Javier Naranjo-Alcazar[1 2], Sergi Perez-Castanos[2], Aaron Lopez-Garcia[2], Pedro Zuccarello[1], Maximo Cobos[2], Francesc J. Ferri[2]

1. Instituto Tecnológico de Informática, Valencia, Spain
2. Universitat de València, Burjassot, Spain

SQUEEZE-EXCITATION CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR AUDIO-VISUAL SCENE CLASSIFICATION

ACOUSTIC SCENE CLASSIFICATION BACKGROUND

Specific task of Machine Listening field

Tag an audio clip into a pre-defined scene

Proposed in the first DCASE Challenge edition (2013)

Different approaches have been addressed

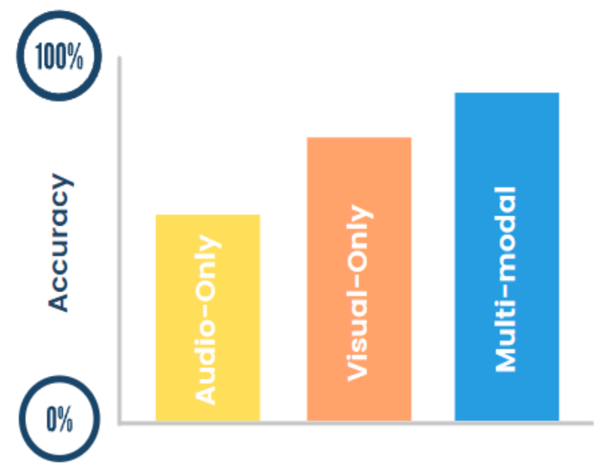
Audio representations, ensembles, data augmentations

MAIN OBJECTIVES

Improve framework accuracy using visual data

Without constraint (number of parameters)

Complexity - accuracy



DATASET

- TAU Urban AudioVisual Scenes 2021
- 10 scenes from 12 European cities
- 10 second audios -> 34 hours of audio data
- Official partition -> 70-30

TRAINING PROCEDURE

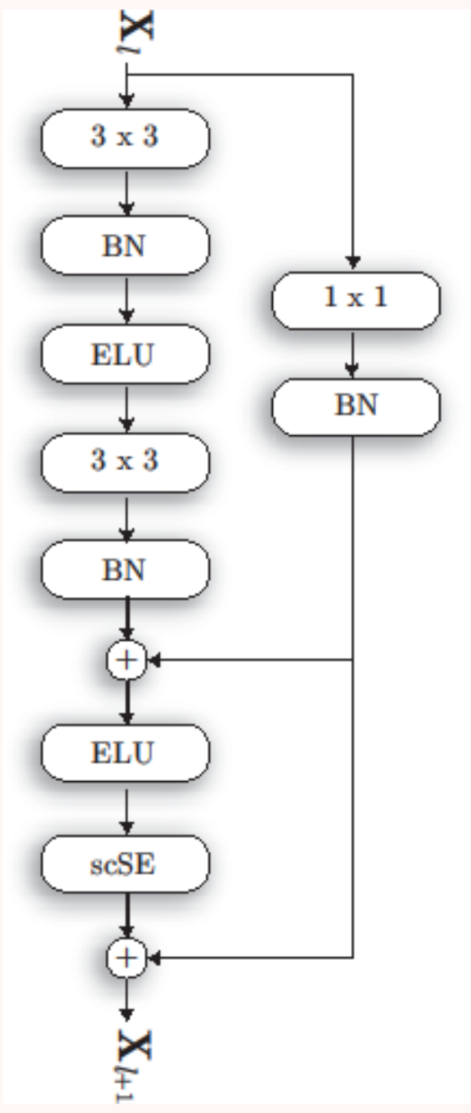
1. Train audio network
2. Train recurrent layer of visual network
3. Train fusion layers of full framework -> final fine-tuning

- 200 epochs, 32 batch size and 16 for full, audio mixup, 1 second

SYSTEM COMPLEXITY

Module	Parameters
Audio	323k
Visual	14M (105k trainable)
Full	15M (272k trainable)

AUDIO MODULE



Input

Following previous submissions: 3-channel audio representation

Mel and Gammatone filterbanks

Audio resampled to 44.1 kHz -> (64, 50, 3)

Network

Fully convolutional -> Conv-StandardPOST block

Max Pooling and Dropouts after each block

Global Average Pooling

VISUAL MODULE

Input

224 x 224 images to match VGG16 input

5 frames per second -> (5, 224, 224, 3)

Network

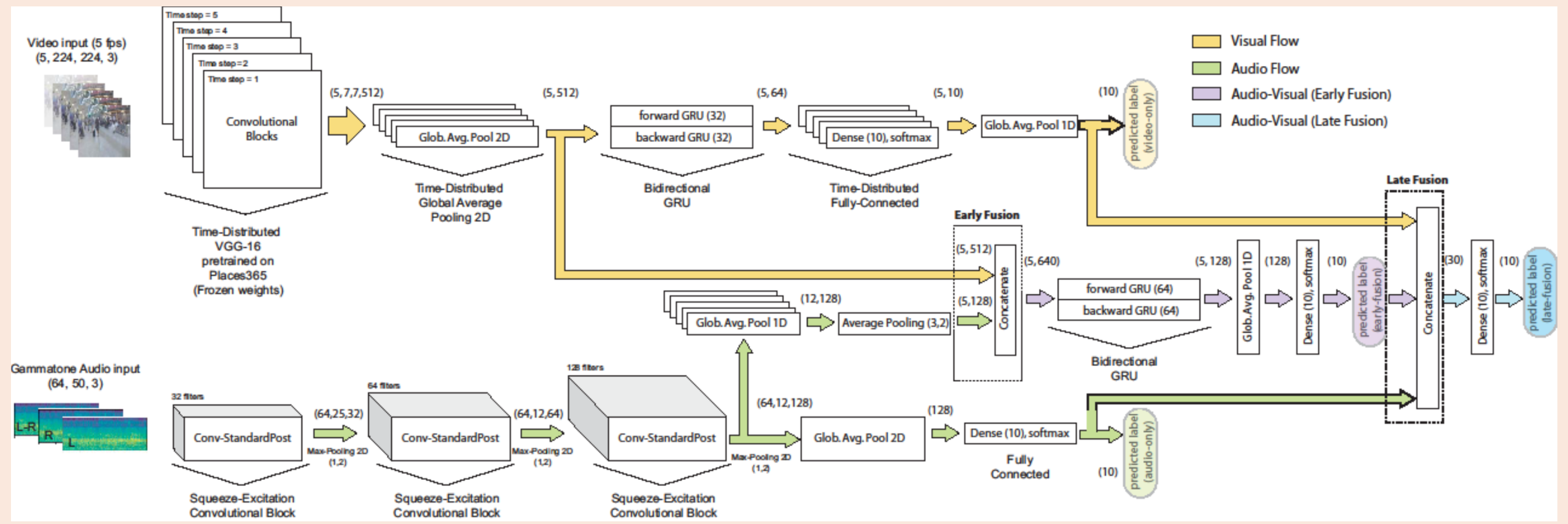
VGG16 pre-trained with places365 -> TimeDistributed

VGG16 as feature extractor -> frozen weights

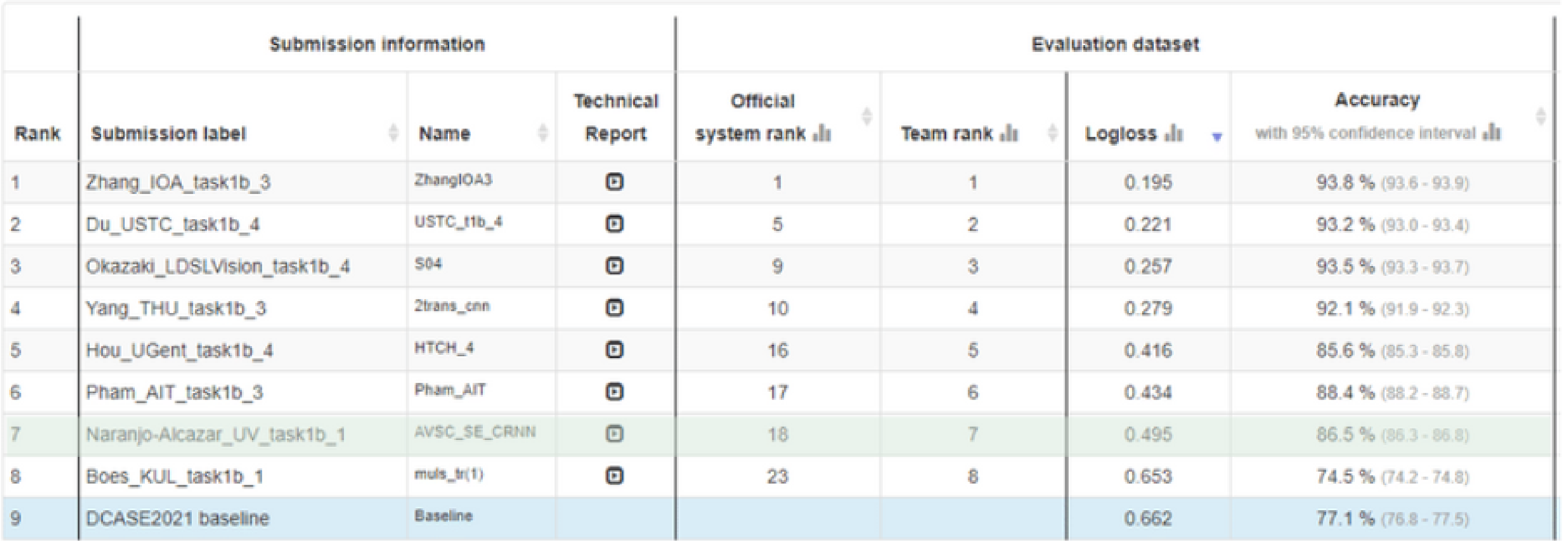
Global Average Pooling -> (5, 512)

Trainable GRU layers and final Dense layers

FULL AUDIO-VISUAL FRAMEWORK



CHALLENGE COMPARISON



Submission information				Evaluation dataset			
Rank	Submission label	Name	Technical Report	Official system rank	Team rank	Logloss	Accuracy with 95% confidence interval
1	Zhang_IOA_task1b_3	ZhangIOA3	▣	1	1	0.195	93.8 % (93.6 - 93.9)
2	Du_USTC_task1b_4	USTC_t1b_4	▣	5	2	0.221	93.2 % (93.0 - 93.4)
3	Okazaki_LDSLVision_task1b_4	S04	▣	9	3	0.257	93.5 % (93.3 - 93.7)
4	Yang_THU_task1b_3	2trans_cnn	▣	10	4	0.279	92.1 % (91.9 - 92.3)
5	Hou_UGent_task1b_4	HTCH_4	▣	16	5	0.416	85.6 % (85.3 - 85.8)
6	Pham_AIT_task1b_3	Pham_AIT	▣	17	6	0.434	88.4 % (88.2 - 88.7)
7	Naranjo-Alcazar_UV_task1b_1	AVSC_SE_CRNN	▣	18	7	0.495	86.5 % (86.3 - 86.8)
8	Boes_KUL_task1b_1	muls_tr(1)	▣	23	8	0.653	74.5 % (74.2 - 74.8)
9	DCASE2021 baseline	Baseline				0.662	77.1 % (76.8 - 77.5)

EXPERIMENTS

	Modality			
	Audio-Only	Visual-Only	Multi-Modal (Early Fusion)	Multi-Modal (Late Fusion)
log-Mel	68.4	87.0	88.5	88.7
Gammatone	69.0	87.0	89.2	90.0

Table 1: Accuracy Results on the TAU Audio-Visual Urban Scenes 2021 validation partition

Parameters		
Audio-Only (Gammatone)	Visual-Only	Multi-Modal (Late Fusion)
66.8	83.2	86.5

Table 2: Accuracy Results on the TAU Audio-Visual Urban Scenes 2021 challenge partition

CONCLUSION

System outperforms baseline accuracy with few parameters compared to other participants

Both models are trained in isolation

The results show the combination of two domains improves system accuracy

Future work -> slim models for real time inference