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

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, València, Spain {jnaranjo, pzuccarello}@iti.es
[2] Universitat de València, Burjassot, Spain, {pecaser, aaron.lopez, maximo.cobos, francesc.ferri}@uv.es

ABSTRACT

The use of multiple and semantically correlated sources can provide complementary information to each other that may not be evident when working with individual modalities on their own. In this context, multi-modal models can help to produce more accurate and robust predictions in machine learning tasks where audio-visual data is available. This paper presents a multi-modal model for automatic scene classification that exploits simultaneously auditory and visual information. The proposed approach makes use of two separate networks which are respectively trained in isolation on audio and visual data, so that each network specializes in a given modality. The visual subnetwork is a pre-trained VGG16 model followed by a bidiretional recurrent layer, whilst the residual audio subnetwork is based on stacked squeeze-excitation convolutional blocks trained from scratch. After training each subnetwork, the fusion of information from the audio and visual streams is performed at two different stages. The early fusion stage combines features resulting from the last convolutional block of the respective subnetworks at different time steps to feed a bidirectional recurrent structure. The late fusion stage combines the output of the early fusion stage with the independent predictions provided by the two subnetworks, resulting in the final prediction. We evaluate the method using the recently published TAU Audio-Visual Urban Scenes 2021, which contains synchronized audio and video recordings from 12 European cities in 10 different scene classes. The proposed model has been shown to provide an excellent trade-off between prediction performance (86.5%) and system complexity (15M parameters) in the evaluation results of the DCASE 2021 Challenge.

Index Terms— Deep Learning, Multi-modal, Convolutional Neural Networks, Scene Classification, Squeeze-Excitation, Gammatone, DCASE 2021

1. INTRODUCTION

The world as it is perceived by humans involves multiple modalities. In general, a sensory modality is understood as a primary channel of communication and sensation, such as vision, hearing or touch. In this context, multi-modal machine learning aims at exploiting datasets including multiple such modalities, building models that can process and relate information among them [1]. The explosion of deep learning and its use in vision, natural language processing and acoustic analysis, makes of multi-modal machine learning a multi-disciplinary area with increasing potential. Obviously, the most abundant multi-modal datasets are those made up of audio-visual data, where the included examples come in the form of videos that include both sound and images.

One important application scenario of multi-modal machine learning is automatic scene classification on videos, which refers to the task of classifying audiovisual data to one of the predefined scene categories (such as airport, park or shopping mall) [2], based on the ambient content provided by the information contributed by both modalities. Note, however, that scene classification has also been a topic of intensive research in the last decade by considering different modalitites on their own [3, 4, 5].

This paper presents a multi-modal approach for scene classification consisting of two specialized components or modules (an audio module and a visual module) that are further trained together to achieve a more robust solution by incorporating two fusion strategies simultaneously. The visual module is based on a VGG16 [6] convolutional neural network (CNN) pre-trained on the *places365* dataset [7, 8]. The training procedure of this component is based on a transfer learning scheme with fine tuning. On the other hand, the audio module is based on a fully convolutional neural network with convolutional blocks implementing residual and squeeze-excitation techniques [9] and gammatone filterbank audio representations as input. Finally, the audio and video modules with frozen weights are combined into a multimodal recurrent structure that performs information fusion both at early and late stages. The early fusion stage combines features resulting from the last convolutional block of the audio and visual subnetworks at different time steps, while the late fusion stage provides a final prediction by combining the output of the early fusion stage with the independent predictions provided by the visual and audio modules.

The model is evaluated by considering the TAU Audio-Visual Urban Scenes 2021 dataset [10], which contains synchronized audio and video recordings from 12 European cities in 10 different scenes classes. For a complete assessment of the model, different input pre-processing alternatives and architecture choices are considered and discussed.

The rest of this paper is structured as follows. Section 2 describes in detail the architecture of the different sections making up the whole learning system. Section 3 describes the experimental set-up and evaluates the proposed system considering some variants with respect to the input and the system architecture. Section 4 compares our system with other Challenge submissions. Finally, Section 5 concludes this work.

2. SYSTEM DESCRIPTION

This section describes the full architecture of the system, providing details on the different modules making up the whole multi-modal network. An schematic view of the full model is depicted in Fig. 1, where the visual and audio flows are represented with different col-

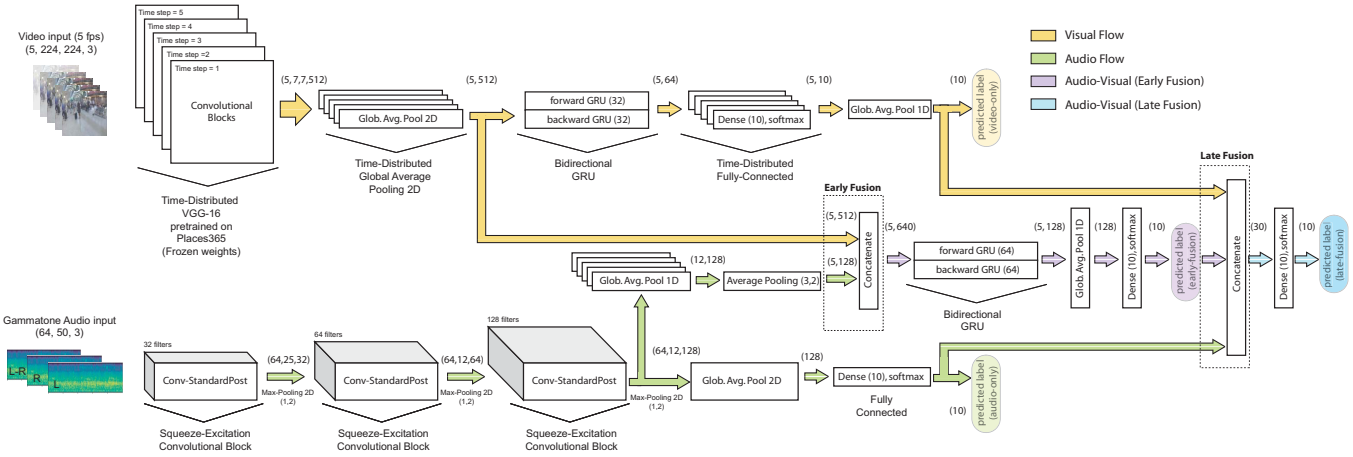

Figure 1: Proposed network achitecture for audiovisual scene classification.

ors.

2.1. Audio Module

2.1.1. Audio Input Representation

Based on previous works by the authors [11, 12], the input to the network consists is a multi-channel 3D audio representation compiling information from the left and right input audio channels as well as their difference. Each channel is converted into a time-frequency domain representation, provided by a Gammatone or a Mel-scale filter bank. Both alternatives have been widely adopted by the machine listening community [13, 14, 15, 16] and we evaluate both options in the experimental section of this paper.

All the considered representations are computed using 64 frequency bands, with a window size of 40 ms and 50% overlap. The audio was resampled from 48 kHz to 44.1 kHz. Gammatone representations were computed by using the Auditory Toolbox presented in [17] with Python implementation and Mel-spectrograms were obtained by using the LibRosa library [18]. Taking the above details into account, one second of audio results in a tensor input of size (64, 50, 3), where the third axis corresponds to the left-right-difference channels.

2.1.2. Audio Subnetwork

The audio module is based on a fully convolutional neural network combining residual connections with squeeze-excitation. More specifically, the convolutional blocks follow the structure of those denoted as *Conv-StandardPOST* in [9] (see Fig. 2), which showed very good performance for acoustic scene classification tasks. The aim of these blocks is to achieve improved accuracy by recalibrating the internal feature maps using residual [19] and squeeze-excitation techniques [20, 21]. An important feature is the use of the scSE (spatial and channel Squeeze and Excitation) module, which performs a spatial and channel-wise recalibration of the block feature maps. The interested reader is referred to [9] for a full description and evaluation of such blocks. All use a 3×3 kernel size, while the number of filters in each block are specified in Fig. 1. In between convolutional blocks, Max Pooling layers are used to halve the resolution of the resulting feature maps along the time axis. Additionally, Dropout [22] with a rate of 0.3 is also included after

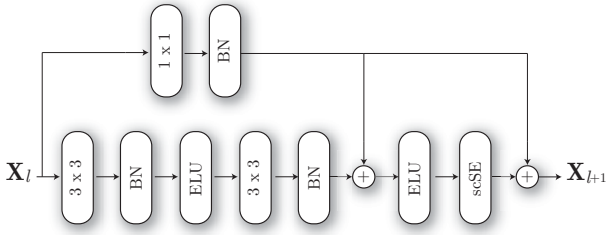

Figure 2: Structure of the *Conv-StandardPOST* block [9]. BN and ELU denote batch normalization and exponential linear unit activation, respectively. The $N \times N$ notation denotes a convolutional layer with the corresponding kernel size. The input to the l-th block is denoted as $\mathbf{X}_l$.

the pooling layers to prevent overfitting. The output feature maps from the last convolutional block are summarized with global average pooling into a 128-dimensional feature vector, which is fed to a fully-connected layer with softmax activation for classification. This subnetwork is trained from scratch using only audio data on the whole dataset by minimizing the cross-entropy loss.

2.2. Visual Module

2.2.1. Visual Input Representation

The visual input is adapted to match the pre-trained VGG16 architecture [6], which accepts color images of size 224×224 pixels. Moreover, as visual scene recognition does not require a very high frame rate (images do not change that much from frame to frame), the videos from the dataset are subsampled for obtaining a frame rate of 5 frames per second (fps). Therefore, a one-second video clip results in a tensor shape of (5, 224, 224, 3).

2.2.2. Visual Subnetwork

The visual module is based on the VGG16 CNN architecture [6] pretrained on the *places365* dataset [7]. With the aim of processing temporal information extracted from multiple frames, a time-distributed structure with frozen weights is considered. It must be

emphasized that the pre-trained model is used as a feature extractor, so that the top fully-connected layers of the network are omitted. The outputs from each time step (5 temporal steps) are globally averaged channel-wise, resulting in a sequence of 512 output features. This sequence is fed to a bidirectional 64-neuron Gated Recurrent Unit (GRU) layer and the returned sequences are processed by a time-distributed fully-connected layer with softmax activation, resulting in a predicted label for each time step. The final label is taken as the temporal average of the predictions. This subnetwork is trained on the visual data only, with trainable weights only on the recurrent and final dense layer.

2.3. Full Audio-Visual Network

The complete audio and visual modules described above are then merged into a full audio-visual framework that combines information from both modalities at two different levels. On an early fusion stage, the output of the last convolutional block of the audio and visual modules are concatenated into a sequence of 640 features. To achieve this, the feature maps of the audio module are turned into a temporal sequence matching the temporal resolution of the visual data (i.e. 5 fps) using global and average pooling operators. A bidirectional GRU processes the sequence and a new prediction is created by stacking a global average pooling and a dense layer. A late fusion stage receives the predictions from the independent modalities as well as the one resulting from their combination and produces the final prediction with a dense layer with softmax activation.

Note that, as observed in Fig. 1, the full network can be used to extract both predictions from the independent modalitites, i.e. labels from the visual (yellow) and audio (green) information flows, and from the fusion flows, i.e. early fusion (purple) and late fusion (blue).

2.4. Dataset

The system is trained on the recently published TAU Urban Audio-Visual Scenes 2021 [10]. This dataset contains fragments of recordings obtained in 12 large European cities corresponding to 10 scene classes: airport, shopping mall (indoor), metro station (underground), pedestrian street, public square, street (traffic), traveling by tram, bus and metro (underground), and urban park. The data was gathered with four devices recording simultaneously. The data examples are provided as segments with a length of 10 seconds, annotated by the corresponding scene class, city and recording location identifier. The dataset contains 34 hours of recordings and it specifies training/test partitions to facilitate comparisons. The training set contains approximately 70% of the data, while the validation set contains the remaining 30%.

2.5. Training Details

The whole network was trained in three steps using the default training and validation partitions provided by the TAU Audio-Visual Urban Scenes 2021 dataset. The first step corresponds to the training of the audio module from scratch using audio data only. The second step trains the recurrent and classification parts of the visual module (the convolutional blocks use frozen weights from the pre-trained network). In the last step, the whole audio-visual network is trained using frozen weights from the audio and visual modules. A fine-tuning strategy is finally followed, unfreezing all the weights and using a very small learning rate. The loss function used at each training step was categorical cross-entropy. The optimizer used was Adam [23] with default parameters. The models were trained with a maximum of 200 epochs. Batch size was set to 32 for training the independent subnetworks and 16 for the complete audio-visual network due to memory constraints. The 10 second examples provided in the dataset were randomly trimmed into 1 second segments in each epoch. The learning rate started with a maximum value of 0.001 decreasing with a factor of 0.5 in case of no improvement in validation accuracy after 20 epochs. In the last fine-tuning with all trainable weights, the starting learning rate was 10^{-5}. The training is considered as early finished in case of no improvement in validation accuracy after 50 epochs. Mixup data-augmentation [24] with $\alpha = 0.4$ was used. All the models were implemented using Keras with Tensorflow backend and trained using NVidia Titan RTX GPU.

2.6. Model Complexity

Assuming that all the weights of the different subsystems are trainable, the number of parameters corresponding to the different modules of the network are as follows: audio module (323k trainable weights), visual module (14M parameters, only 105k trainable) and full audio-visual system (15M parameters, only 272k trainable). Note that, although the number of parameters used by the visual module is considerably higher than that of the audio module, the visual one uses frozen weights in all its convolutional blocks. Thus, all subsystems can be trained considerably fast, as only 272k weights are trainable. Additionally, the final fine-tuning step in which all the weights are unfrozen only requires 2 epochs and, therefore, it does not require too much extra training time.

3. EXPERIMENTS

This section evaluates the proposed multi-modal framework over the default validation partition of the TAU Audio-Visual Urban Scenes 2021 dataset. For those systems submitted to the DCASE 2021 Challenge, we also provide the performance reported by the organizers over the evaluation/test dataset. Note that no custom test partition was created in order to facilitate comparisons with other competing systems using the same dataset. The performance of the proposed system is analyzed considering three different aspects:

- Audio input representation: log-Mel spectrogram and gammatone filterbank.
- Independent modalities: audio-only and visual-only.
- Multi-modal fusion: early fusion and late fusion.

Table 1 shows the accuracy results obtained for the different subsystems involved in the audio-visual framework on the default validation set. Similarly, Table 2 shows the results obtained for the evaluation partition used in DCASE 2021 Task1b. Comparisons to other competing approaches can be directly accessed via the DCASE 2021 Challenge website[1].

In general, the results clearly highlight that the visual-only modality is much more accurate than the audio-only one (e.g. 87.0% vs 69.0% in the development validation partition). Although it is true that the visual network departs from previous knowledge provided by a pre-trained model, this confirms that, as of today, acoustic scene recognition is a more challenging problem than visual scene recognition. Nonetheless, the audio module used in this

[1] http://dcase.community/challenge2021/task-acoustic-scene-classification-results-b

	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.

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

Table 2: Accuracy results on the DCASE 2021 Task1b evaluation set.

work was ranked as the best performing model from all the submissions considering only the audio modality in terms of log-loss performance. In addition, it is observed that although both audio input representations, log-Mel and Gammatone spectrograms, led to very similar performances, the results were slightly better with the use of Gammatone filterbanks (69.0% vs 68.4% in the validation partition).

Despite the fact that the multi-modal models provide the best performance, their accuracy is only slightly better than the best of the individual modalities, which is particularly the visual one. In any case, for our specific case and despite the significant performance gap between the audio and video modalities, the multi-modal approaches achieved a performance gain of approximately 3 percentage points. In this context, although both the early and late fusion stages were able to exploit the information from the audio and visual data, the late fusion stage performed consistently (slightly) better in all our experiments.

In order to provide further insight, Fig. 3 shows the performance achieved by each modality and their combination in the DCASE 2021 Task1b evaluation set on each class. Note that the audio module only outperforms the visual one for the tram class and it can be clearly observed for this case that multi-modality allows to exploit significantly both types of information. Interestingly, although the multi-modal system usually outperforms the individual modalities alone, there are some classes at which this does not happen, as in public square or metro.

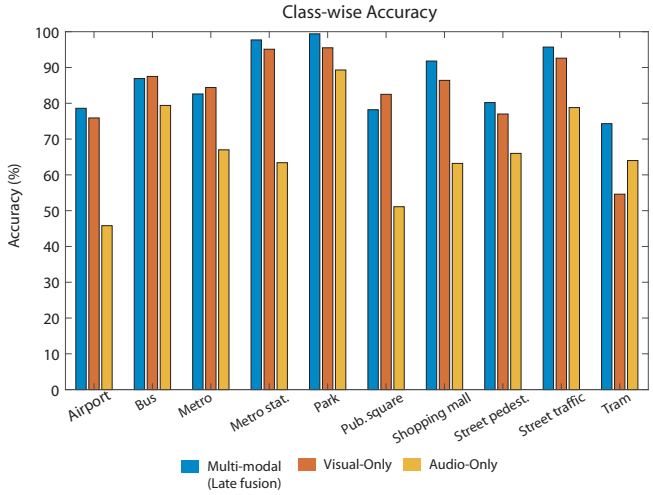

Figure 3: Class-wise performance in DCASE 2021 Task1b evaluation set.

4. CHALLENGE COMPARISON

The presented multi-modal framework ranked 7th in Task1b of the DCASE 2021 Challenge. It is important to remark that only 7 teams exceeded 85% accuracy. The team ranking 8th achieved an accuracy of 74% [25], 12 percentage points lower than our system. Moreover, our model presents a great performance-complexity balance. For example, the system ranking 6th [26] (with 88.4% accuracy) has 140M parameters, while ours has only 15M. All systems above 90% accuracy have more than 40M parameters, with the most complex one having 1B parameters [27]. Thus, the model presented in this paper allows very good accuracy with moderate complexity.

5. CONCLUSION

This paper presented a multi-modal system for audio-visual scene classification based on convolutional recurrent neural networks. The full system is based on two individual modules that are trained in isolation on the audio and visual modalities, respectively. The visual module is based on a time-distributed VGG16 model pre-trained on the places365 dataset, followed by a bidirectional recurrent layer. The audio module is a convolutional neural network that incorportes residual and squeeze-excitation techniques, working over Gammatone input representations. After training both modules, both are incorporated into a full audio-visual architecture that performs information fusion at early and late stages. Early fusion combines features extracted from both modalities at each time step into a bidirectional recurrent layer. Late fusion decides a final label after receiving the predictions obtained from each independent modality and that resulting from early fusion. The results show that the proposed framework is able to exploit successfully information from both modalities, even though the visual modality is considerably more accurate than the audio one. The results obtained in the DCASE 2021 Challenge confirm that the proposed system provides an excellent trade-off between prediction performance and system complexity.

6. ACKNOWLEDGEMENTS

This work is partially supported by ERDF and the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-097045-B-C21, as well as grants AICO/2020/154 and AEST/2020/012 from Generalitat Valenciana.

7. REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] D. Zeng, M. Liao, M. Tavakolian, Y. Guo, B. Zhou, D. Hu, M. Pietikäinen, and L. Liu, "Deep Learning for Scene Classification: A Survey," *arXiv preprint arXiv:2101.10531*, 2021.

[3] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[4] J. Abeßer, "A review of Deep Learning based methods for Acoustic Scene Classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[5] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[6] K. Simonyan and A. Zisserman, "Very deep Convolutional Networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[8] G. Kalliatakis, "Keras-vgg16-places365," https://github.com/GKalliatakis/Keras-VGG16-places365, 2017.

[9] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic Scene Classification with squeeze-excitation residual networks," *IEEE Access*, vol. 8, pp. 112 287–112 296, 2020.

[10] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, accepted. [Online]. Available: https://arxiv.org/abs/2011.00030

[11] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, M. Cobos, and F. J. Ferri, "CNN depth analysis with different channel inputs for Acoustic Scene Classification," *arXiv preprint arXiv:1906.04591*, 2019.

[12] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Task 1 DCASE 2020: ASC with mismatch devices and reduced size model using residual squeeze-excitation CNNs," DCASE2020 Challenge, Tech. Rep, Tech. Rep., 2020.

[13] S. Tabibi, A. Kegel, W. K. Lai, and N. Dillier, "Investigating the use of a Gammatone filterbank for a cochlear implant coding strategy," *Journal of Neuroscience Methods*, vol. 277, pp. 63–74, 2017.

[14] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep Convolutional Neural Network with mixup for environmental sound classification," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[15] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous Sound Detection using Unsupervised and Semi-Supervised Autoencoders and Gammatone Audio Representation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 145–149.

[16] ——, "Listen Carefully and Tell: An Audio Captioning System Based on Residual Learning and Gammatone Audio Representation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 150–154.

[17] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, 1998.

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference*, vol. 8, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00745

[21] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[25] W. Boes and H. Van hamme, "Multi-source transformer architectures for audiovisual scene classification," DCASE2021 Challenge, Tech. Rep., June 2021.

[26] L. Pham, A. Schindler, M. Schutz, J. Lampert, and R. King, "DCASE 2021 task 1B: Technique report," DCASE2021 Challenge, Tech. Rep., June 2021.

[27] Q. Wang, S. Zheng, Y. Li, Y. Wang, Y. Wu, H. Hu, C.-H. H. Yang, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "A model ensemble approach for audio-visual scene classification," DCASE2021 Challenge, Tech. Rep., June 2021.