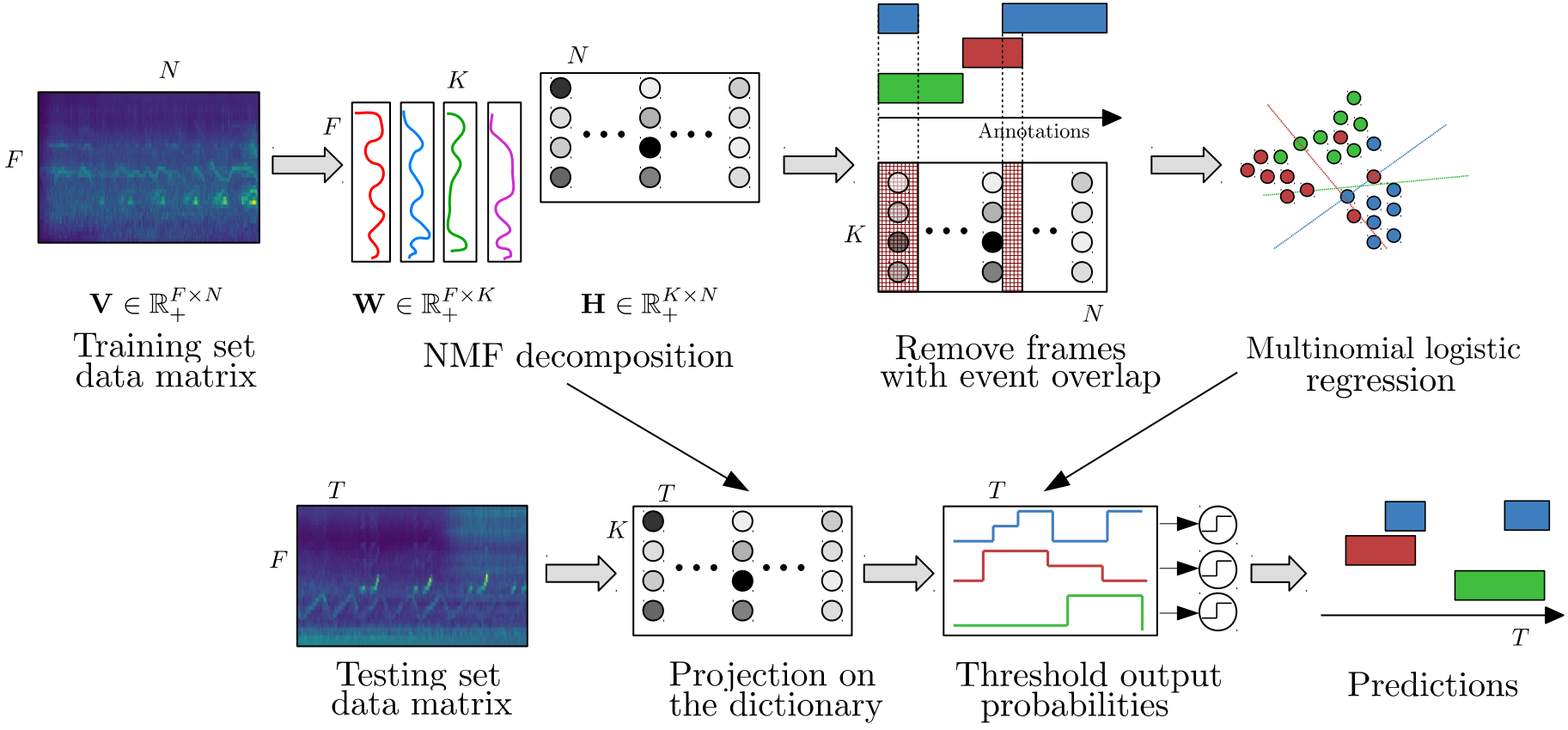


- TUT SED Synth 2016: Synthetic dataset
  - → 100 synthetic mixtures of 5 minutes each
  - → Mixtures of isolated events collected for online sound banks
  - → 16 different labels with various degrees of overlap



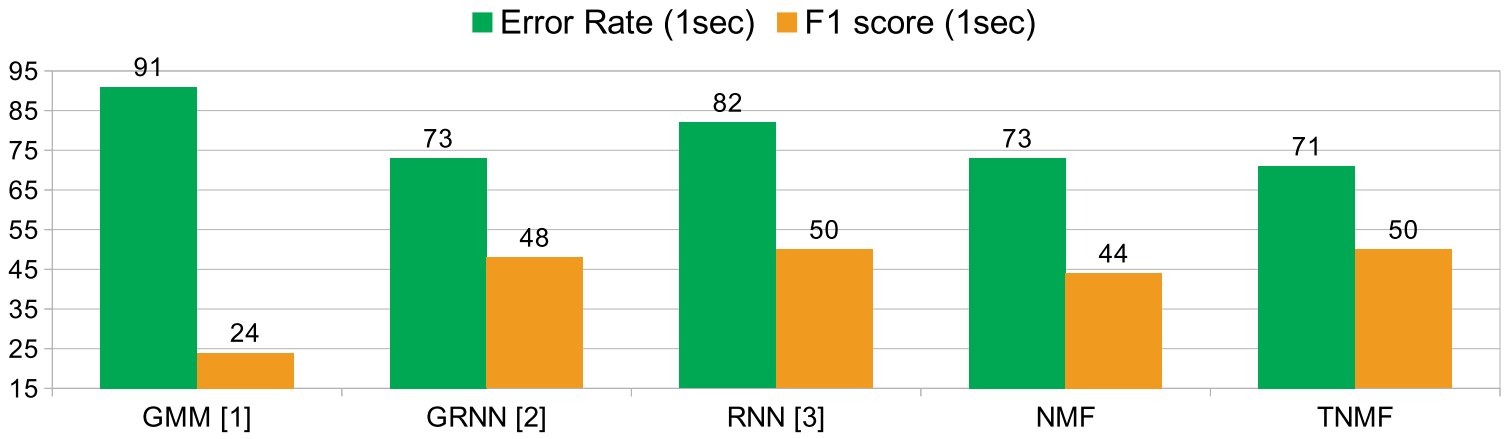
NMF-based system for event detection

- Turn multi-label problem into a simple multi-class problem for training
- Threshold output probabilities → predict multiple labels per frame



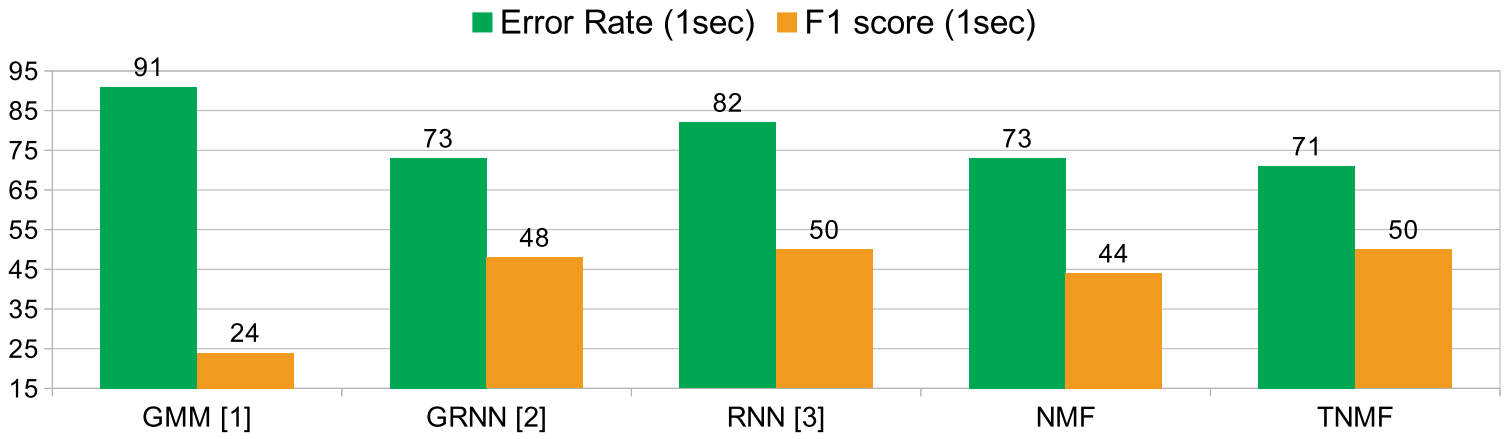
Event detection results

- DCASE 2016 Task3 dataset: event detection in real life conditions

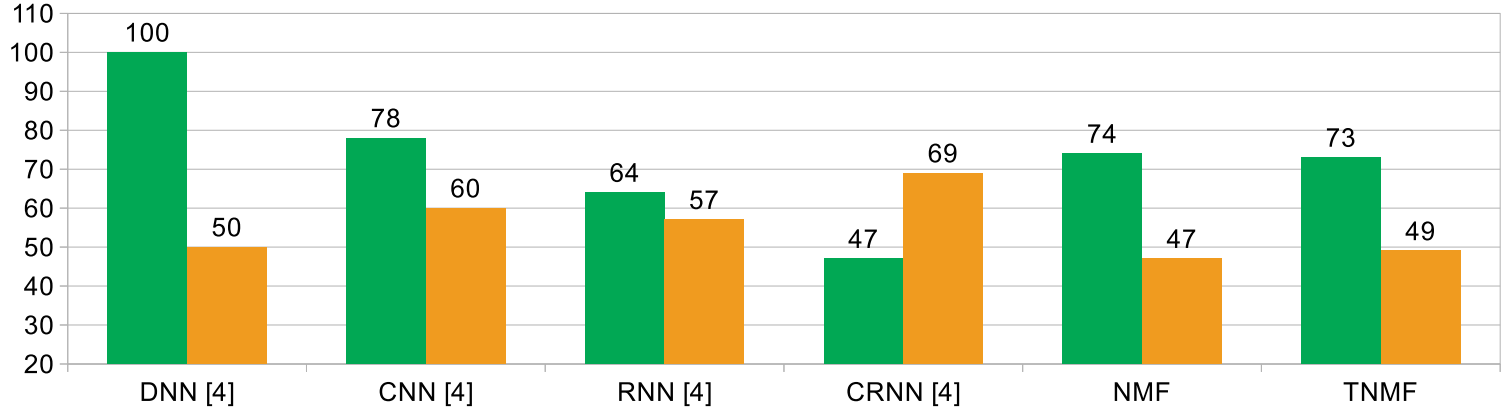


Event detection results

- DCASE 2016 Task3 dataset: event detection in real life conditions



- TUT SED synth 2016: synthetic event detection dataset



[1] Mesaros et al. 2016 [2] Zohrër and Penkopf, 2016 [3] Vu and Wang, 2016 [4] Cakir et al. 2017

Outline

Unsupervised feature learning

Task-driven NMF

Deep learning and NMF

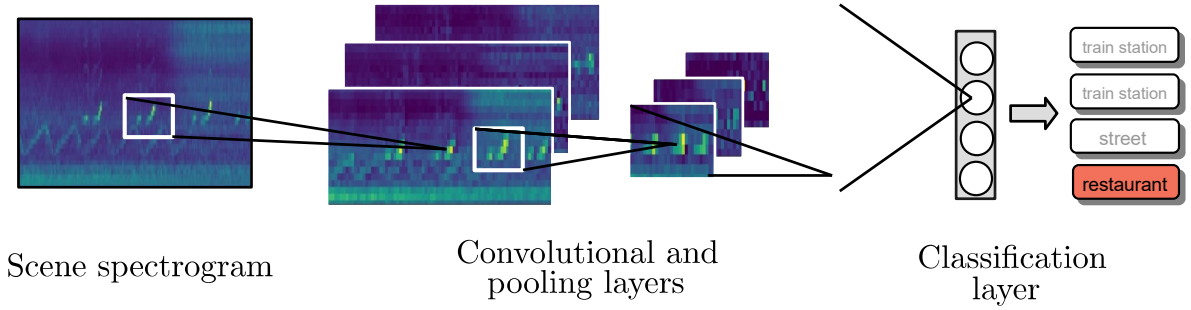
- Scene classification evaluation
- Event detection evaluation

Conclusion

DNN for sound scene analysis

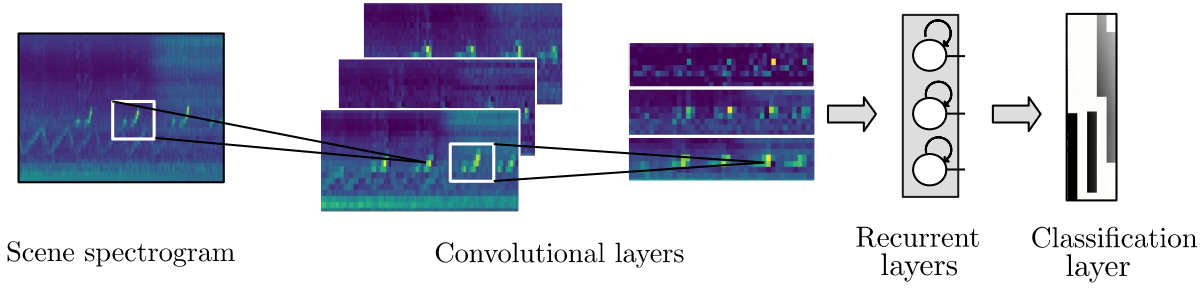
Scene and event classification

- Convolutional neural networks on time-frequency representations



Acoustic event detection

- Convolutional recurrent neural networks (CRNN) on time-frequency representations

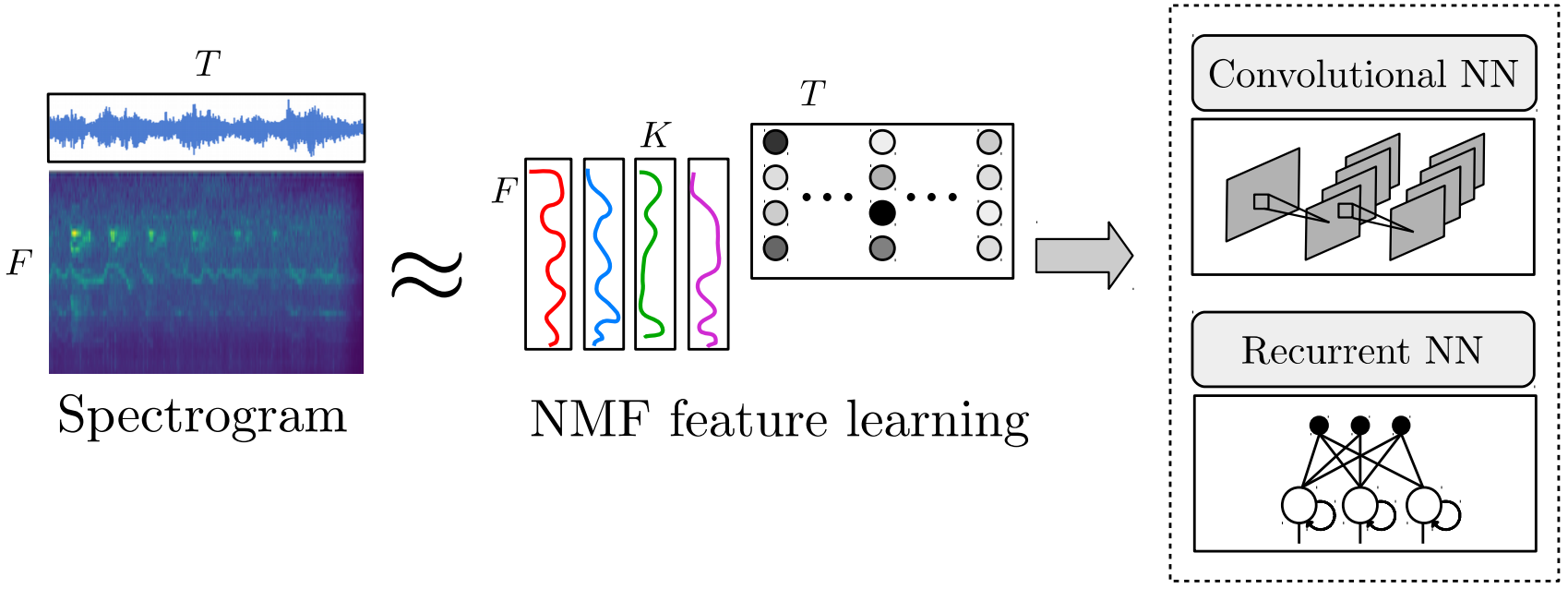


NMF as input of DNNs

- NMF $\rightarrow$ good representation learning tools for acoustic scenes
- Deep neural networks $\rightarrow$ powerful classification and detection models

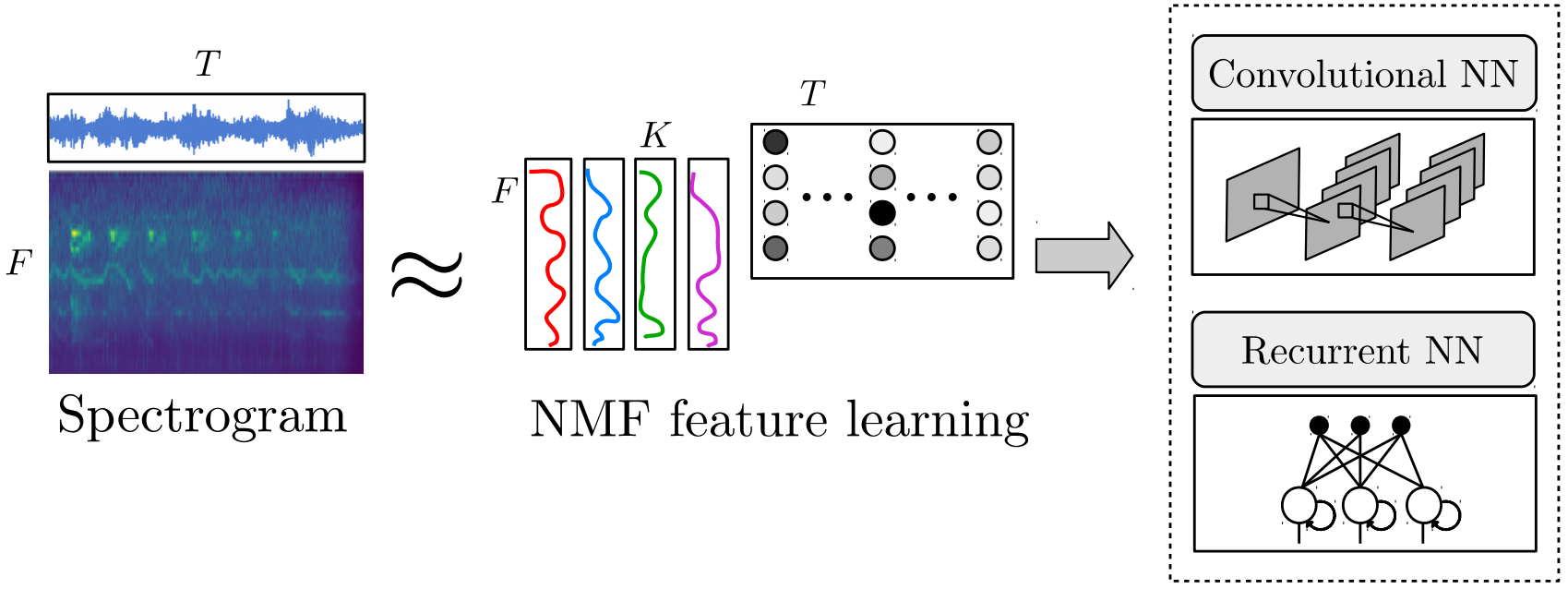
NMF as input of DNNs

- NMF $\rightarrow$ good representation learning tools for acoustic scenes
- Deep neural networks $\rightarrow$ powerful classification and detection models



NMF as input of DNNs

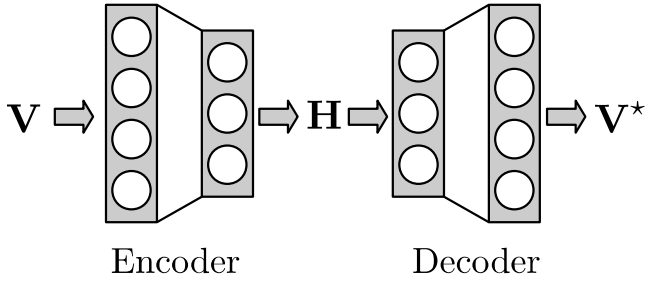
- NMF $\rightarrow$ good representation learning tools for acoustic scenes
- Deep neural networks $\rightarrow$ powerful classification and detection models



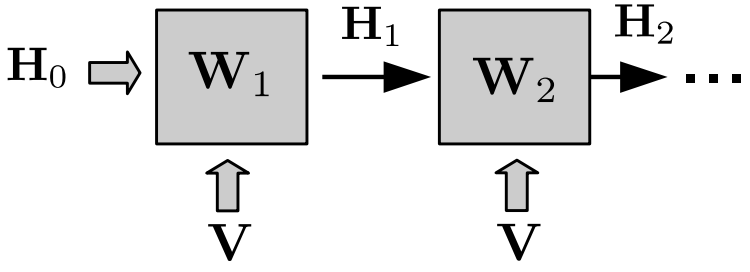
- Compatibility between NMF and standard fully-connected layers
- NMF features with RNN and CNN layers for event detection
- Jointly learning NMF and DNN parameters in the TNMF framework

Other works linking NMF to DNNs

- Replace NMF with deep auto-encoders for source separation [Smaragdis et al., 2017]

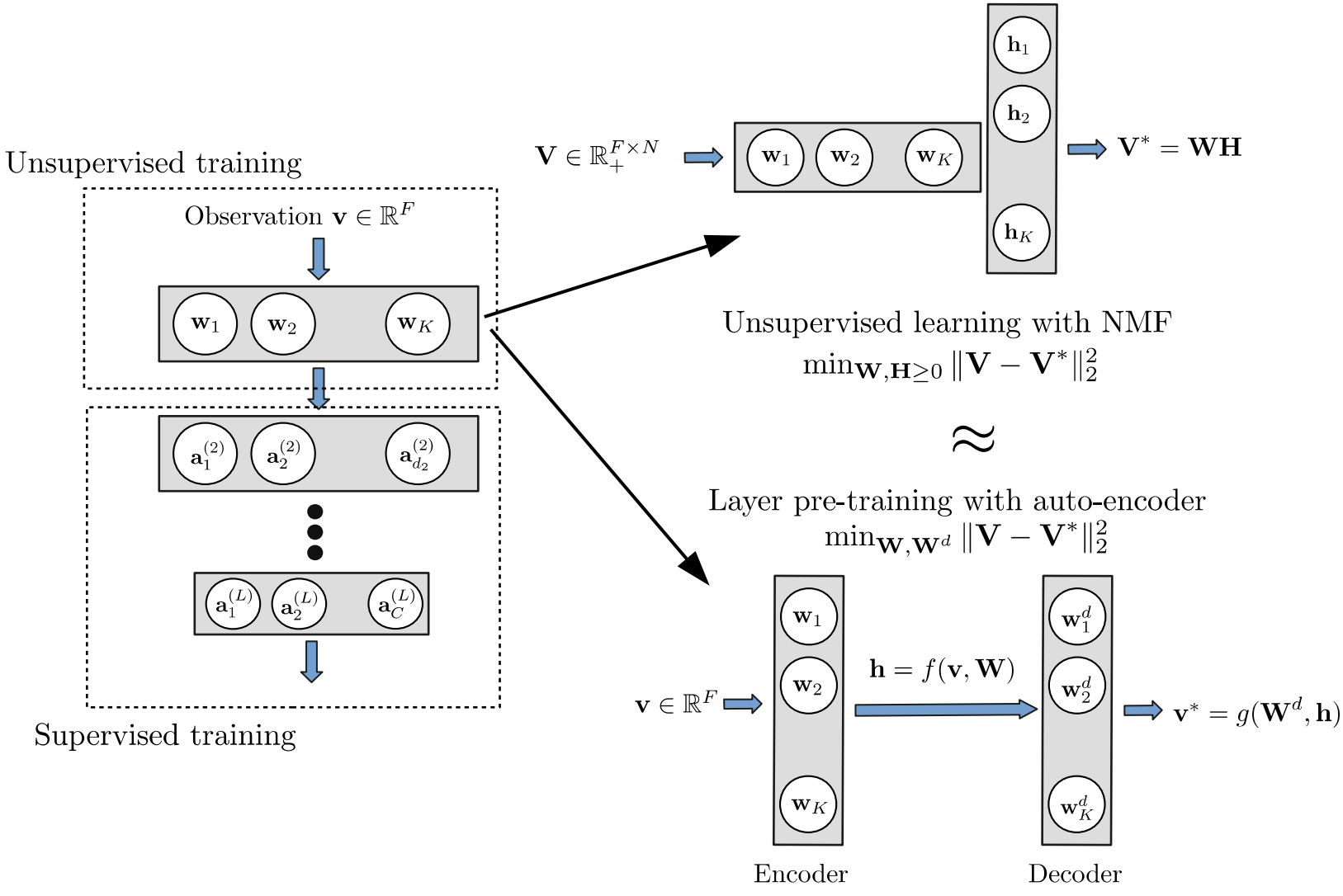


- Build a deep NMF by unfolding multiplicative update algorithm for NMF [Le Roux et al., 2015; Wisdom et al. 2017]

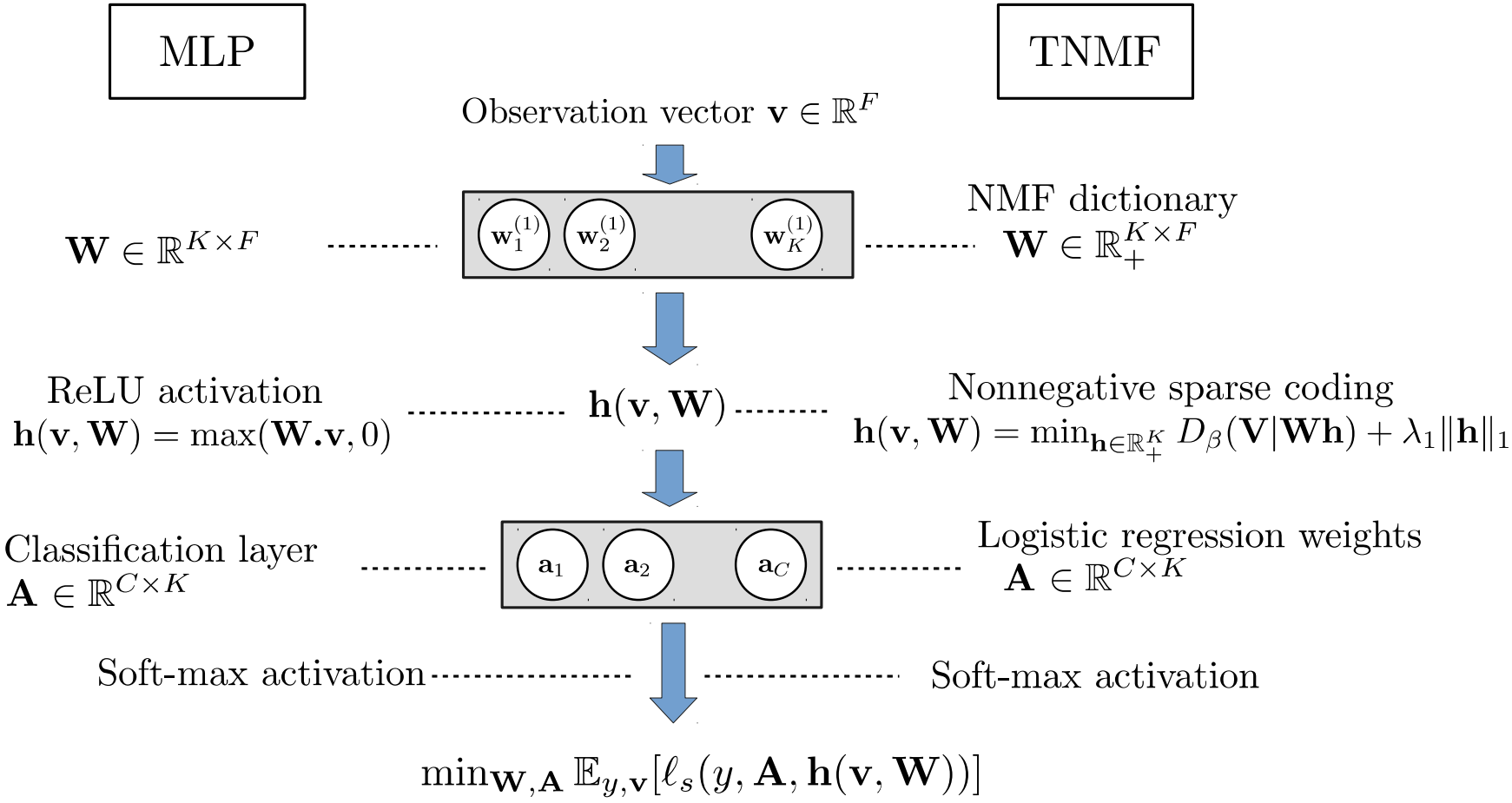


$$\mathbf{H}_t = \mathbf{H}_{t-1} \otimes \frac{\mathbf{W}_t^T((\mathbf{W}_t\mathbf{H}_{t-1})^{(\beta-2)} \otimes \mathbf{V}}{\mathbf{W}_t^T(\mathbf{W}_t\mathbf{H}_{t-1})^{(\beta-1)}}$$

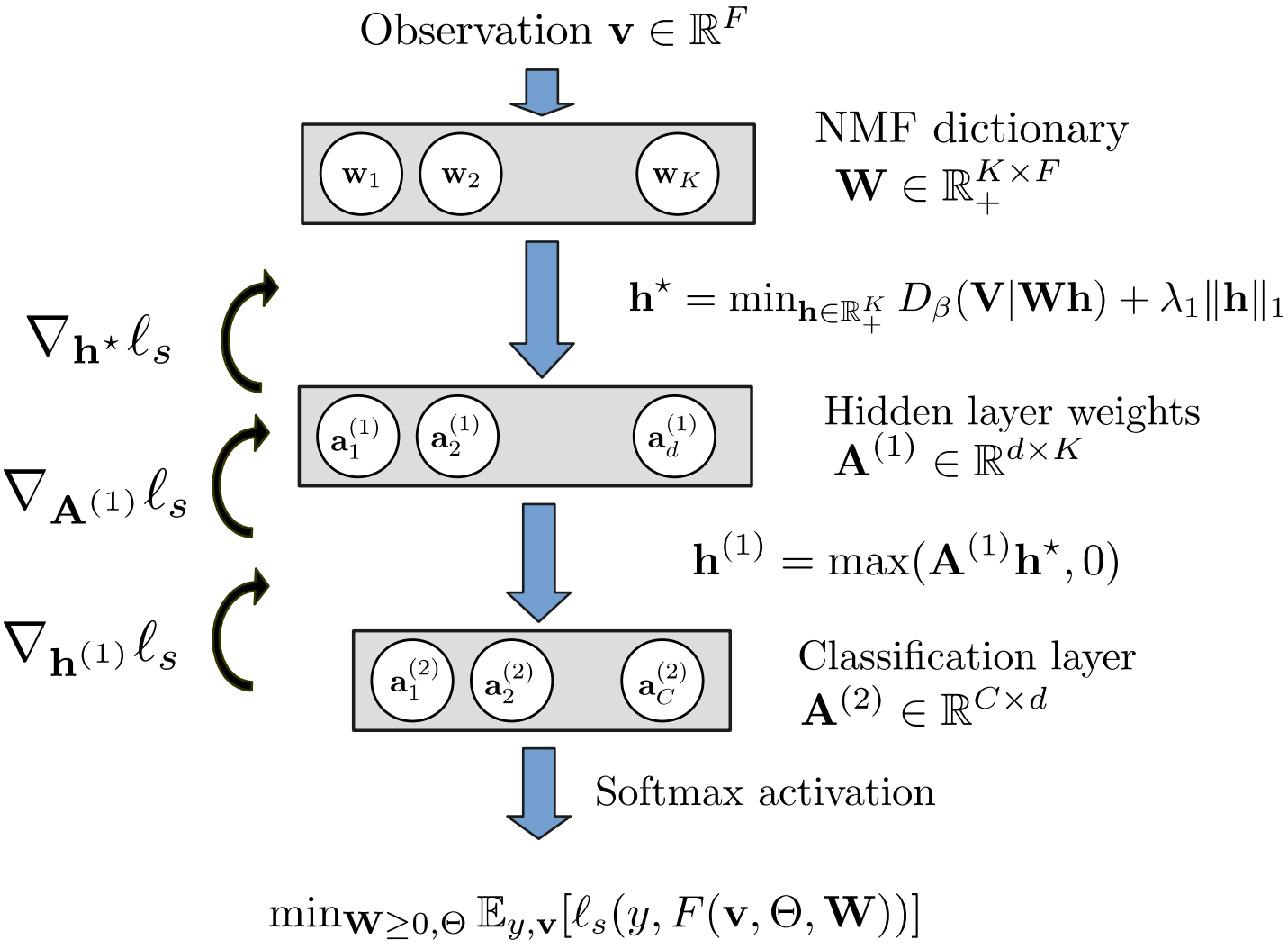
Layer-wise pre-training and NMF



TNMF = one hidden layer MLP



DNN-TNMF



$$\min_{\mathbf{W} \geq 0, \Theta} \mathbb{E}_{y,\mathbf{v}}[\ell_s(y, F(\mathbf{v}, \Theta, \mathbf{W}))]$$

Outline

Unsupervised feature learning

Task-driven NMF

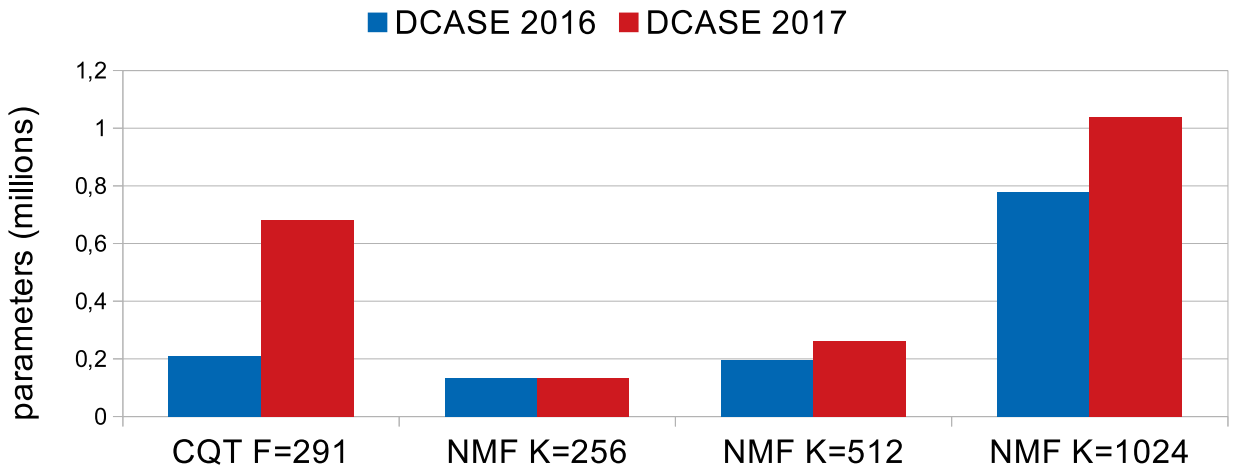
Deep learning and NMF
- Scene classification evaluation
- Event detection evaluation

Conclusion

DNN architecture search

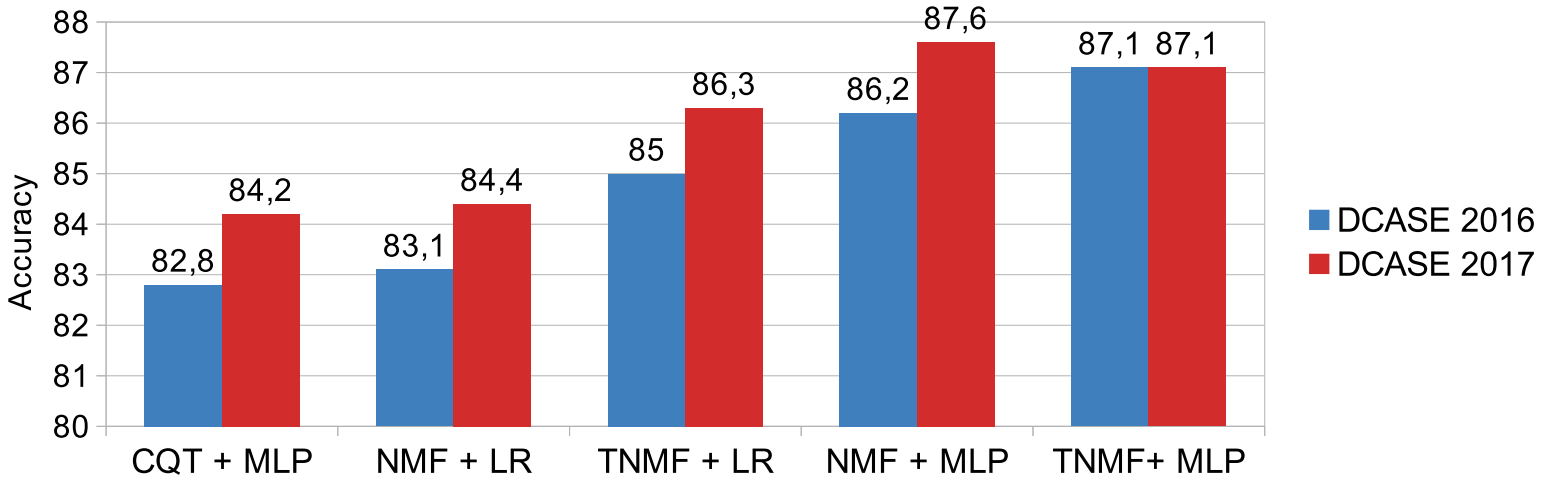
- Grid-search of network best architecture for all input representations
  - → Simple multi-layer perceptron as a classifier trained with SGD
  - → ReLU activations and dropout

	Dcase 2016		Dcase2017	
	Layers	Units	Layers	Units
CQT	3	256	3	512
NMF $K = 256$	2	256	2	256
NMF $K = 512$	2	256	3	256
NMF $K = 1024$	2	512	3	512



Scene classification results

- DCASE development sets results

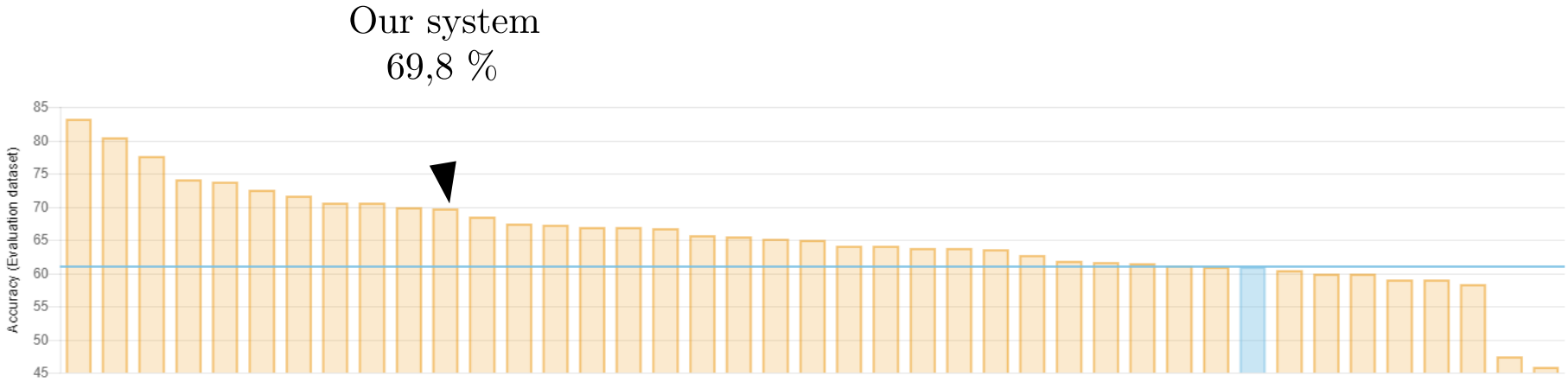


Dcase 2016 Challenge set				
	Input	Classifier	Accuracy	Rank/49
	TNMF	Logistic regression	87.7	3
[1]	MFCC + Mel	CNN + I vector	**89.7**	1
	CQT	DNN	86.7	-
	NMF	DNN	88.5	-
	TNMF	DNN	**90.5**	-

[1] Eghbal-Zadeh et al. 2016

Room for improvement

- DCASE 2017 challenge for scene classification



- More challenging conditions: mismatch between training and challenge set
  → Drop in performance for all methods on challenge set
- What did the top systems do that we didn't?
  - Use of augmentations and vary input representations
  - Include temporal modeling in neural networks (see NMF+CRNN)

Outline

Unsupervised feature learning

Task-driven NMF

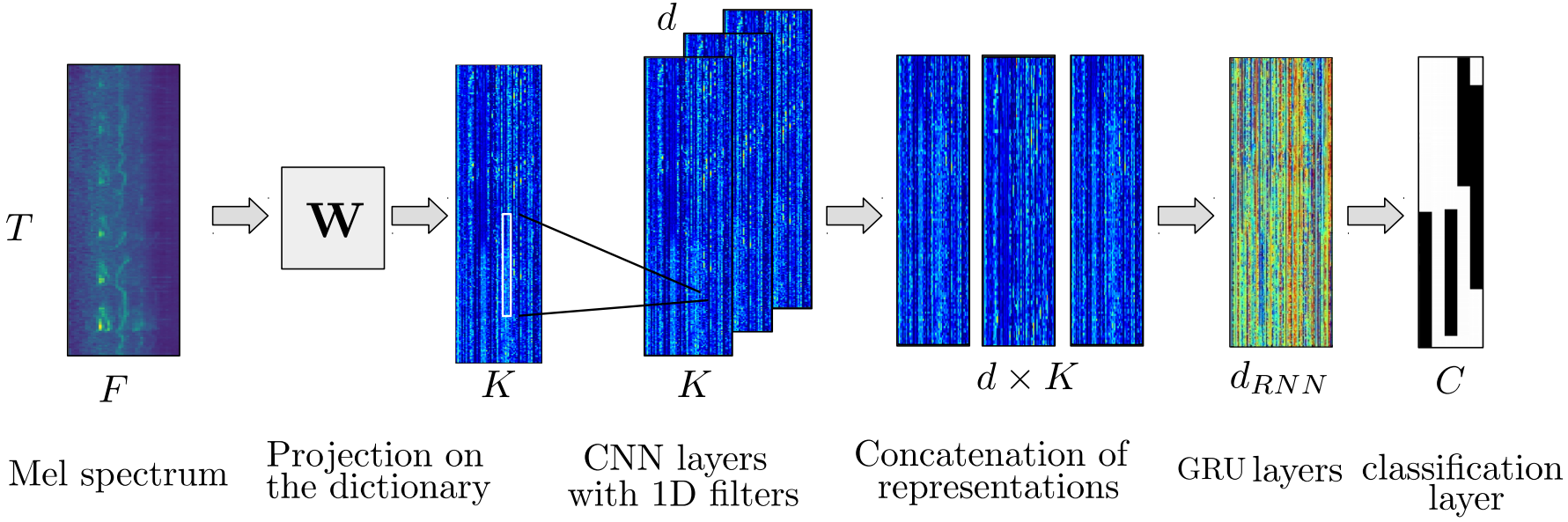
Deep learning and NMF

- Scene classification evaluation
- Event detection evaluation

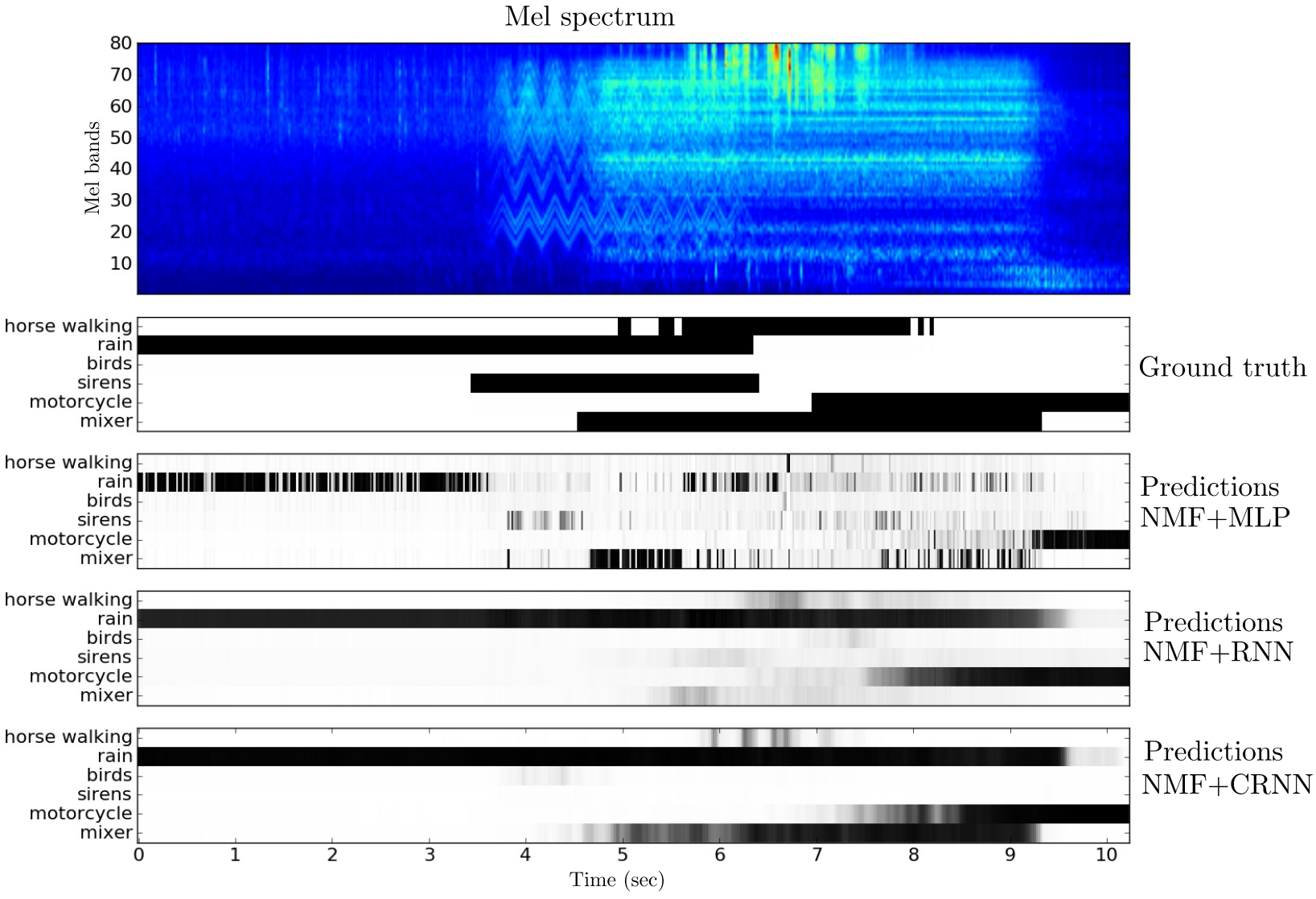
Conclusion

NMF+CRNN for event detection

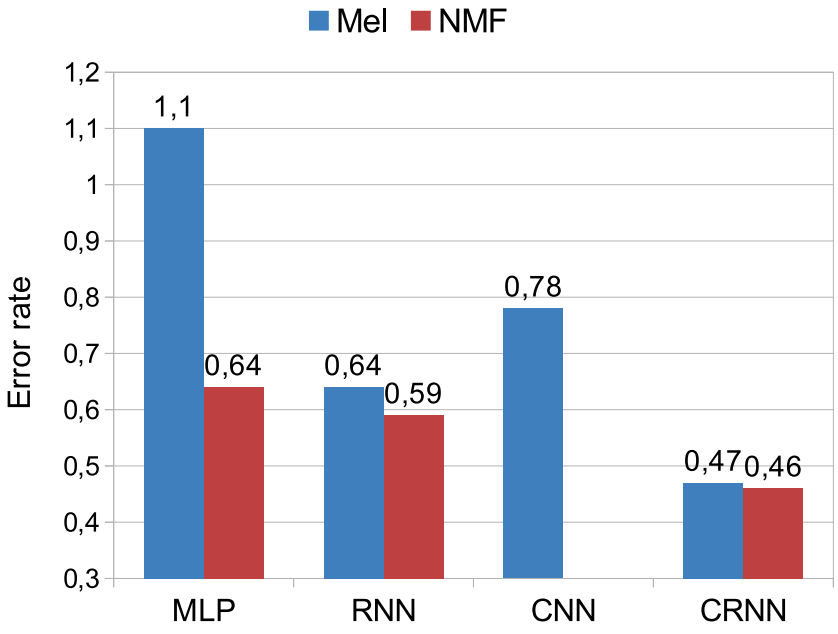
- Inspired from state of the art event detection CRNNs [Cakir et al., 2017]
- Separation of representation learning in two steps
  - → NMF representation on the frequency axis
  - → Model temporal information with CNN and RNN layers

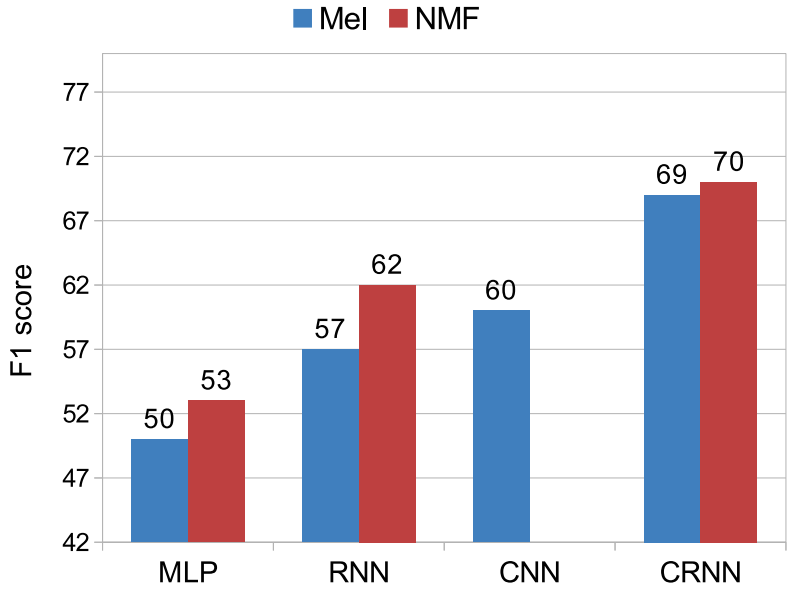


Results (1)

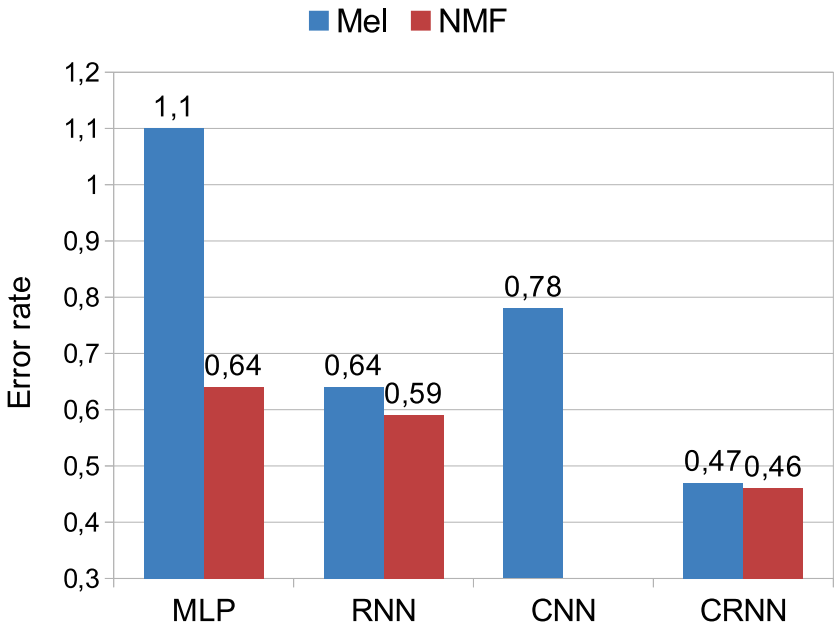

Mel spectrum
Mel bands
80
70
60
50
40
30
20
10
horse walking
rain
birds
sirens
motorcycle
mixer
Ground truth
Predictions
NMF+MLP
Predictions
NMF+RNN
Predictions
NMF+CRNN
0
1
2
3
4
5
6
7
8
9
10
Time (sec)

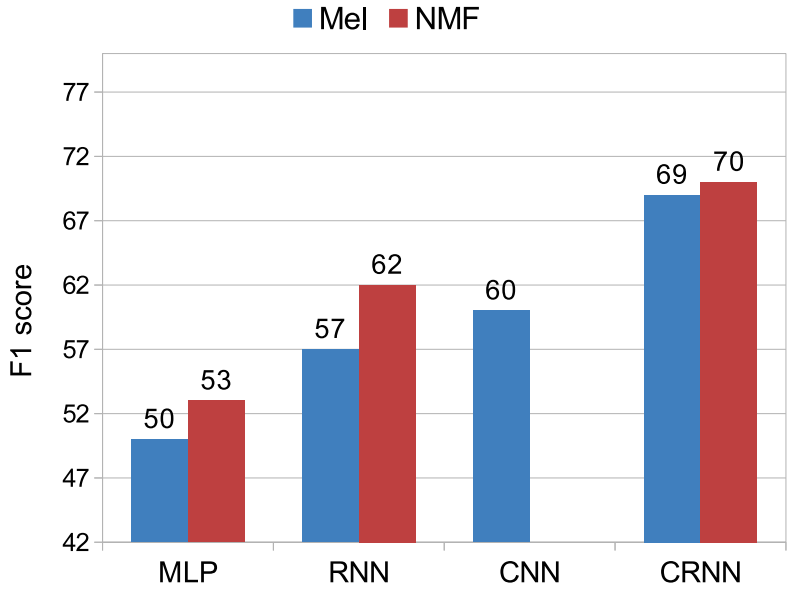
Results (2)

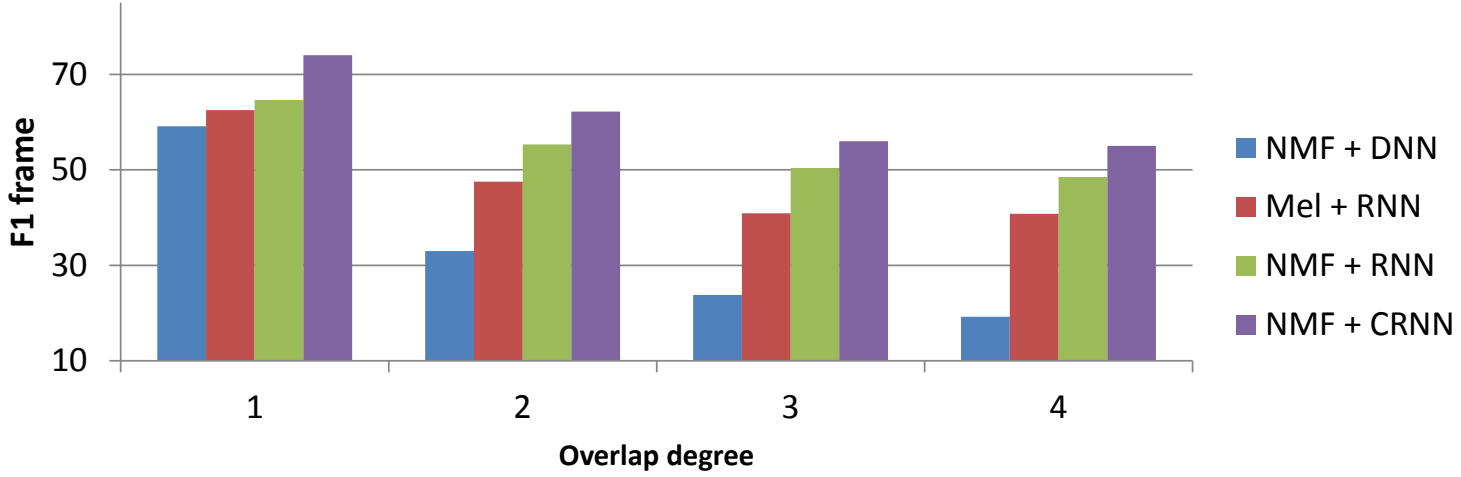

Mel
NMF
Error rate
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
1,1
0,64
0,64
0,59
0,78
0,47
0,46
MLP
RNN
CNN
CRNN


Mel
NMF
F1 score
77
72
67
62
57
52
47
42
50
53
57
62
60
69
70
MLP
RNN
CNN
CRNN

Results (2)


Mel
NMF
Error rate
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
1,1
0,64
0,64
0,59
0,78
0,47
0,46
MLP
RNN
CNN
CRNN

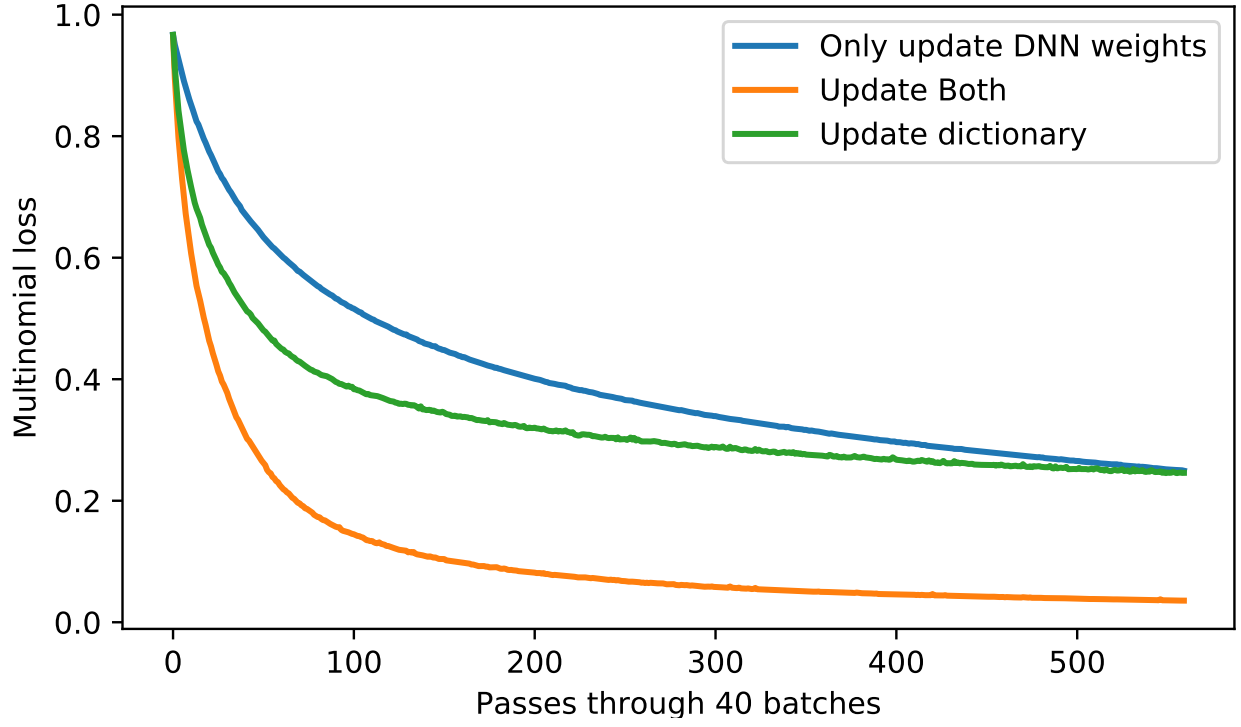

Mel
NMF
F1 score
77
72
67
62
57
52
47
42
50
53
57
62
60
69
70
MLP
RNN
CNN
CRNN


F1 frame
70
50
30
10
1
2
3
4
Overlap degree
NMF + DNN
Mel + RNN
NMF + RNN
NMF + CRNN

DNN-TNMF first results

- Jointly learn NMF and 1 hidden layer DNN for scene classification

	Dcase 2017			
	Sparse NMF	TNMF	NMF + DNN	DNN-TNMF
$K = 256$	79.3	85.0	84.3	85.5
$K = 512$	83.1	86.3	86.3	86.1



Outline

Unsupervised feature learning

Task-driven NMF

Deep learning and NMF

Conclusion

Summary of contributions

- A combination of spectrogram image features for improved scene classification
- Simple and efficient unsupervised NMF feature learning systems
  - → Clear benefit compared to hand-crafted audio features
- Adapt NMF feature learning to the task with TNMF
  - → Learn smaller more discriminative dictionaries
  - → An efficient alternative to deep learning on smaller datasets
- NMF as a better input to train deep networks
  - → Leave representation learning role to NMF for multi-source environments
  - → Allows for better performance with simpler networks
  - → Steps towards an end-to-end system with DNN-TNMF

Perspectives

- Make our systems more robust to noisy and unseen data
  - Vary input time-frequency representations to vary NMF features
  - Explore use of compatible augmentations with our approach
  - Simply perform fusion with various networks

Perspectives

- Make our systems more robust to noisy and unseen data
  - Vary input time-frequency representations to vary NMF features
  - Explore use of compatible augmentations with our approach
  - Simply perform fusion with various networks
- Large scale weakly-labeled data and tagging problems
  - Recent release of larger scale event classification datasets with Audioset
  - Adapt supervised NMF models to multiple instance learning for tagging
  - Extend our models to networks with attention classification layers

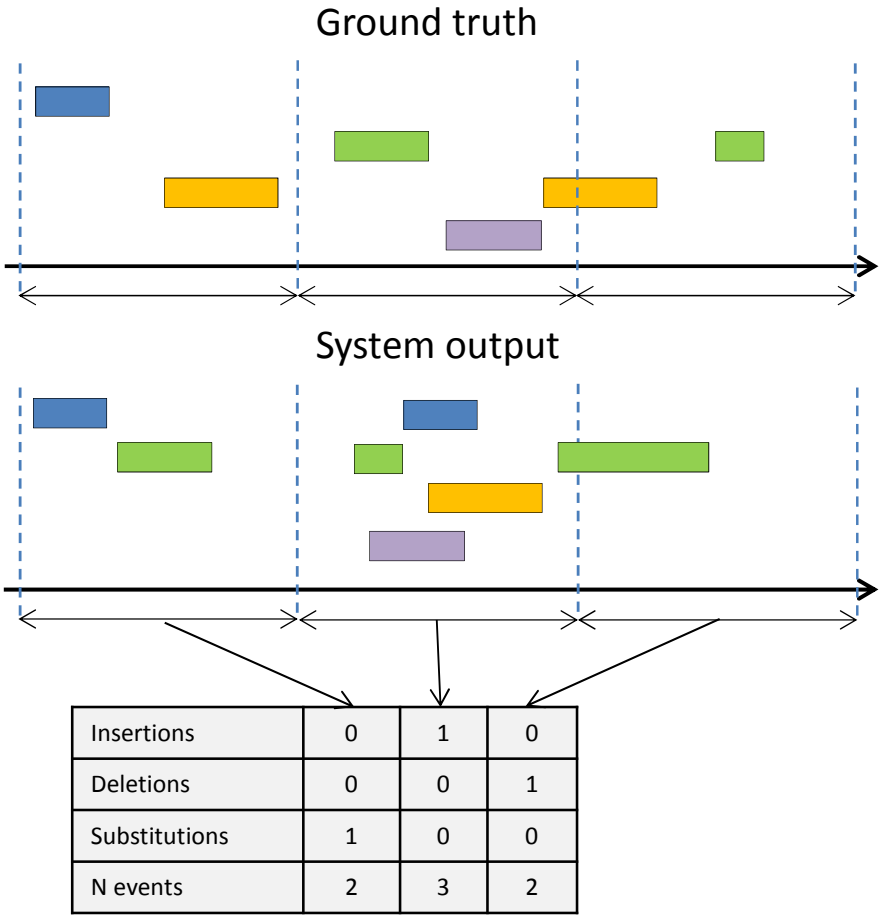
Perspectives

- Make our systems more robust to noisy and unseen data
  - Vary input time-frequency representations to vary NMF features
  - Explore use of compatible augmentations with our approach
  - Simply perform fusion with various networks
- Large scale weakly-labeled data and tagging problems
  - Recent release of larger scale event classification datasets with Audioset
  - Adapt supervised NMF models to multiple instance learning for tagging
  - Extend our models to networks with attention classification layers
- Towards an end-to-end approach with DNN-TNMF
  - Exploration of efficient algorithms on large scale datasets
  - Study compatibility with recurrent and convolutional layers

Publications

- Journal article and book chapter
  - V. Bisot, R. Serizel, S. Essid et G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2017.
  - R. Serizel, V. Bisot, S. Essid et G. Richard, "Acoustic features for environmental sound analysis," in *Computational analysis of sound scenes and events*, Springer, 2018.
- Conference proceedings
  - V. Bisot, R. Serizel, S. Essid et G. Richard, "Nonegative feature learning methods for acoustic scene classification," in *Proc. of DCASE Workshop*, 2017.
  - V. Bisot, R. Serizel, S. Essid et G. Richard, "Leveraging deep neural networks with nonnegative representations for improved environmental sound classification," in *Proc. of MLSP*, 2017.
  - V. Bisot, R. Serizel, S. Essid et G. Richard, "Overlapping sound event detection with supervised nonnegative matrix factorization," in *Proc. of ICASSP*, 2017.
  - V. Bisot, R. Serizel, S. Essid et G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proc. of ICASSP*, 2016.
  - V. Bisot, S. Essid et G. Richard, "HOG and Subband Power Distribution Image Features for Acoustic Scene Classification," in *Proc. of EUSIPCO*, 2015.

Evaluation metrics



Insertions	0	1	0
Deletions	0	0	1
Substitutions	1	0	0
N events	2	3	2

Total Error Rate [1]

$$\mathrm{ER} = \frac{\sum_k \mathrm{I}_k + \sum_k \mathrm{S}_k + \sum_k \mathrm{D}_k}{\sum_k \mathrm{N}_k}$$

Different variants to model different aspects

Sparse NMF: promote sparse activations

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \sum_i D_\beta(\mathbf{v}_i, \sum_k \mathrm{h}_{ki} \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}) + \lambda_1 \sum_{i,k} \mathrm{h}_{ik},$$

Convolutive NMF: temporal context in the decomposition

$$\mathbf{S} \approx \sum_{t=0}^{\tau-1} \mathbf{W}_t \overset{t\rightarrow}{\mathbf{H}},$$

- $\mathbf{S} \in \mathbb{R}_+^{F\times T}$ a scene spectrogram
- $\mathbf{W}_t[:,k] \rightarrow$ frame t for basis k

Kernel decompositions: decomposing non-linearly separable data

$$\Phi(\mathbf{V}) \approx \Phi(\mathbf{V})\mathbf{W}\mathbf{H}$$

- Φ kernel mapping function

Event detection settings

- Draw fixed length sequences randomly for training
- Trained with ADAM algorithm with early stopping
- Parameter search for all networks on development set

NMF + X	**MLP**	**RNN**	**CRNN**
MLP layers	3	2	-
GRU layers	-	3	2
CNN layers	-	-	3
MLP/RNN units	256	512	512
CNN units	-	-	32
CNN filter size	-	-	10
Input sequence length (sec)	-	5	10

Adaptive scaling TNMF algorithm

- Goal: scale the projections for each mini-batch with updated statistics to improve model training and performance
  - N the number of training examples in $\mathbf{V}$ divided into B batches $\mathbf{V}_b$
  - (μ, σ) mean and standard deviation of optimal projection features $\mathbf{H}$
  - (μ_b, σ_b) mean and standard mini-batch $\mathbf{H}_b$

Classifier update

- Compute and scale optimal projection on full training data $\mathbf{H}^{'} = \frac{1}{\sigma}(\mathbf{H}^{\star} - m)$
- Update $\mathbf{A}$ with one iteration of L-BFGS for logistic regression

Dictionary update on one epoch

1. Draw a random data point $\mathbf{v}$ with label y
2. Compute optimal projection $\mathbf{H}_b^{\star}(\mathbf{V}_b, \mathbf{W})$ and statistics (μ_b, σ_b)
3. Update global statistics $m = m - \frac{1}{B}(m - m_b)$ and $\sigma^2 = \sigma^2 - \frac{1}{B}(\sigma^2 - \sigma_b^2)$
4. Scale mini-batch projections $\mathbf{H}_b^{'} = \frac{1}{\sigma}(\mathbf{H}_b - m)$
5. Update $\mathbf{W}$ by projected gradient with $\mathbf{H}_b^{'}$ as previously