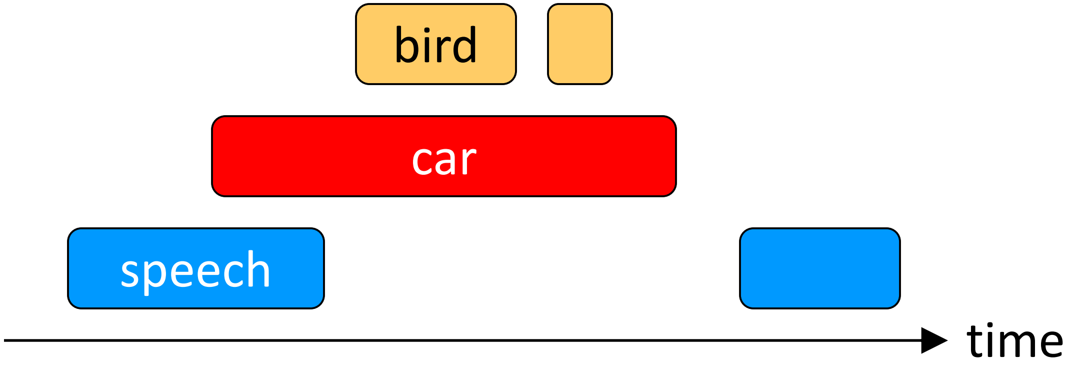
A **Comparison** of **Five Multiple Instance Learning** Pooling Functions for Sound Event Detection with **Weak Labeling**

Yun Wang, **Juncheng Li**, Florian Metze

May 14, 2019

Sound Event Detection

- Detection = audio tagging + localization



- Strong labeling is expensive to obtain

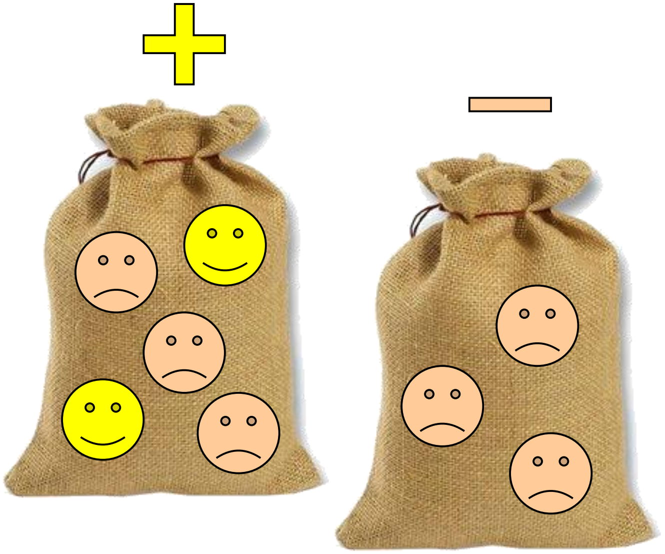
Sound Event Detection

- Train with weak labeling

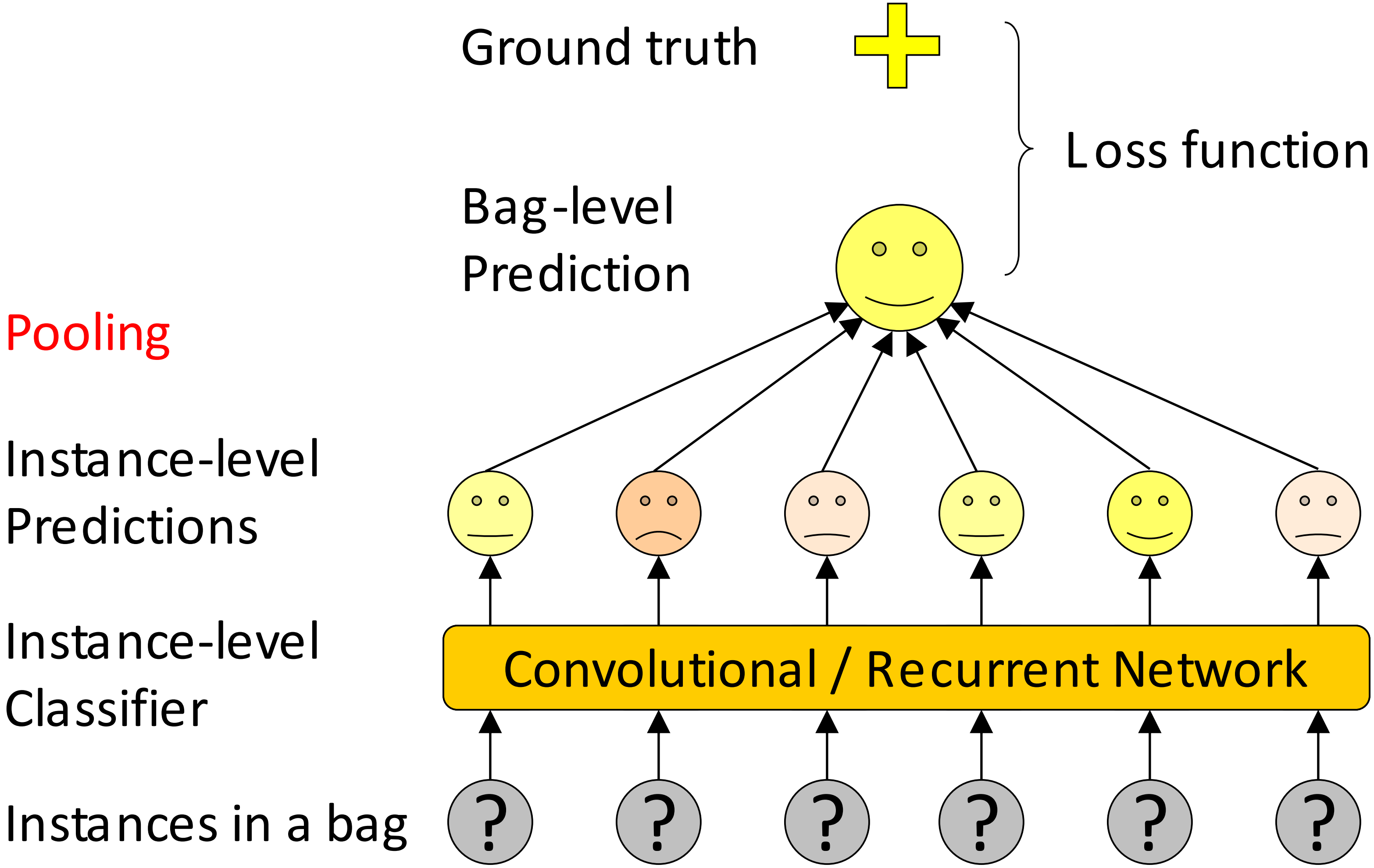
- But still, we want both tagging and localization output

Multiple Instance Learning

- SED with weak labeling is a Multiple Instance Learning (MIL) problem
  - Bag is positive ⇔ any instance is positive
  - Recording = bag, frames = instances



Multiple Instance Learning


Ground truth
Loss function
Bag-level Prediction
Pooling
Instance-level Predictions
Instance-level Classifier
Convolutional / Recurrent Network
Instances in a bag
?
?
?
?
?
?

Pooling Functions

1

Max pooling

$$y = \max_i y_i$$

3

Linear softmax

$$y = \frac{\sum_i y_i^2}{\sum_i y_i}$$

4

Exp. softmax

$$y = \frac{\sum_i y_i \exp(y_i)}{\sum_i \exp(y_i)}$$

2

Average pooling

$$y = \frac{1}{n} \sum_i y_i$$

Weighted Average

One frame gets all the weight

Larger probs get larger weight

All frames get equal weight

$$y = \frac{\sum_i y_i w_i}{\sum_i w_i}$$

5

Attention:

Learn the weights!

Pooling Functions

- We found linear softmax best for localization!

$$y = \frac{\sum_i y_i^2}{\sum_i y_i} \qquad \frac{\partial y}{\partial y_i} = \frac{2y_i - y}{\sum_j y_j}$$

Positive when $y_i > y/2$

- When bag is positive:
  - □ y_i gets away from $y/2$ ☺
  - □ Only boosts frames with $y_i > y/2$ – nice localization!
- When bag is negative:
  - □ y_i approaches $y/2$ – finally converges to zero ☺

Pooling Functions

- What's wrong with attention?

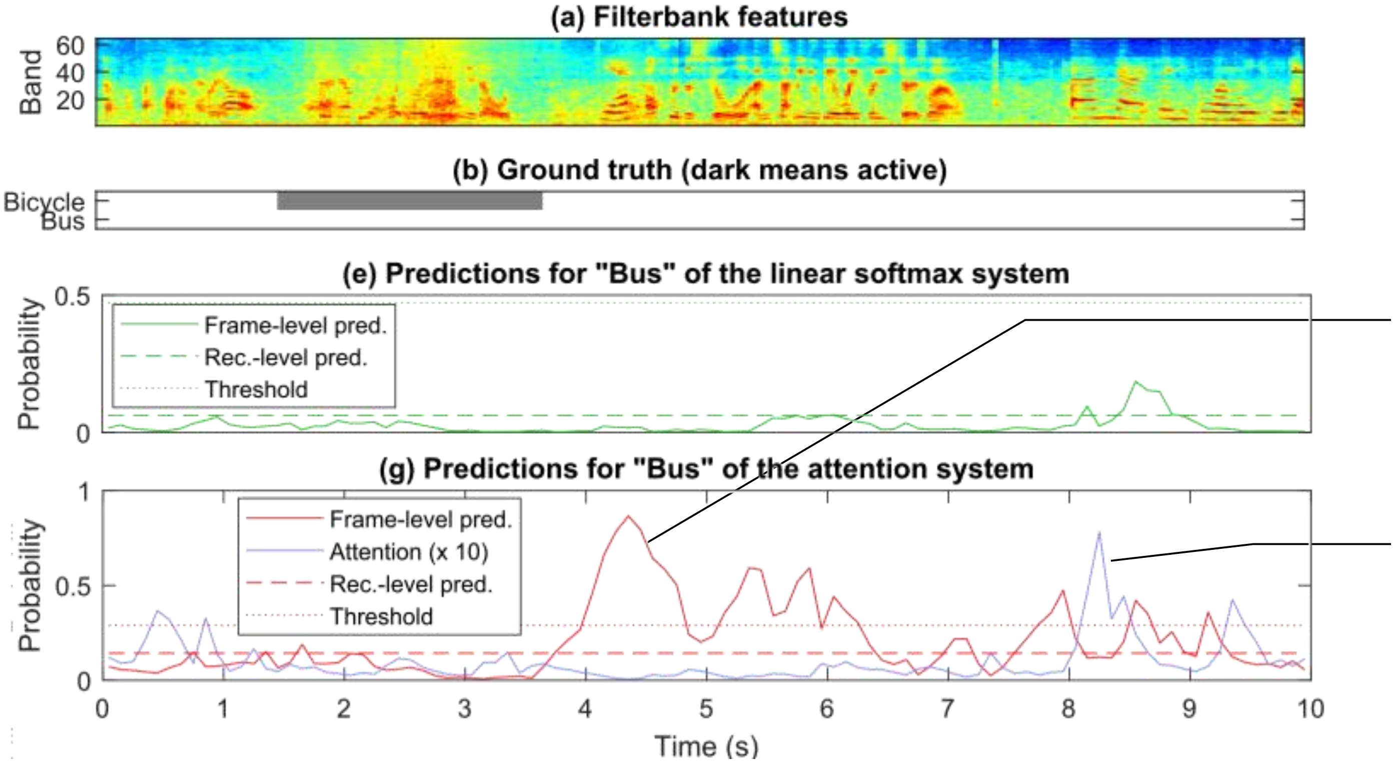
$$y = \frac{\sum_i y_i w_i}{\sum_i w_i} \qquad \frac{\partial y}{\partial y_i} = \frac{w_i}{\sum_j w_j} \qquad \frac{\partial y}{\partial w_i} = \frac{y_i - y}{\sum_j w_j}$$

Always positive

Positive when ***y***

- When bag is positive:
  - All y_i increase 🙂 attention focuses where $y_i > y$ 🙂
- When bag is negative:
  - All y_i decrease 🙂 attention focuses where $y_i < y$ 😱
  - Smaller probs get larger weight!

Failure Mode of Attention



False positives in unattended regions

Attention focuses here

- Too many frame-level false positives
- Inconsistent recording-level and frame-level predictions

EVALUATION I:

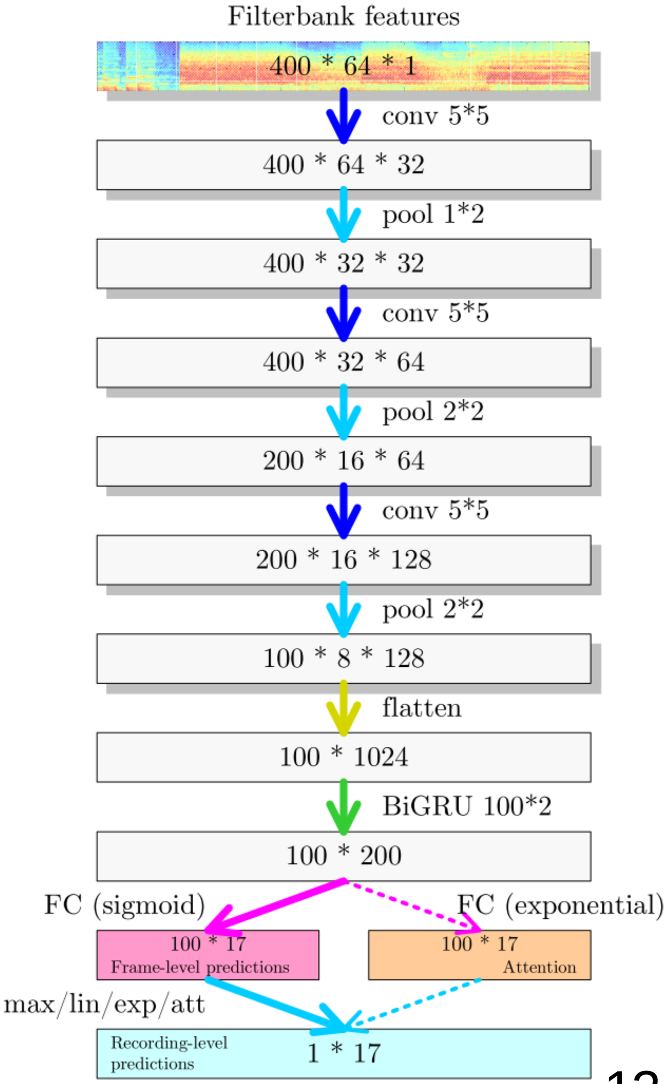
DCASE 2017 Challenge, Task 4

DCASE 2017: Task

- **17 event types**
    - Vehicles, warnings
- **Training data:**
    - ~50k recordings * 10 seconds each = ~140 hours
    - Weakly labeled
- **Test data:**
    - 488 recordings * 10 seconds each = ~1.4 h
    - Strongly labeled
- **Evaluation metrics:**
    - Tagging: F1
    - Localization: error rate & F1 on 1s segments

DCASE 2017: Model

- Input:
  - Logmel features @ 40 Hz
- Structure:
  - 3 conv layers + 1 GRU layer
- Output:
  - Frame-level event probs at 10 Hz
  - For tagging: pooled globally into recording-level event probs
  - For localization: pooled over 1s segments



DCASE 2017: Results

Pooling Func	Tag F1	Loc ER	Loc F1	Loc #FN	Loc #FP
Max	45.3	84.7	35.4	3,154	1,253
Linear softmax	49.5	84.3	43.7	2,528	2,187
Attention	49.2	102.5	40.1	2,434	3,309

- Max: too many false negatives (FNs) hurt F1
- Attention: too many false positives (FPs) hurt ER
- Linear softmax: balanced FNs and FPs

EVALUATION II:

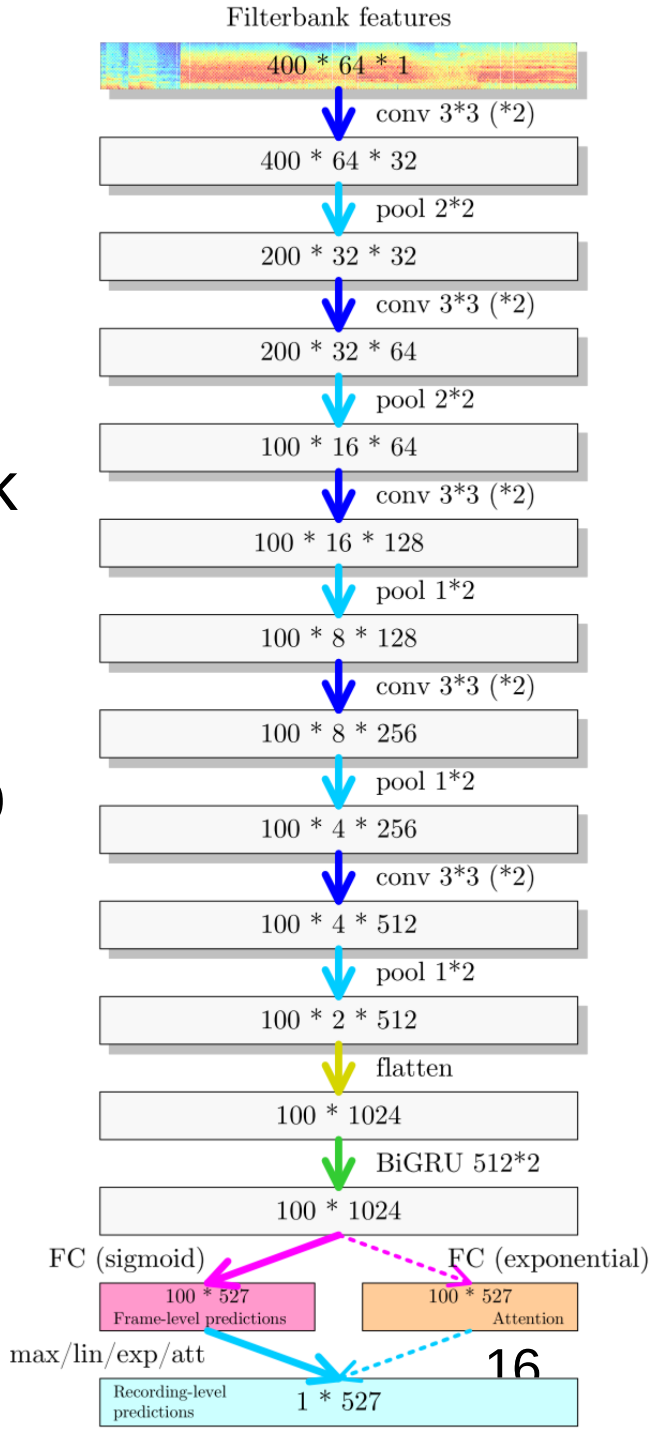
Google Audio Set

Audio Set: Task

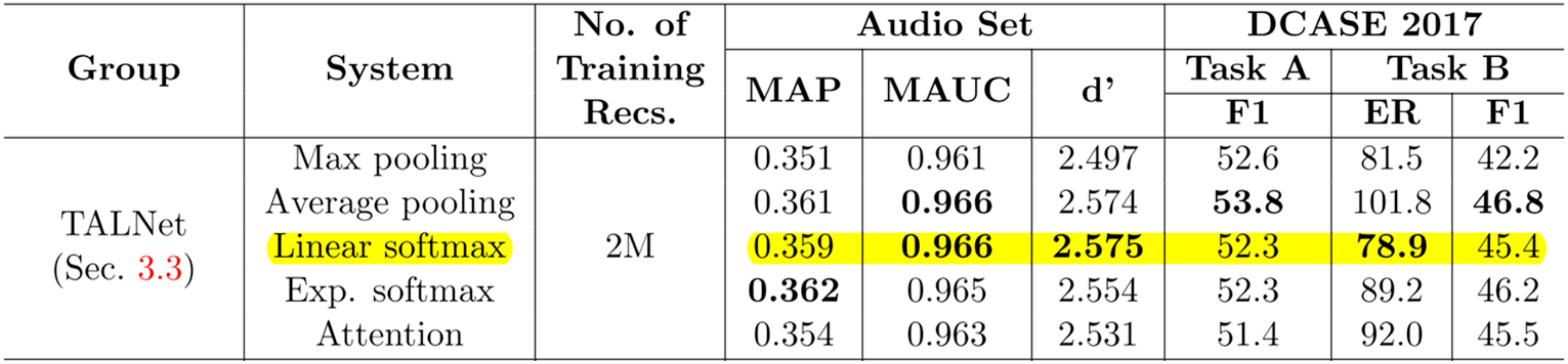
- Data:
  - 527 event types (include the 17 events of DCASE)
  - Weakly labeled
  - Training: ~2M recordings * 10s = 8 months
  - Test: ~20k recordings * 10s = 56 hours
- Evaluation metrics:
  - Audio Set only measures tagging
    - MAP, MAUC, d'
  - Reuse DCASE data & metrics for tagging & localization
    - Tag F1, Loc ER, Loc F1 over 1s segments

Audio Set: Model

- TALNet:
  - Tagging and Localization Network
  - 10 conv layers, 1 GRU layer
  - Same input & output as before
- No fine-tuning when applied to DCASE data



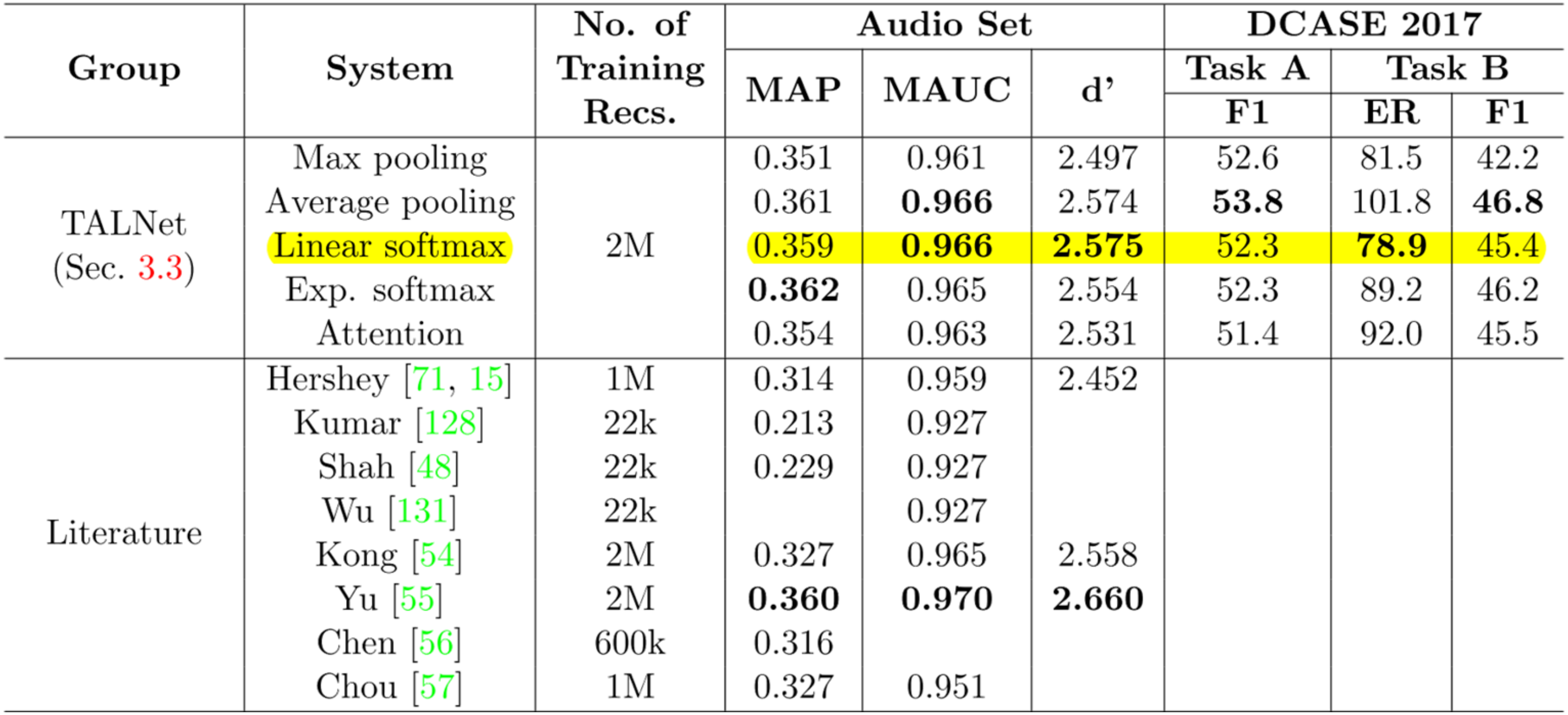
Audio Set: Result 1/3



Group	System	No. of Training Recs.	Audio Set			DCASE 2017		
			MAP	MAUC	d'	Task A	Task B	
						F1	ER	F1
TALNet (Sec. 3.3)	Max pooling	2M	0.351	0.961	2.497	52.6	81.5	42.2
	Average pooling		0.361	**0.966**	2.574	**53.8**	101.8	**46.8**
	Linear softmax		0.359	**0.966**	**2.575**	52.3	**78.9**	45.4
	Exp. softmax		**0.362**	0.965	2.554	52.3	89.2	46.2
	Attention		0.354	0.963	2.531	51.4	92.0	45.5

- TALNet works out of the box on DCASE
- Linear softmax is best for localization
  - And good enough for tagging

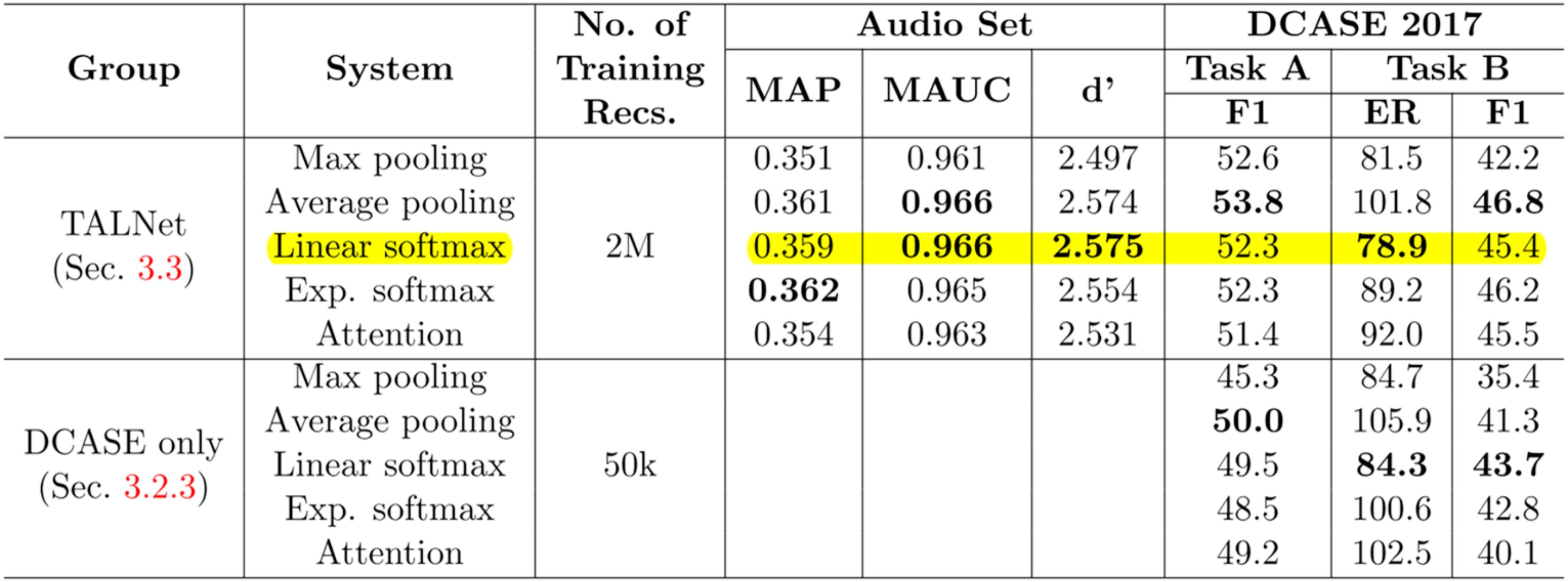
Audio Set: Result 2/3



Group	System	No. of Training Recs.	Audio Set			DCASE 2017		
			MAP	MAUC	d’	Task A	Task B	
						F1	ER	F1
TALNet (Sec. 3.3)	Max pooling	2M	0.351	0.961	2.497	52.6	81.5	42.2
	Average pooling		0.361	**0.966**	2.574	**53.8**	101.8	**46.8**
	Linear softmax		0.359	**0.966**	**2.575**	52.3	**78.9**	45.4
	Exp. softmax		**0.362**	0.965	2.554	52.3	89.2	46.2
	Attention		0.354	0.963	2.531	51.4	92.0	45.5
Literature	Hershey [71, 15]	1M	0.314	0.959	2.452			
	Kumar [128]	22k	0.213	0.927				
	Shah [48]	22k	0.229	0.927				
	Wu [131]	22k		0.927				
	Kong [54]	2M	0.327	0.965	2.558			
	Yu [55]	2M	**0.360**	**0.970**	**2.660**			
	Chen [56]	600k	0.316					
	Chou [57]	1M	0.327	0.951				

- TALNet closely matches state of the art on tagging
  - Yu’s system uses multi-level attention and can’t do localization!
- Amount of training data matters!

Audio Set: Result 3/3



Group	System	No. of Training Recs.	Audio Set			DCASE 2017		
			MAP	MAUC	d’	Task A	Task B	
						F1	ER	F1
TALNet (Sec. 3.3)	Max pooling	2M	0.351	0.961	2.497	52.6	81.5	42.2
	Average pooling		0.361	**0.966**	2.574	**53.8**	101.8	**46.8**
	Linear softmax		0.359	**0.966**	**2.575**	52.3	**78.9**	45.4
	Exp. softmax		**0.362**	0.965	2.554	52.3	89.2	46.2
	Attention		0.354	0.963	2.531	51.4	92.0	45.5
DCASE only (Sec. 3.2.3)	Max pooling	50k				45.3	84.7	35.4
	Average pooling					**50.0**	105.9	41.3
	Linear softmax					49.5	**84.3**	**43.7**
	Exp. softmax					48.5	100.6	42.8
	Attention					49.2	102.5	40.1

- Adding more data helps the 17 DCASE events
  - Even though most of it belongs to 510 other events

Summary

- Linear softmax is the best for localization
  - Better than max: unobstructed gradient flow
  - Better than attention:
    - Balanced false negatives and false positives
    - Consistent frame-level & recording-level predictions
- We built TALNet
  - First simultaneous audio tagging and localization
  - Closely matches state of the art on Audio Set
  - Good performance on DCASE 2017 out of the box
- Future work
  - Attention pooling with monotonicity constraint?

Thanks!

Questions?