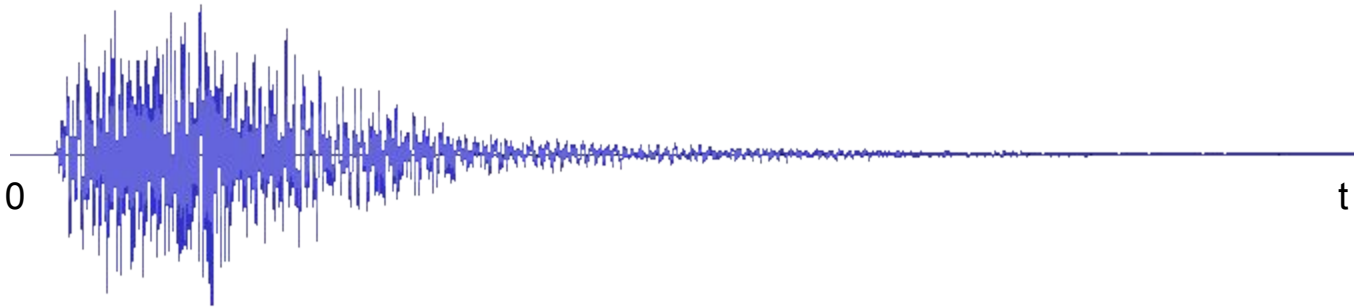
Sound event detection using weakly labeled dataset with convolutional and recurrent neural network

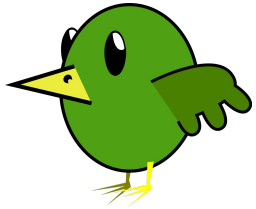
Sharath Adavanne, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland

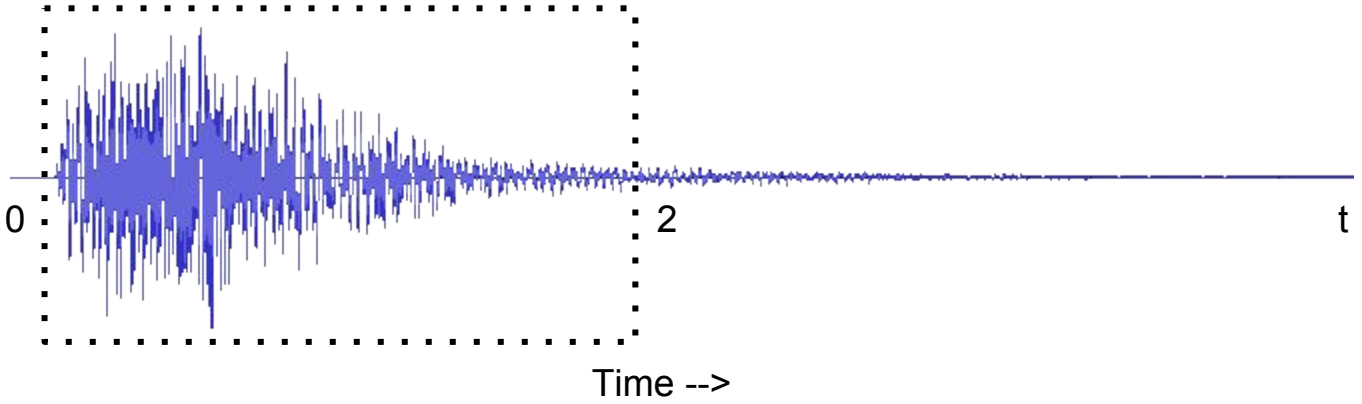
Outline

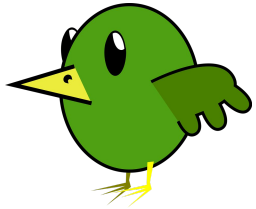
- Introduction
  - Sound tagging / Weak labels
  - Sound event detection / Strong labels
- Dataset
- Proposed neural network
- Results


0
t
Time -->

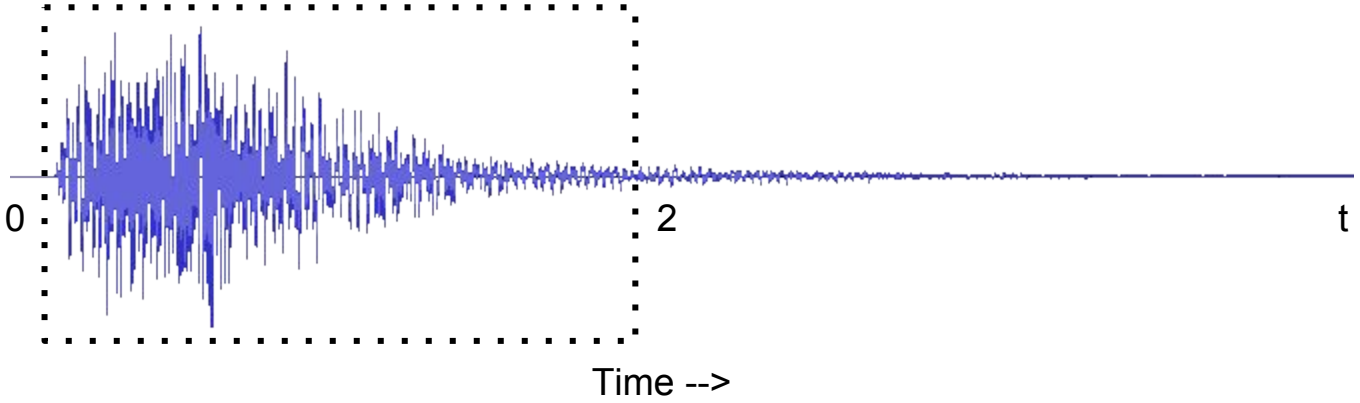


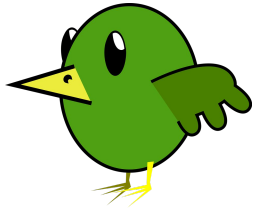
Sound event detection



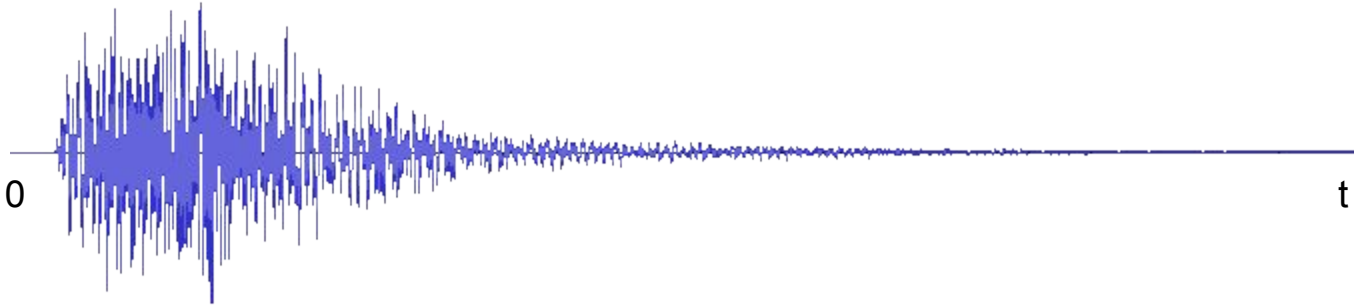


Strong label
Sound event detection

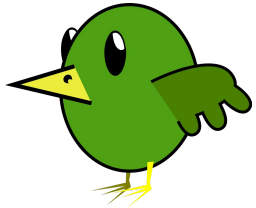




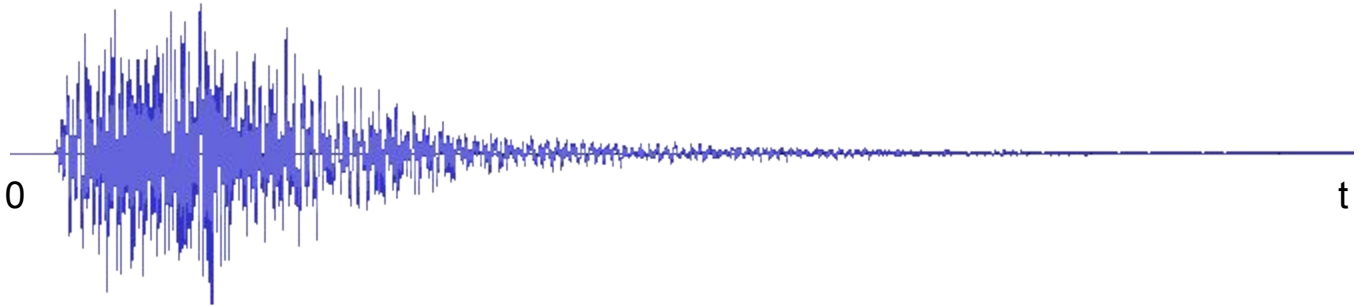
Weak label



Time -->



Weak label
Sound tagging



Time -->

Sound event detection using weakly labeled dataset with convolutional and recurrent neural network

Sound event detection using weakly labeled dataset with convolutional and recurrent neural network

Sound event detection using sound tagging dataset with convolutional and recurrent neural network

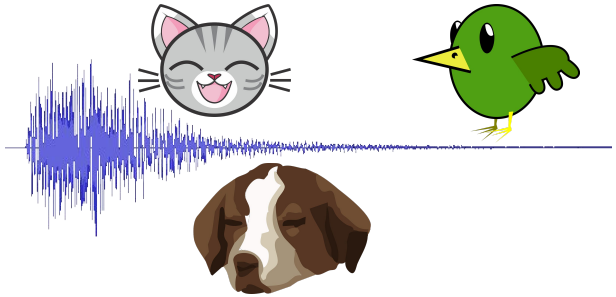
Dataset

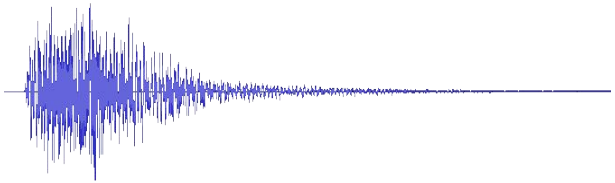
- Subset of Google’s AudioSet
- 51,172 for training, 488 for testing
- 10s clips (padded with zeros, if not 10s)

Dataset

- Subset of Google's AudioSet
- 51,172 for training, 488 for testing
- 10s clips (padded with zeros, if not 10s)

- Weak labeled
- 17 classes - Car, Bus, Train, Truck etc..
- Single recording can have more than one sound source

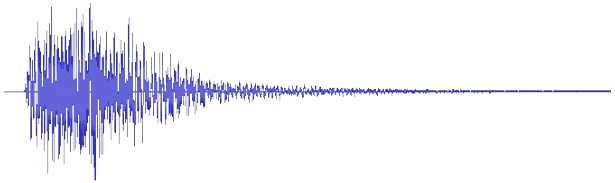




Neural network

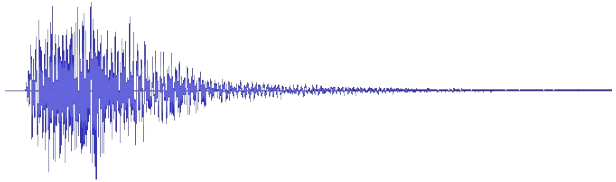
Weak label
Sound Tagging

Training procedure



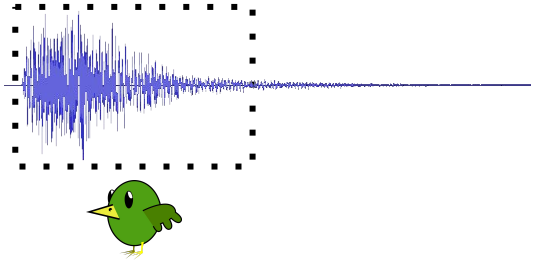
Training

Weak label
Sound Tagging



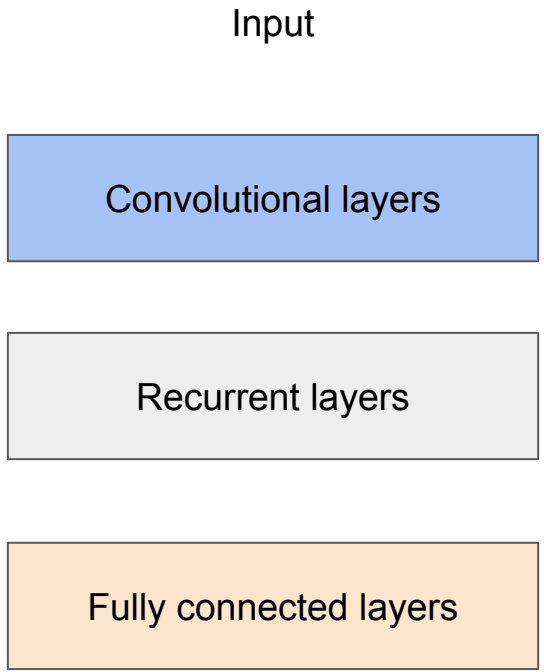
Testing

Sound tagging

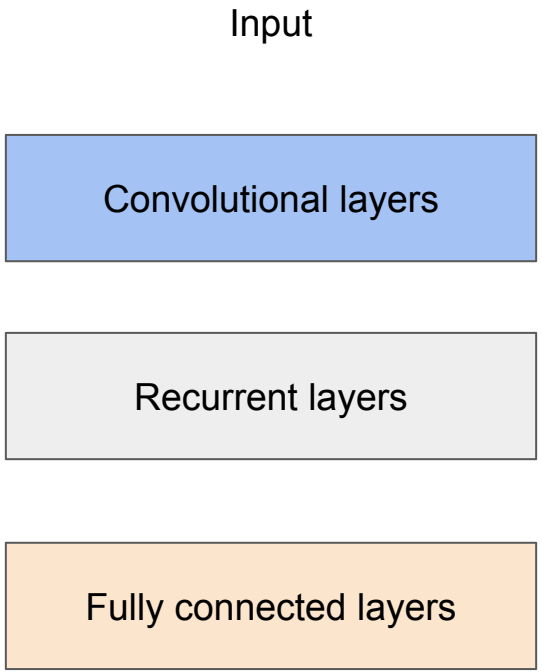


Sound event detection

Convolutional and recurrent neural network (CRNN) architecture



Convolutional and recurrent neural network (CRNN) architecture



Why?

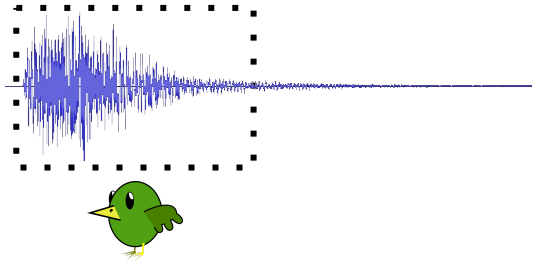
Input

Convolutional and recurrent neural network (CRNN) architecture

Convolutional layers

Recurrent layers

Fully connected layers

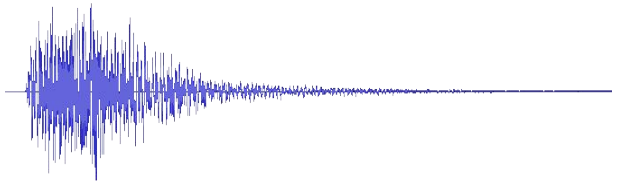


Sound tagging

Output

Sound event detection

Convolutional and recurrent neural network (CRNN) architecture for sound event detection from weak labels

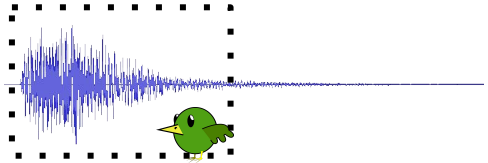


Input

Convolutional layers

Recurrent layers

Fully connected layers

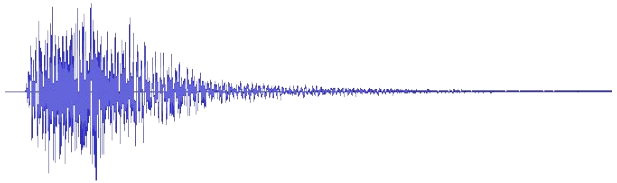


Output 1 : Sound event detection

Fully connected layers

Output 2 : Sound tagging

Convolutional and recurrent neural network (CRNN) architecture for sound event detection from weak labels



Input

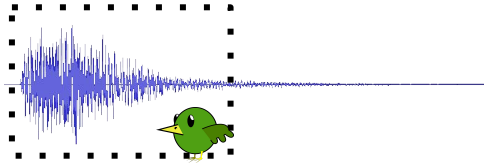
Log mel-band energies of complete 10 second audio

Convolutional layers

Recurrent layers

Fully connected layers

Time-distributed layers: Sequence input and sequence output



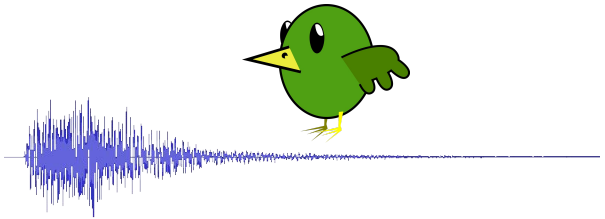
Output 1 : Sound event detection

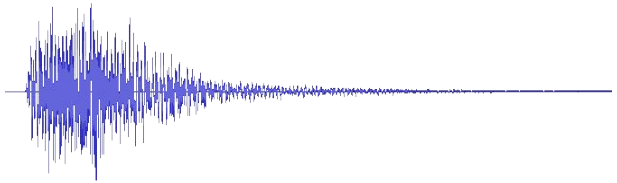
Fully connected layers

Sequence input and number of classes output

Output 2 : Sound tagging

Convolutional and recurrent neural network (CRNN) architecture for sound event detection from weak labels





Input

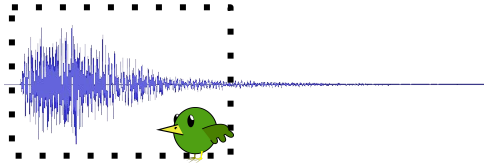
Log mel-band energies of complete 10 second audio

Convolutional layers

Recurrent layers

Fully connected layers

Time-distributed layers: Sequence input and sequence output



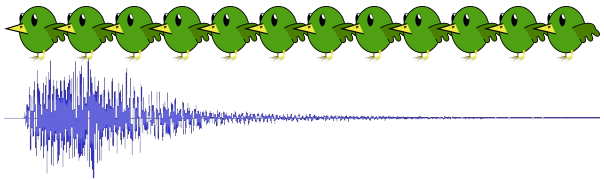
Output 1 : Sound event detection

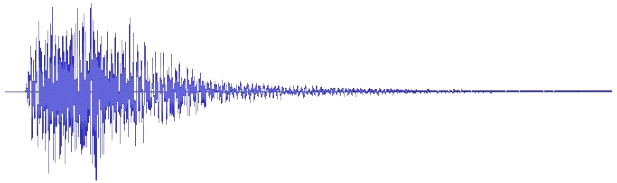
Fully connected layers

Sequence input and number of classes output

Output 2 : Sound tagging

Convolutional and recurrent neural network (CRNN) architecture for sound event detection from weak labels





Input

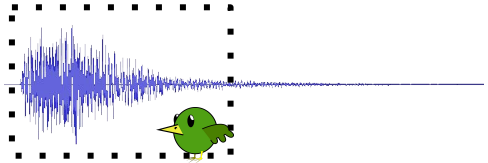
Log mel-band energies of complete 10 second audio

Convolutional layers

Recurrent layers

Fully connected layers

Time-distributed layers: Sequence input and sequence output

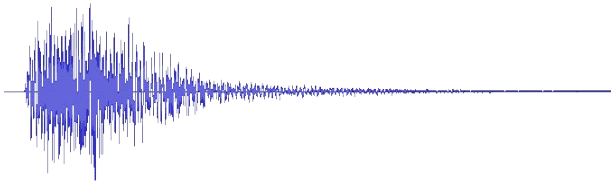


Output 1 : Sound event detection

Fully connected layers

Sequence input and number of classes output

Output 2 : Sound tagging



Input

Log mel-band energies of complete 10 second audio

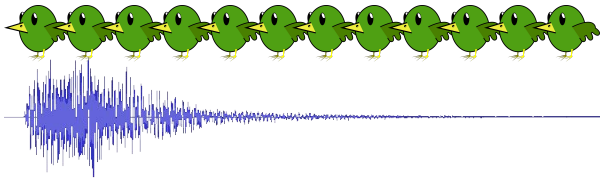
Convolutional and recurrent neural network (CRNN) architecture for sound event detection from weak labels

Convolutional layers

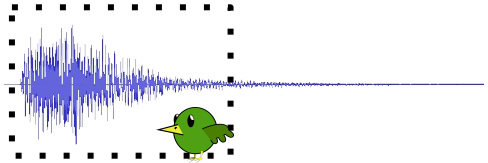
Recurrent layers

Fully connected layers

Time-distributed layers: Sequence input and sequence output



Loss 1 : strong label loss



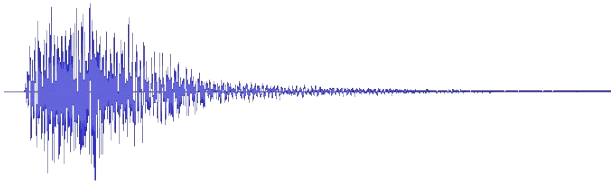
Output 1 : Sound event detection

Fully connected layers

Sequence input and number of classes output

Loss 2 : weak label loss

Output 2 : Sound tagging



Input

Log mel-band energies of complete 10 second audio

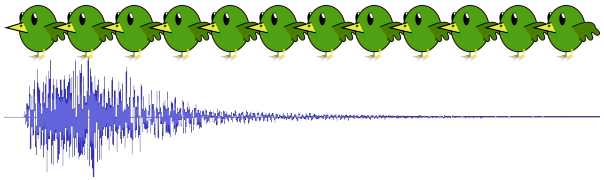
Convolutional and recurrent neural network (CRNN) architecture for sound event detection from weak labels

Convolutional layers

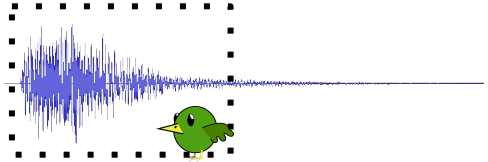
Recurrent layers

Fully connected layers

Time-distributed layers: Sequence input and sequence output



Loss 1 : strong label loss



Output 1 : Sound event detection

Fully connected layers

Sequence input and number of classes output

Loss 2 : weak label loss

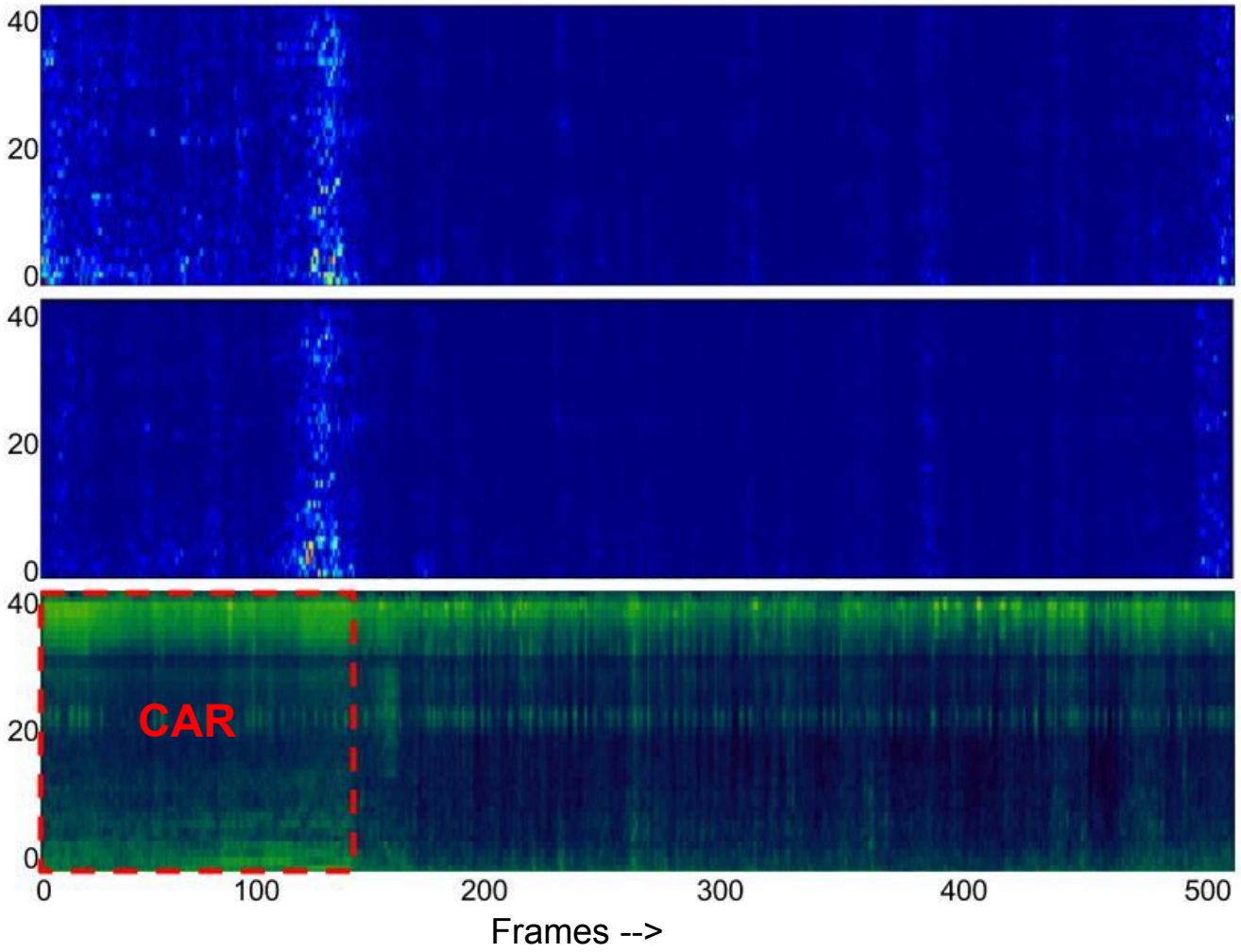
Output 2 : Sound tagging

Saliency map for –jc0NAxK8M_30.000_40.000 recording

Gradients for strong label

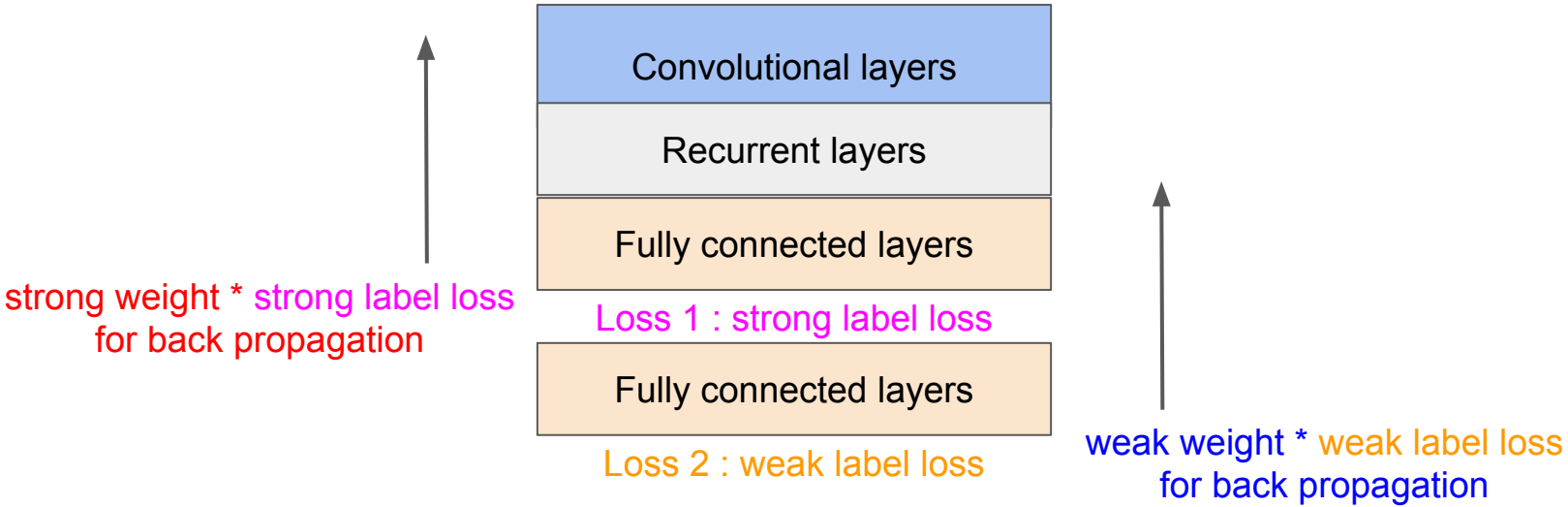
Gradients for weak label

Log mel-band energy



Results

Strong Weight	Weak Weight	Sound tagging				Sound Event detection		
		Precision	Recall	F-score		Error rate	F-score	
0.002	1	44.9	37.0	40.5		1.38	10.9	



Results

Strong Weight	Weak Weight	Sound tagging				Sound Event detection		
		Precision	Recall	F-score		Error rate	F-score	
0.002	1	44.9	37.0	40.5		1.38	10.9	
0.02	1	44.2	36.5	40.0		1.13	17.0	
0.2	1	**47.5**	39.6	43.2		0.84	38.1	
1	1	**47.5**	**39.7**	**43.3**		0.84	38.8	
1	0.2	47.3	39.5	43.0		0.84	38.6	
1	0.02	25.5	20.6	22.8		**0.81**	41.1	
1	0.002	20.5	16.5	18.3		**0.81**	**42.4**	

Results

Strong Weight	Weak Weight	Sound tagging				Sound Event detection		
		Precision	Recall	F-score		Error rate	F-score	
0.002	1	44.9	37.0	40.5		1.38	10.9	
0.02	1	44.2	36.5	40.0		1.13	17.0	
0.2	1	**47.5**	39.6	43.2		0.84	38.1	
1	1	**47.5**	**39.7**	**43.3**		0.84	38.8	
1	0.2	47.3	39.5	43.0		0.84	38.6	
1	0.02	25.5	20.6	22.8		**0.81**	41.1	
1	0.002	20.5	16.5	18.3		**0.81**	**42.4**	

Results

Strong Weight	Weak Weight	Sound tagging				Sound Event detection		
		Precision	Recall	F-score		Error rate	F-score	
0.002	1	44.9	37.0	40.5		1.38	10.9	
0.02	1	44.2	36.5	40.0		1.13	17.0	
0.2	1	**47.5**	39.6	43.2		0.84	38.1	
1	1	**47.5**	**39.7**	**43.3**		0.84	38.8	
1	0.2	47.3	39.5	43.0		0.84	38.6	
1	0.02	25.5	20.6	22.8		**0.81**	41.1	
1	0.002	20.5	16.5	18.3		**0.81**	**42.4**	

Results

Strong Weight	Weak Weight	Sound tagging				Sound Event detection		
		Precision	Recall	F-score	F-score chlng	Error rate	F-score	Error rate chlng
0.002	1	44.9	37.0	40.5		1.38	10.9	
0.02	1	44.2	36.5	40.0		1.13	17.0	
0.2	1	**47.5**	39.6	43.2	**46.6**	0.84	38.1	0.80
1	1	**47.5**	**39.7**	**43.3**	45.5	0.84	38.8	0.81
1	0.2	47.3	39.5	43.0	44.5	0.84	38.6	0.82
1	0.02	25.5	20.6	22.8		**0.81**	41.1	
1	0.002	20.5	16.5	18.3	26.3	**0.81**	**42.4**	**0.79**

Results

Strong Weight	Weak Weight	Sound tagging				Sound Event detection		
		Precision	Recall	F-score	F-score chlng	Error rate	F-score	Error rate chlng
0.002	1	44.9	37.0	40.5		1.38	10.9	
0.02	1	44.2	36.5	40.0		1.13	17.0	
0.2	1	**47.5**	39.6	43.2	**46.6**	0.84	38.1	0.80
1	1	**47.5**	**39.7**	**43.3**	45.5	0.84	38.8	0.81
1	0.2	47.3	39.5	43.0	44.5	0.84	38.6	0.82
1	0.02	25.5	20.6	22.8		**0.81**	41.1	
1	0.002	20.5	16.5	18.3	26.3	**0.81**	**42.4**	**0.79**
					55.6			0.66

Outline

- **Introduction**
  - Sound tagging / Weak labels
  - Sound event detection / Strong labels

Outline

- **Introduction**
  - Sound tagging / Weak labels
  - Sound event detection / Strong labels
- **Dataset**
  - Weak labels
  - More than one label for each recording

Outline

- **Introduction**
  - Sound tagging / Weak labels
  - Sound event detection / Strong labels
- **Dataset**
  - Weak labels
  - More than one label for each recording
- **Proposed neural network**
  - Convolutional recurrent neural network (CRNN)
  - Two sequential outputs - strong label followed by weak

Outline

- **Introduction**
  - Sound tagging / Weak labels
  - Sound event detection / Strong labels
- **Dataset**
  - Weak labels
  - More than one label for each recording
- **Proposed neural network**
  - Convolutional recurrent neural network (CRNN)
  - Two sequential outputs - strong label followed by weak
- **Results**
  - CRNN was shown to learn temporal information, given just weak labels
  - Best result was against our intuition: equal scaling of strong and weak loss

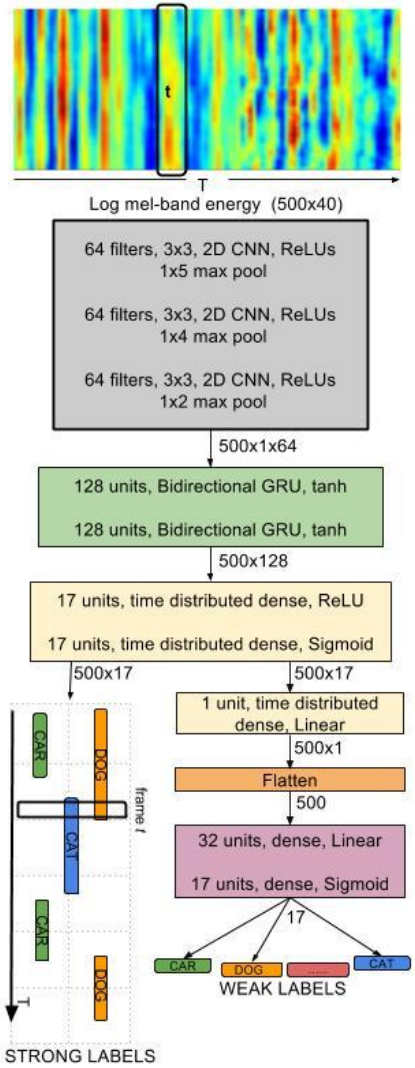
Outline

- **Introduction**
  - Sound tagging / Weak labels
  - Sound event detection / Strong labels
- **Dataset**
  - Weak labels
  - More than one label for each recording
- **Proposed neural network**
  - Convolutional recurrent neural network (CRNN)
  - Two sequential outputs - strong label followed by weak
- **Results**
  - CRNN was shown to learn temporal information, given just weak labels
  - Best result was against our intuition: equal scaling of strong and weak loss
- **Future work**
  - More fine tuning
  - Strong labels for high energy regions only
  - Attention layers

Thank you.

Sharath Adavanne, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland



218,000 trainable weights