This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI

10.1109/ACCESS.2020.3002761, IEEE Access

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

**Acoustic Scene Classification with Squeeze-Excitation Residual Networks**

**JAVIER NARANJO-ALCAZAR1,2, (Student Member, IEEE), SERGI PEREZ-CASTANOS1, PEDRO ZUCCARELLO 1, AND MAXIMO COBOS.2, (Senior Member, IEEE)**

1 Visualfy, 46181 Benisanó, Spain (e-mail: javier.naranjo, sergi.perez, pedro.zuccarello@visualfy.com)

* Computer Science Department, Universitat de Valencia, 46100 Burjassot, Spain (e-mail: maximo.cobos@uv.es)

Corresponding author: Javier Naranjo-Alcazar (e-mail: javier.naranjo@visualfy.com).

“This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 779158. The participation of Javier Naranjo-Alcazar and Dr. Pedro Zuccarello in this work is partially supported by Torres Quevedo fellowships DIN2018-009982 and PTQ-17-09106 respectively from the Spanish Ministry of Science, Innovation and Universities. The participation of Dr. Cobos is supported by FEDER and the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-097045-B-C21.”

 **ABSTRACT** Acoustic scene classification (ASC) is a problem related to the field of machine listeningwhose objective is to classify/tag an audio clip in a predefined label describing a scene location (e. g. park, airport, etc.). Many state-of-the-art solutions to ASC incorporate data augmentation techniques and model ensembles. However, considerable improvements can also be achieved only by modifying the architecture of convolutional neural networks (CNNs). In this work we propose two novel squeeze-excitation blocks to improve the accuracy of a CNN-based ASC framework based on residual learning. The main idea of squeeze-excitation blocks is to learn spatial and channel-wise feature maps independently instead of jointly as standard CNNs do. This is usually achieved by combining some global grouping operators, linear operators and a final calibration between the input of the block and its learned relationships. The behavior of the block that implements such operators and, therefore, the entire neural network, can be modified depending on the input to the block, the established residual configurations and the selected non-linear activations. The analysis has been carried out using the TAU Urban Acoustic Scenes 2019 dataset presented in the 2019 edition of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge. All configurations discussed in this document exceed the performance of the baseline proposed by the DCASE organization by 13% percentage points. In turn, the novel configurations proposed in this paper outperform the residual configurations proposed in previous works.

 **INDEX TERMS** Acoustic Scene Classification, Deep Learning, Machine Listening, Pattern Recognition,Squeeze-Excitation

**I. INTRODUCTION**

The analysis of everyday ambient sounds can be very useful when developing intelligent systems in applications such as domestic assistants, surveillance systems or autonomous driving. Acoustic scene classification (ASC) is one of the most typical problems related to machine listening [1]–[4]. Machine listening is understood as the field of artificial intel-ligence that attempts to create intelligent algorithms capable of extracting meaningful information from audio data. There-fore, ASC can be defined as the area of machine listening that attempts to tag an audio clip in one of the predefined tags related to the description of a scene (for example, airport, park, subway, etc.).

The first approaches to the ASC problem were centered on the design of proper inputs to the classifier, this is, feature engineering [5]. Most research efforts tried to create mean-ingful representations of the audio data to later feed gaussian mixture models (GMMs), hidden Markov models (HMM) or support vector machines (SVMs) [6]. In this context, a wide range of input representations were proposed such as Mel-frequency cepstral coefficients (MFCCs) [7], [8], Wavelets [8], constant-Q transform (CQT) or histograms of oriented gradientes (HOG) [9], among others.

With the years and the emergence of convolutional net-works in the field of image and computer vision, CNNs have become a preferred option for the design of machine listening

|  |  |
| --- | --- |
| VOLUME 4, 2016 | 1 |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

systems, usually fed with a 2D audio representation such as log-Mel spectrograms [1], [10]. These networks have shown very satisfactory results, especially when they are trained on large datasets. This is why data augmentation techniques are commonly applied, such as mixup strategies [11] or temporal cropping [12]. In addition, to improve the final accuracy, many studies use ensembles, combining the output from different classifiers to obtain a single more robust prediction. Unfortunately, the use of ensembles makes it more difficult to analyze the contribution to the classification performance of a new CNN architecture integrated within the proposed ensemble. To avoid such issue, this work considers isolated contributions of several CNN architectures implemented with different residual blocks based on squeeze-excitation, with-out any extra modifications during the training or inference phases.

CNNs are built with stacked convolutional layers. These layers learn its filter coefficients by capturing local spatial relationships (neighbourhood information) along the input channels and generate features maps (filtered inputs) by jointly encoding the spatial and channel information. In all application domains (image classification/segmentation, audio classification/tagging, etc.), the idea of encoding the spatial and the channel information independently has been less studied, despite having shown promising results [13], [14].

In order to provide insight about the behaviour of CNNs when analyzing spatial and channel information indepen-dently, several squeeze-excitation (SE) blocks have been presented in the image classification literature [13], [14]. In [14], a block that “squeezes” spatially and “excites” channel-wise with linear relationships was presented. The idea behind this block, denoted as cSE in this work, is to model the interdependencies between the channels of feature maps by exciting in a channel-wise manner. This type of block showed its effectiveness in image classification tasks, outperforming other state-of-the-art networks only by inserting it at a spe-cific point of the network. Following this idea, two more blocks were presented in [13]. The first one, denoted as sSE, “squeezes” along the channels and “excites” spatially, whereas the last block, scSE, combines both strategies. The scSE block recalibrates the feature maps along spatial and channel dimensions independently (cSE and sSE) and then combines the information of both paths by adding their outputs. This last block showed the most promising results in image-related tasks. According to [13], this block forces the feature maps to be more informative, both spatially and channel-wise.

This work analyzes the performance of conventional SE blocks for addressing the ASC problem and proposes two novel block configurations in this context. The new configu-rations are intended to enhance the benefits of residual learn-ing and feature map recalibration in a jointly fashion. This is achieved by a double short-cut connection that enforces residual learning both with and without recalibrated outputs. The use of SE techniques allows the network to extract

more meaningful information during training, while resid-ual learning facilitates the training procedure by mitigating vanishing gradient problems. The results show that, by using the proposed block configurations, results are considerably improved. Moreover, it is shown that all the residual SE configurations perform better than a classical convolutional residual block in the considered task.

The following of the paper is organized as follows. Sec-tion II presents the the background for the techniques used in this work in the context of ASC, namely Squeeze-Excitation and residual learning. Section III introduces the different SE blocks analyzed in this work and the baseline CNN architec-ture. Section IV describes the dataset used in the experiments, the audio pre-processing and the training procedure of the CNN. Section V discusses the experimental results, while Section VI concludes our work.

**II. BACKGROUND**

This section summarizes the technical background for this work and describes the ideas underlying SE blocks and residual networks.

**A. RELATED WORK**

Some previous works have shown that the use of SE mod-ules can be a simple and effective approach to tackle audio classification problems. In [15], a multi-scale fusion and channel weighted CNN was proposed within an ASC con-text. The framework consists of two stages: a multi-scale feature fusion scheme that integrates a hierarchy of semantic-features extracted from a simplified Xception architecture, and a final SE-based channel weighting stage. However, such work considers only channel recalibration by using a cSE-like block at a final stage, without further integration of other SE-based calibration modules. In contrast, the configurations proposed in this work consider both spatial and channel-wise weighting within a residual learning framework jointly and at multiple depths within the network architecture.

Another work using SE techniques in the audio domain is [16], which presented a VGG-style CNN and compared its performance with an enhanced version including residual connections and SE modules. In contrast to the work pre-sented in this paper, an end-to-end 1D architecture accepting raw audio inputs was proposed, with cSE channel-wise recal-ibration. The results over three different tasks (music auto-tagging, speech command recognition and acoustic event de-tection) confirmed the superiority of the enhanced network.

Finally, although some technical reports could not corrob-orate the improvements offered by SE modules over plain residual networks in audio-oriented tasks [17], few details were given, which motivates further the analysis carried out in this work.

**B. SQUEEZE-EXCITATION BLOCKS**

Squeeze-excitation (SE) blocks can be understood as mod-ules for channel recalibration of feature maps [13]. Let’ s assume an input feature map, X 2 RH W C0 , that feeds any

2 VOLUME 4, 2016

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

convolutional block, usually implemented by convolutional layers and non-linearities, and generates an output feature map U 2 RH W C. Here, U could also be expressed as

* = [u1; u2; : : : ; uC ], being ui 2 RH W a channel output. Considering this notation, H and W represents the height and the width, while C0 and C defines the number of input and output channels, respectively. The convolutional process function can be defined as F( ), so that F(X) = U. The output U is generated by combining the spatial and channel

information of X. The objective of SE blocks is to recalibrate

() ^ (): !^

U with FSE to generate U, i.e. FSE U U.

^

This recalibrated feature map, U, can be stacked after every convolutional block and then used as input to the forthcoming pooling layers. This recalibration can be carried out with dif-ferent types of block functions FSE( ), as it is next explained.

1) Spatial Squeeze and Channel Excitation Block (cSE)

In a cSE module (depicted in Fig. 1(a)) for spatial squeeze and channel excitation, a unique feature map of each channel from U is first obtained by means of global average pooling. This operator produces a vector z 2 R1 1 C . The kth element of such vector can be expressed as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | H | W |  |
| zk = | XXuk(i; j); k = 1; : : : ; C; (1) | |  |
|  |  |
| H W |  |
|  |  | i | j |  |

where uk(i; j) denotes the (i; j) element of the kth channel feature map.

As suggested by Eq. (1), global spatial information is embedded in vector z. This representation is then used to extract channel-wise dependencies using two fully-connected layers, obtaining the transformed vector z^. Therefore, z^ can be expressed as z^ = W1( (W2z)), where represents ReLU activation. W1 2 RC C and W2 2 RC C are the weights of the fully-connected layers, and is a ratio parameter. As last step, the activation range is compressed to the interval [0; 1] using a sigmoid activation function, . This final step indicates the importance of each channel and how they should be rescaled. The purpose of this recalibration is to let the network ignore channels with less information and emphasize the ones that provide more meaningful informa-

^

tion. Then, the rescaled feature maps, U, can be expressed as [13], [14]:

|  |  |  |
| --- | --- | --- |
| ^ | ; : : : ; (^zC )uC ]; (2) |  |
| UcSE = FcSE(U) = [ (^z1)u1 |  |

where z^k are the elements of the transformed vector z^.

2) Channel Squeeze and Spatial Excitation Block (sSE)

In the case of an sSE block [13], as shown in Fig. 1(b), a unique convolutional layer with one filter and (1; 1) kernel size is implemented to obtain a channel squeeze and spatial excitation effect. Here, it is assumed an al-ternative representation of the input tensor as U =

[u1;1; u1;2; : : : ; ui;j; : : : ; uH;W ] where ui;j 2 R1 1 C. The

convolution can be expressed as q = W ? U, being

* 2 R1 1 C 1 and q 2 RH W . Each qi;j represents the combination of all channels in location (i; j). As done with cSE, the output of this convolution is passed through a sigmoid function. Each (qi;j) determines the importance of the specific location (i; j) across the feature map. Like the previous block, this recalibration process indicates which locations are more meaningful during the training procedure. As a result, the output of the SE block can be expressed as [13]:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ^ | 1;1 | ; : : : ; (qH;W )u | H;W | ]: |  |
| UsSE = FsSE(U) = [ (q1;1)u |  |  |  |
|  |  |  |  | (3) |  |

1. Spatial and Channel Squeeze & Excitation block (csSE)

The csSE block [13] is implemented by declaring cSE and sSE blocks in parallel and adding both outputs (see Fig. 1(c)). It has been reported that the csSE block shows better perfor-mance than cSE and sSE used independently. In this case, a location (i; j; c) gets a higher sigmoid or activation value when both channel and spatial recalibration get it at the same time [13]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ^ | ^ | ^ | (4) |  |
| UcsSE = UcSE + UsSE: | | |  |

In this case, the network focuses on feature maps that are meaningful from both a spatial and channel-wise point of view.

**C. RESIDUAL NETWORKS**

Residual networks were first proposed in [18]. A network of this kind replaces the standard stacked convolutional layers [19] by residual blocks. Residual layers are designed to approximate a residual function: F(X) := H(X) X, where H( ) represents the mapping to be fit by a set of stacked layers and X represents the input to the first of such stacked layers. The original function H can therefore be defined as H(X) = F(X) + X. The main motivation of choosing this kind of network corresponds to the intuition that op-timizing a residual mapping may be easier than optimizng the original unreferenced one, as in a classical convolutional network. A simple way of implementing residual learning in CNNs is by adding a shortcut connection that performs as an identity mapping, adding back the input X to the output of the residual block F(X). In the first proposition of the residual block, Rectified ReLU activation is applied after the addition and the result of such activation becomes the input for the next residual block. Note, that in the first configura-tion, shortcut connections do not add more parameters nor extra computational cost. Therefore, deeper networks can be trained with little additional effort, reducing vanishing-gradient problems. As it will be later explained, in this work, the identity mapping is replaced with a 1 1 convolutional layer as it is explained in Section III. Therefore, this work function can be expressed as H(X) = F(X)+g(X), where

|  |  |
| --- | --- |
| VOLUME 4, 2016 | 3 |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

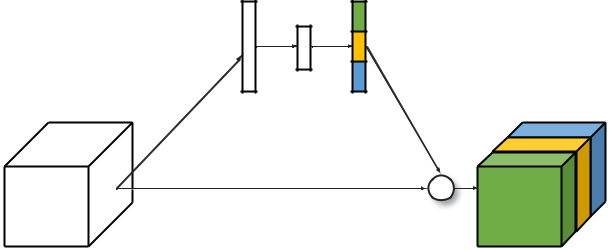
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

**cSE**

|  |  |  |  |
| --- | --- | --- | --- |
| ***C***+***1*** | ***C*** |  |  |
|  | ***C* // *r*** |  |  |
|  | **ReLU** |  |  |
| **U** | **sigmoid** | **U**ˆ |  |
| *C* |  |  |  |
| *H* |  | + |  |
|  |  |  |

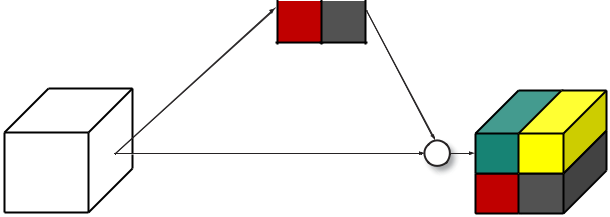


*W*

**(a)**

**sSE**

*H* 



|  |  |  |
| --- | --- | --- |
|  | *W* |  |
| **U** | **U**ˆ |  |
| *C* |  |  |
| *H* | + |  |
|  |  |

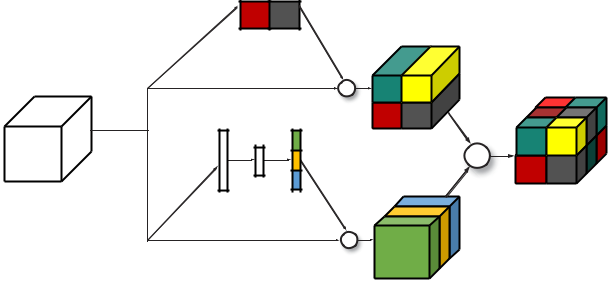
*W*

**(b)**

**csSE**



*H* 



*W*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **U** |  |  | + | **U**ˆ |  |
| *C* |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  | ***C***+***1*** | ***C*** |  |  |
|  |  |  |  | ***C // r*** |  |  |
| *H* |  |  |  |  | + |  |
|  |  |  |  |  |  |
| *W* |  | **averge** | **pooling** | **ReLU** |  |  |
|  |  | **sigmoid** |  |  |
|  |  |  |  |  |  |
|  | **global** |  |  |  |  |  |
|  |  |  |  | + |  |  |

**(c)**

**FIGURE 1.** Diagram of different SE blocks: (a) describes cSE blockprocedure, (b) ilustrates sSE block framework and (c) shows csSE block by combining (a) and (b).

**A. SE BLOCK DESCRIPTION**

All the configurations analyzed in this work are depicted in Fig. 2. In the following, we describe in details these blocks.

a: Conv-residual

Shown in Fig. 2(a), is inspired by [18]. It is used as a baseline in order to validate the network performance without any SE and how much it can be improved when incorporating these blocks. In the present work some slight modifications for a more convenient implementation were introduced: the shortcut connection was implemented with a 1 1 convolu-tional layer and the activation after the addition was set to an exponential linear unit (ELU) function [20], [21].

b: Conv-POST

Shown in Fig. 2(b), is inspired by the block referred to as se-POST in [14]. The scSE block is included at the end and is equivalent to a recalibration of the Conv-residual block.

c: Conv-POST-ELU

Shown in Fig. 2(c), is very similar to the above Conv-POST block, but the recalibration is performed over the ELU-activated output of the residual block.

d: Conv-Standard

Shown in Fig. 2(d), is inspired by [14], where the scSE block is stacked after the convolutional block for recalibrating prior to adding the shortcut branch.

e: Conv-StandardPOST

g( ) represents the convolutional process with the learnt filter coefficients.

1. **CONFIGURATIONS FOR SQUEEZE-AND-EXCITATION RESIDUAL NETWORKS**

According to [14], SE blocks exhibit better performance when deployed on networks with residual configuration than on VGG-style networks. Therefore, two novel residual blocks implementing scSE modules are presented in this paper. The performance of these two newly proposed blocks is compared against other state-of-the-art residual configurations that in-corporate SE modules.

4

Shown in Fig. 2(e) is proposed in this work to create a double shortcut connection, one before SE calibration and one after. The idea is to let the network learn residual map-pings simultaneously with and without SE recalibration, thus, affecting the way in which the block optimizes the residual by considering jointly standard and post SE-calibrated outputs.

f: Conv-StandardPOST-ELU

Shown in Fig. 2(f) is the other proposed block, correspond-ing to the above explained Conv-StandardPOST block, but followed by ELU activation.

To summarize, the output Xl+1 of each block for an input

X is given by:

VOLUME 4, 2016

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

|  |  |  |
| --- | --- | --- |
| a) | Xl+1 = | R (F(X) + g(X))) ; |
|  |  | (5) |
| b) | Xl+1 = | FSE (F(X) + g(X))) ; |
|  |  | (6) |
| c) | Xl+1 = | FSE (R (F(X) + g(X)))) ; |
|  |  | (7) |
| d) Xl+1 = | | F(SE)(X) + g(X); |
|  |  | (8) |

1. Xl+1 =FSE (R (F(X) + g(X)))) + g(X);

(9)

1. Xl+1 = R (FSE (R (F(X) + g(X)))) + g(X)) ; (10)

where R( ) refers to ELU activation function with pa-rameter set to 1 and F(SE) denotes a residual function that includes SE calibration. As it will be discussed in Section V, the two proposed configurations have been shown to outper-form the rest in the considered acoustic scene analysis task.

In order to avoid possible duplications or expansion pro-cesses in the channel dimension, the identity branch is re-placed by a convolutional layer with a (1; 1) kernel size and with the same number of filters as the residual branch. Including such convolutional layer in the shortcut branch creates a projection that avoids dimensionality conflicts in the residual block addition.

By looking at Fig. 2, it can be clearly observed that the most representative feature of the two proposed blocks, (e) and (f), resides in the use of two skip connections: one before SE re-calibration and one after. This double short-cut connection leads the network towards the learning of a global residual function embedding an inner and SE-calibrated par-tial residual. The objective is to facilitate the learning of calibration weightings by using the same residual rationale.

In general, the presence or absence of relevant acoustic events within an input audio clip can be very important when addressing the ASC problem. The use of spatial and channel-wise recalibration at different depths of the network adds a mechanism to allow the network weight properly, according to their importance, the different dimensions of the information flowing throughout the network. Therefore, SE modules are expected to add flexibility for identifying relevant acoustic textures or events, making easier to infer the type of underlying acoustic scene.

**B. NETWORK ARCHITECTURE**

The CNN implemented in order to validate the behaviour of the different SE configurations has been inspired on

1. where a VGG-style [19] network with 3 convolutional blocks followed by different max-pooling and dropout [23] operators is implemented. In the present work, the original convolutional blocks have been replaced with the different residual squeeze-excitation blocks proposed in this study. The max-pooling, dropouts and linear layers are configured

**TABLE 1.** Proposed network for validating the scSE configurations of Fig. 2.Values preceeded by # correspond to the number of filters. Kernel sizes are set as indicated in Fig. 2. This architecture is inspired by the work in [22]

Residual-scSE block (#32, = 2)

MaxPooling(2,10)

Dropout(0.3)

Residual-scSE block (#64, = 2)

MaxPooling(2,5)

Dropout(0.3)

Residual-scSE block (#128, = 2)

MaxPooling(2,5)

Dropout(0.3)

Flatten

[Dense(100), batch normalization, ELU]

Dropout(0.4)

[Dense(10), batch normalization, softmax]

**TABLE 2.** Parameters and FLOPs analysis from the studied networkconfigurations

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Residual block |  | Network parameters |  | Network FLOPs |  |
| Conv-Residual |  | 506606 |  | 1009115 |  |
|  |  |  |
|  |  |  |  |  |  |
| Conv-POST |  |  |  |  |  |
| Conv-POST-ELU |  |  |  |  |  |
| Conv-Standard | 528334 | | 1052571 | |  |

Conv-StandardPOST

Conv-StandardPOST-ELU

with the same parameters as in [22]. The network architecture can be found in Table 1.

As the database used in the current work is much smaller than the one in [14], some of the hyperparameters that define the components of the csSE block had to be modified. The number of elements in the Dense layer with ReLU activation in Fig. 1(a) has been set to 16 in the first Residual-scSE block, the same as in [14] in its cSE block, but the number of filters at the input, C, has been set to C = 32. Therefore, the ratio between these parameters throughout the network is two, as observed Table 1. The number of network parameters that implement SE residual blocks, i.e. those represented in Fig. 2(b)-(f), is 528,334. On the other hand, the network that does not integrate SE modules has 506,606. Note, therefore, that there is only a slight increase of approximately 4% in the SE networks. Table 2 shows as well the number of floating point operations (FLOPs) involved in each network.

**IV. EXPERIMENTAL DETAILS**

This section describes in detail the experimental implemen-tation carried out to conduct the analysis of the presented SE residual blocks, including the datasets, the audio rep-resentation selected to feed the network and the training configuration.

|  |  |
| --- | --- |
| VOLUME 4, 2016 | 5 |

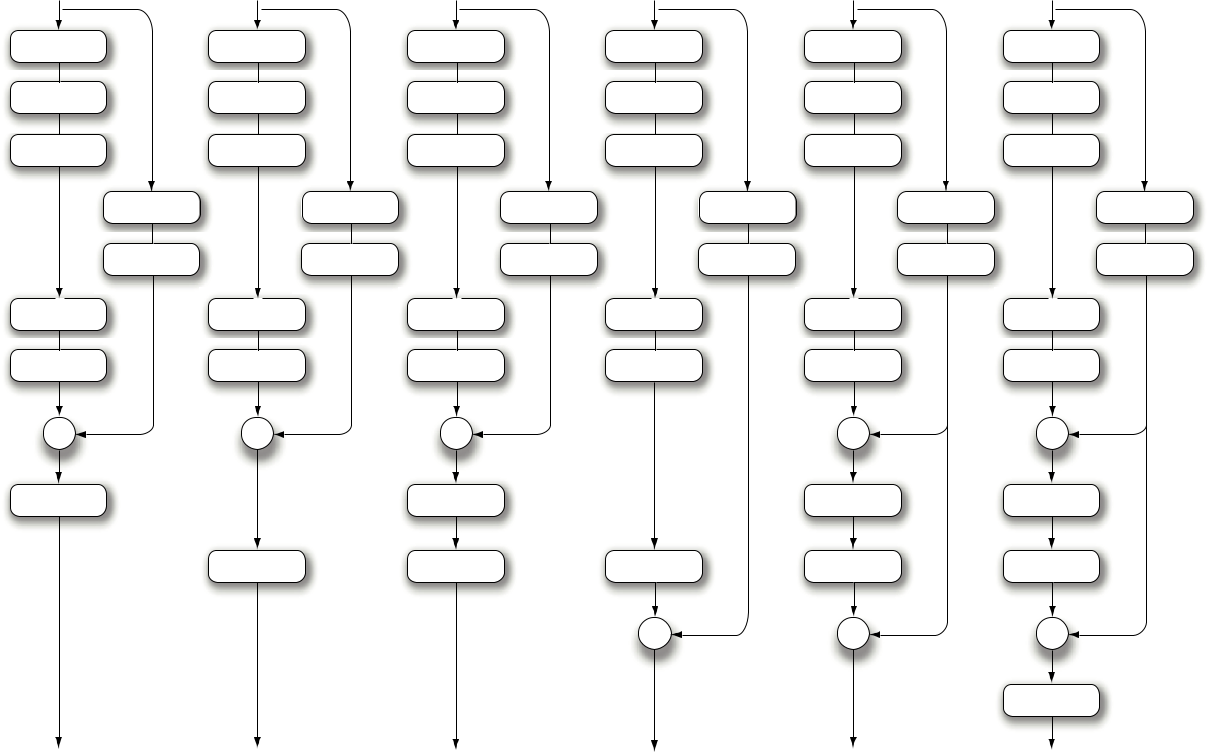
This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **X**l | **X**l | **X**l | **X**l | **X**l |  | **X**l |
| 3 x 3 | 3 x 3 | 3 x 3 | 3 x 3 | 3 x 3 |  | 3 x 3 |
| BN | BN | BN | BN | BN |  | BN |
| ELU | ELU | ELU | ELU | ELU |  | ELU |
| 1 x 1 | 1 x 1 | 1 x 1 |  | 1 x 1 | 1 x 1 | 1 x 1 |
| BN | BN | BN |  | BN | BN | BN |
| 3 x 3 | 3 x 3 | 3 x 3 | 3 x 3 | 3 x 3 |  | 3 x 3 |
| BN | BN | BN | BN | BN |  | BN |



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| + | + | + |  | + | + |
| ELU |  | ELU |  | ELU | ELU |
|  | scSE | scSE | scSE | scSE | scSE |
|  |  |  | + | + | + |
|  |  |  |  |  | ELU |
| **X**l+1 | **X**l+1 | **X**l+1 | **X**l+1 | **X**l+1 | **X**l+1 |
| **(a)** | **(b)** | **(c)** | **(d)** | **(e)** | **(f)** |
| **Conv-Residual** | **Conv-POST** | **Conv-POST-ELU** | **Conv-Standard** | **Conv-StandardPOST Conv-StandardPOST-ELU** | |

**FIGURE 2.** Different residual squeeze-excitation blocks analyzed in this work: (a) is inspired by the first residual block proposed in [18]; (b), (c) and (d) are inspiredby the work done in [14]; (e) and (f) are the two novel configurations proposed in this work.

**A. DATASET**

To check the behavior of these implementations in an ASC problem, the TAU Urban Acoustic Scenes 2019, Devel-opment dataset presented in Task 1A of the 2019 edition of DCASE has been used [10]. The database consists of 40 hours of stereo audio-recording in different urban en-vironments and landscapes such as parks, metro stations, airports, etc. making a total of 10 different scenes. These have been recorded in different cities such as Barcelona, Paris or Helsinki, among others. All audio clips are 10-second long. They are divided into two subsets of 9185 and 4185 clips for training and validation, respectively. Although there are a slightly different number of samples available for each class, the data set is not severely unbalanced.

**B. AUDIO PROCESSING**

The input to the network is a 2D log-Mel spectrogram representation with 3 audio channels. The three channels are composed of the harmonic and percussive component [24], [25] of the signal converted to mono and the differ-ence between left (L) and right (R) channels. That is, the first channel corresponds to the log-Mel spectrogram of the harmonic source, the second channel corresponds to the same representation but over the percussive source and the last

6

one to the log-Mel spectrogram of the difference between channels calculated by subtracting left and right channels (L R). This representation, known as HPD, was presented in [22]. The log-Mel spectrogram is calculated using 64 Mel filters with a window size of 40 ms and 50% overlap. Therefore, an audio clip becomes a 64 T 3 array with

* being the number of time frames. In this specific dataset, the input audio representation corresponds to an array of dimension 64 500 3.
  1. **TRAINING PROCEDURE**

The training process was optimized using the Adam op-timizer [26]. The cost function used was the categorical crossentropy. Training was limited to a maximum of 500 epochs but early stopping is applied if the validation accuracy does not improve by 50 epochs. If this same metric does not improve in 20 epochs, the learning rate is decreased by a factor of 0.5. The batch size used was 32 samples.

**V. RESULTS**

In order to analyze the contributions of this work with respect to other state-of-the-art approaches, the results obtained with the different configurations presented in this work (see Fig. 2) are compared to the ones obtained by different authors in

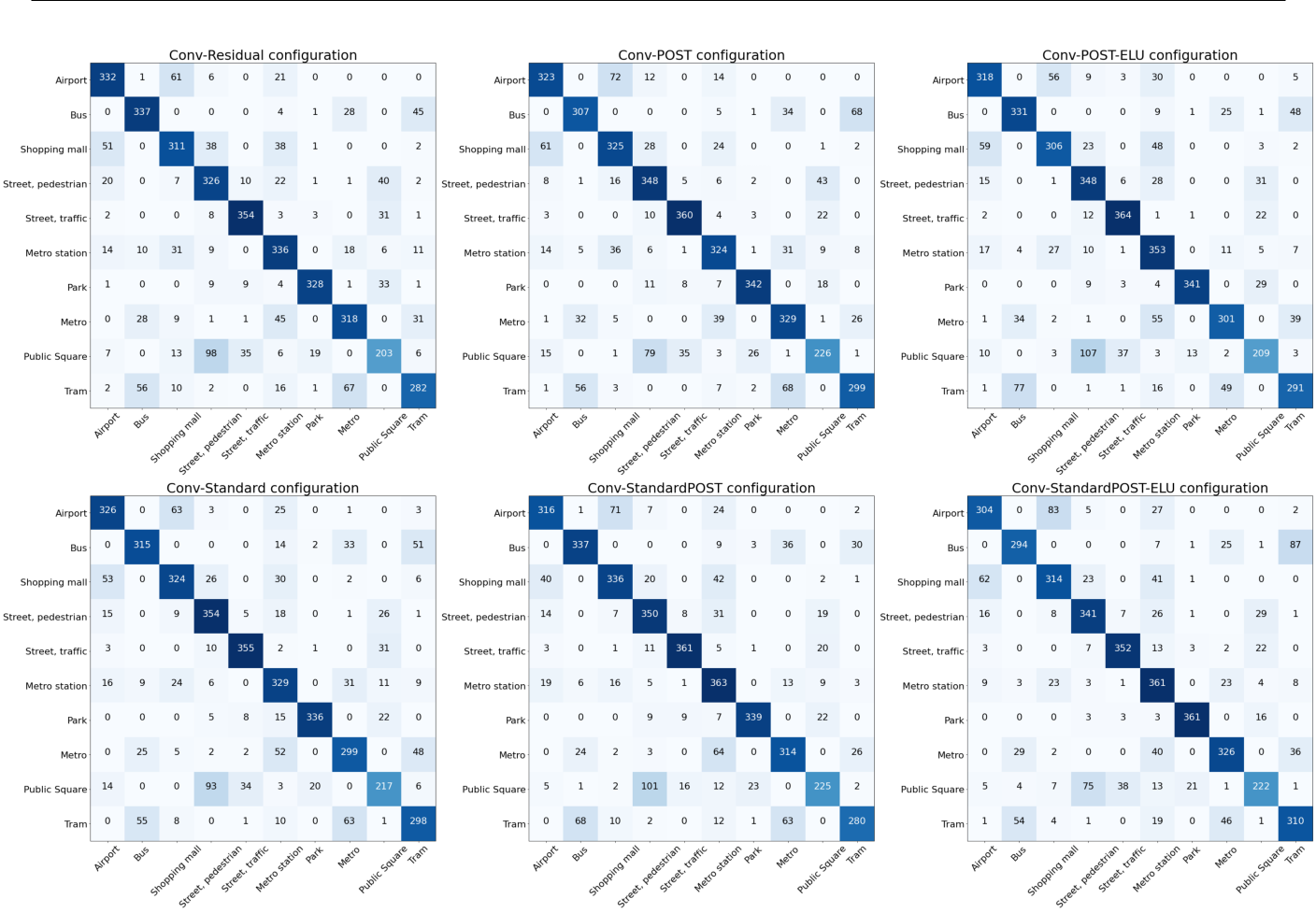
VOLUME 4, 2016

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks



**FIGURE 3.** Confusion matrices for the generated models over the evaluation dataset.

**TABLE 3.** Accuracy results from the validation partition in development phase.

|  |  |
| --- | --- |
| System | Development accuracy (%) |
|  |  |
|  |  |
| Baseline [10] | 62.5 |
|  |  |
| Wang\_NWPU\_task1a [27] | 72.4 |
|  |  |
| Fmta91\_KNToosi\_task1a [28] | 70.49 |
|  |  |
| MaLiu\_BIT\_task1a [29] | 76.1 (evaluation) |
|  |  |
| DSPLAB\_TJU\_task1a [30] | 64.3 |
|  |  |
| Kong\_SURREY\_task1a [31] | 69.2 |
|  |  |
| Liang\_HUST\_task1a [32] | 70.70 |
|  |  |
| Salvati\_DMIF\_task1a [33] | 69.7 |
|  |  |
| Conv-Residual | 74.51 0.65 |
| Conv-Standard | 75.16 0.33 |
| Conv-POST | 75.84 0.65 |
| Conv-POST-ELU | 75.81 0.47 |
| Conv-StandardPOST | 76.72 0.59 |
| Conv-StandardPOST-ELU | 76.00 0.55 |

Task 1A of DCASE 2019 using the same dataset. For a fair comparison, only submissions not making use of data augmentation techniques are considered. In the case of sub-missions that presented an ensemble of several models, only the results of the best performing model making up the en-semble are taken into account. For example, in [27] a global development accuracy of 78.3% is reported, but that value was obtained by averaging 5 models. The best individual model obtained 72.4%, so this is the value presented in Table

1. This said, please be aware that the accuracy of the final submission1 may differ from that presented in Table 3. Next, we summarize some important features of the competing approaches.

Wang\_NWPU\_task1a [27]: the audio representation considers two channels using a log-Mel Spectrogram from harmonic and percussive sources similar to our representation. The number of Mel filters is set to 256. The window size is set to 64 ms and the hop size to 15 ms. Mel filters are calculated with cutoff frequencies from 50 Hz to 14 kHz. A VGG-style CNN [19] is used as a classifier.

1http://dcase.community/challenge2019/task-acoustic-scene-classification-results-a

|  |  |
| --- | --- |
| VOLUME 4, 2016 | 7 |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

Fmta91\_KNToosi\_task1a [28]: wavelet scattering spectral features are extracted from the mono audio signal. A random subspace method is used as classifier.

MaLiu\_BIT\_task1a [29]: Deep Scattering Spectra fea-tures (DSS) are extracted from each stereo channel. Classification is performed with a Convolutional Recur-rent Neural Network (CRNN). For this network, Table 3 does not report the accuracy on the development set (only on the evaluation set). This is because of some mismatch reported by the authors in the validation pro-cedure with the configuration of the dataset.

DSPLAB\_TJU\_task1a [30]: this submission ap-proaches the problem in a more classical way extracting audio statistical features such as ZRC, RMSE, spectro-gram centroid, etc. A GMM is used as a classifier.

Kong\_SURREY\_task1a [31]: this submssion can be defined as the state-of-the-art framework in ASC prob-lem. The audio representation considers also the log-Mel spectrogram. The classifier is a VGG-based [19] CNN. This network is a fully convolutional network with no linear layers implemented. The feature maps are reshaped into a one dimensional vector using a global average pooling before the decision layer.

Liang\_HUST\_task1a [32]: in this method, the log-Mel spectrogram is first extracted after converting the audio signal to mono. Interestingly, the log-Mel spectrogram is divided into two-seconds spectrograms, that means that spectrogram shapes change from [F T 1] to [F (T =5) 1]. This configuration allows training with audio samples consisting of 5 different spectrograms instead of one. A CNN with frequency attention mecha-nism is implemented as classifier. For more detail of the attention implementation, see [32].

Salvati\_DMIF\_task1a [33]: unlike the other submis-sions, this one works directly on the audio vector. To this end, a 1D convolutional network is implemented. Although some recent efforts have been made in this direction [34], the state-of-the-art literature shows that 2D audio representations, such as spectrograms, still obtain the better classification results [35].

DCASE baseline [10]: the audio is first converted to mono and a log-Mel spectrogram is extracted. In this case, only 40 Mel bins are calculated instead of 64, which is the typical state-of-the-art choice. A CNN is used as a classifier with 2 convolutional layers. The 1D conversion before classification layers is performed by a flatten layer. A dense layer is stacked before the decision layer.

1. **GLOBAL PERFORMANCE**

Although the results of the DCASE challenge only report the mean accuracy value, we consider 10 runs to provide not only the mean accuracy value, but also the standard deviation. As it can be seen in Table 3, all the configurations detailed in Fig. 2 obtain better accuracy than the DCASE baseline. The contribution of the scSE block is easily justified as Conv-

Residual gets the lowest performance among the studied configurations. In general, POST configurations show a slight improvement compared to the Standard configuration. This behaviour differs from what was reported in the original paper, [14], in which these blocks were analyzed in the image domain, where the Standard block outperforms the POST block. There is no remarkable difference between Conv-POST and Conv-POST-ELU. It is also shown that the networks that incorporate the two novel blocks presented in this work, the ones depicted in Figs. 2(e) and (f), exhibit the best accuracy values. The shortcut addition at two differente points of the residual block, this is, before and after the scSE block, allowed the network to obtain a more precise classification in this ASC task.

**B. CLASS-WISE PERFORMANCE**

Fig. 3 shows confusion matrices for each of the analyzed residual blocks in this work. In general, the performance across the different classes is considerably balanced. The “Public square” class is the one showing the worst perfor-mance, tending to be misclassified as “Street, Pedestrian”. Other similar classes such as “Airport” and “Shopping mall” or “Tram” and “Bus” or “Metro” tend also to produce com-mon errors in the analyzed networks.

By analyzing the class-wise performance of the two pro-posed blocks with respect to the conventional Conv-Residual block, substantial improvements are observed. Considering the proposed Conv-StandardPOST block, a significant im-provement is observed for the classes “Metro station” and “Street, Pedestrian”. Other classes showing slight improve-ments are “Shopping mall”, “Park” or “Public square”. The class showing the worst relative result was “Airport”. On the other hand, the second proposed block Conv-StandardPOST-ELU provides substantial improvements in “Street traffic” and “Park”, but other classes like “Airport” or “Bus” were degraded.

Finally, when considering the performance of networks implementing SE blocks together, from a general perspective, it is noticed that classes like “Street, Pedestrian”, “Park” or “Public square” are improved with respect to the con-ventional residual network. Only the class “Airport” shows the best performance in the conventional network, followed by “Bus”. The remaining classes are improved or worsened across all configurations in a degree not as significant as the aforementioned ones.

**C. SIGNIFICANCE TEST**

To determine if there are statistically significant differences in the performance of the different blocks analyzed in this work, a McNemar’s test has been carried out [36]. This test, which is a paired non-parametric hypothesis test, has been widely recommended for evaluating deep learning models, which are often trained on very large datasets. The test is based on a contingency table created from the results obtained for two methods trained on exactly the same training test and evaluated on the same test set. The null hypothesis of

8 VOLUME 4, 2016

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Residual |  |  | POST |  |  | POST |  |  |  | Standard |  |  | StandardPOST StandardPOST | | - | ELU |  |
|  |  |  |  |  |  | - | ELU |  |  |  |  |  |  |
| Conv | - |  | Conv | - |  | Conv | - |  |  | Conv | - |  | Conv | - | Conv | - |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Conv-Residual

Conv-POST

Conv-POST-ELU

Conv-Standard

Conv-StandardPOST

Conv-StandardPOST-ELU

*p* < 0.05

**FIGURE 4.** Pairwise analysis of the studied residual networks usingMcNemar’s test. Gray cells indicate p-values below a 0.05 significance level.

the test is that the performance of the two analyzed systems disagree to the same amount. If the null hypothesis is re-jected, there is evidence to suggest that the two systems have different performance when trained on a particular training set. Given a significance level , if p < , there may be sufficient evidence to claim that the two classifiers show different proportions of errors. The result of applying the McNemar’s test to all the available system pairs is shown in Fig. 4. Gray cells indicate p-values below a significance level of 0.05. It is confirmed that the two proposed blocks, Conv-StandardPOST and Conv-StandardPOST-ELU, show significant differences in performance with respect to all the other blocks but Conv-POST, which was the third best performing block. However, no significant differences can be observed between these new blocks, which only differ in the final ELU activation.

**VI. CONCLUSION**

The use of squeeze-excitation blocks in convolutional neural networks allows to perform a spatial and channel-wise recal-ibration of its inner feature maps. This work presented the use of squeeze-excitation residual networks for addressing the acoustic scene classification problem, and presented two novel block configurations that consider residual learning of standard and recalibrated outputs jointly. Results over the well-known DCASE dataset confirm that the proposed blocks provide meaningful improvements by adding a slight architecture modification, outperforming other competing approaches when no data augmentation or model ensembles are considered.

**REFERENCES**

1. M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, “A convolutional neural network approach for acoustic scene classification,” in 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017, pp. 1547–1554.
2. S. H. Bae, I. Choi, and N. S. Kim, “Acoustic scene classification using parallel combination of lstm and cnn,” in Proceedings of the Detec-tion and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), 2016, pp. 11–15.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| [3] L. | Pham, H. Phan, T. Nguyen, R. | Palaniappan, | A. Mertins, | | and |
| I. | McLoughlin, “Robust acoustic | scene classification | | using | a |
| multi-spectrogram encoder-decoder | | framework,” | arXiv | preprint | |
| arXiv:2002.04502, 2020. | |  |  |  |  |

1. Y. Han and K. Lee, “Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation,” arXiv preprint arXiv:1607.02383, 2016.
2. I. Martín-Morató, M. Cobos, and F. J. Ferri, “A case study on feature sensitivity for audio event classification using support vector machines,” in 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2016, pp. 1–6.
3. I. Martín-Morató, M. Cobos, and F. J. Ferri, “Adaptive mid-term represen-tations for robust audio event classification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 12, pp. 2381–2392, 2018.
4. G. Roma, W. Nogueira, and P. Herrera, “Recurrence quantification analysis features for auditory scene classification,” DCASE2013 Challenge, Tech. Rep., June 2013.
5. D. Li, J. Tam, and D. Toub, “Auditory scene classification using machine learning techniques,” DCASE2013 Challenge, Tech. Rep., June 2013.
6. A. Rakotomamonjy and G. Gasso, “Histogram of gradients of time-frequency representations for audio scene classification,” DCASE2013 Challenge, Tech. Rep., June 2013.
7. A. Mesaros, T. Heittola, and T. Virtanen, “Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups,” 2019.
8. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.
9. W. Gao and M. McDonnell, “Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths,” DCASE2019 Challenge, Tech. Rep., June 2019.
10. 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 Inter-vention. Springer, 2018, pp. 421–429.
11. 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
12. L. Yang, X. Chen, L. Tao, and X. Gu, “Multi-scale fusion and channel weighted cnn for acoustic scene classification,” in Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning, 2019, pp. 41–45.
13. J. Lee, T. Kim, J. Park, and J. Nam, “Raw waveform-based au-dio classification using sample-level cnn architectures,” arXiv preprint arXiv:1712.00866, 2017.
14. O. Akiyama and J. Sato, “Multitask learning and semisupervised learning with noisy data for audio tagging,” DCASE2019 Challenge, Tech. Rep, Tech. Rep., 2019.
15. 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.
16. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
17. A. Shah, E. Kadam, H. Shah, S. Shinde, and S. Shingade, “Deep resid-ual networks with exponential linear unit,” in Proceedings of the Third International Symposium on Computer Vision and the Internet, 2016, pp. 59–65.
18. D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” arXiv preprint arXiv:1511.07289, 2015.
19. 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.
20. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdi-nov, “Dropout: a simple way to prevent neural networks from overfitting,” The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.
21. D. Fitzgerald, “Harmonic/percussive separation using median filtering,” in Proc. of DAFX, vol. 10, no. 4, 2010.

|  |  |
| --- | --- |
| VOLUME 4, 2016 | 9 |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002761, IEEE Access



Javier Naranjo-Alcazar et al.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks

1. J. Driedger, M. Müller, and S. Disch, “Extending harmonic-percussive separation of audio signals.” in ISMIR, 2014, pp. 611–616.
2. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
3. M. Wan, R. Wang, B. Wang, J. Bai, C. Chen, Z. Fu, J. Chen, X. Zhang, and S. Rahardja, “Ciaic-asc system for dcase 2019 challenge task1,” Tech. Rep., DCASE2019 Challenge, Tech. Rep., 2019.
4. F. Arabnezhad and B. Nasersharif, “Urban acoustic scene classification using binaural wavelet scattering and random subspace discrimination method,” DCASE2019 Challenge, Tech. Rep., June 2019.
5. S. Ma and W. Liu, “Acoustic scene classification based on binaural deep scattering spectra with neural network,” DCASE2019 Challenge, Tech. Rep., June 2019.
6. B. Ding, G. Liu, and J. Liang, “Acoustic scene classification based on ensemble system,” DCASE2019 Challenge, Tech. Rep., June 2019.
7. Q. Kong, Y. Cao, T. Iqbal, W. Wang, and M. D. Plumbley, “Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems,” DCASE2019 Challenge, Tech. Rep., June 2019.
8. H. Liang and Y. Ma, “Acoustic scene classification using attention-based convolutional neural network,” DCASE2019 Challenge, Tech. Rep., June 2019.
9. D. Salvati, C. Drioli, and G. L. Foresti, “Urban acoustic scene classifi-cation using raw waveform convolutional neural networks,” DCASE2019 Challenge, Tech. Rep., June 2019.
10. J. Naranjo-Alcazar, S. Perez-Castanos, I. Martin-Morato, P. Zuccarello, and M. Cobos, “On the performance of residual block design alternatives in convolutional neural networks for end-to-end audio classification,” arXiv preprint arXiv:1906.10891, 2019.
11. S. Dieleman and B. Schrauwen, “End-to-end learning for music audio,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 6964–6968.
12. T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” Neural computation, vol. 10, no. 7, pp. 1895–1923, 1998.

JAVIER NARANJO-ALCAZAR (S’ 20) received the telecommunications degree and the master´s degree in telecommunications engineering from the Universitat Politècnica de València, València, Spain in 2016 and 2018, respectively. He was the recipient of the Best MSc Thesis Award from the Regional Telecommunications Engineering Asso-ciation in 2019. Currently, he is working toward the Ph.D. degree with the Department of Computer Science, Universitat de València, funded by the



Torres Quevedo program and the valencian start-up Visualfy. His research interests lie in the field of machine listening, few-shot learning and open-set recognition.

SERGI PEREZ-CASTANOS received the telecommunications degree and the master´s de-gree in telecommunications engineering from the Universitat Politècnica de València, València, Spain in 2016 and 2018, respectively. Currently, he is working as a machine learning engineer at Visualfy. His research interests lie in the field of machine listening, anomaly detection and audio captioning.



PEDRO ZUCCARELLO received an Electronics Engineering degree from the University of Buenos Aires (Argentina), a M.Sc. degree in Telecom-munications from the Universitat Politècnica de València (Valencia, Spain) and a Ph.D. degree from the Universitat de València (Valencia, Spain). He developed most his career as a researcher in several public R&D institutions such as the Insti-tute of Microelectronics of Barcelona (Barcelona, Spain) or the Institute of Corpuscular Physics



(Valencia, Spain). Since 2017 he is the Head of the Artificial Intelligence group at Visualfy private startup company. He received several postdoctoral fellowships such as Val-I+D, from the Valencian Government, or Torres Quevedo, from the Spanish Ministry of Science and Education. He co-authored nearly 30 papers in international peer-review journals and confer-ences in topics that include artificial intelligence, machine learning, signal processing, electronics and microelectronic design.

MAXIMO COBOS (SM’12) received the mas-ter’s degree in telecommunications and the Ph.D. degree in telecommunications engineering from the Universitat Politècnica de València, València, Spain, in 2007 and 2009, respectively. He com-pleted with honors his studies under the Univer-sity Faculty Training program (FPU) and was the recipient of the Ericsson Best Ph.D. Thesis Award from the Spanish National Telecommunications Engineering Association. In 2010, he received a



Campus de Excelencia postdoctoral fellowship to work at the Institute of Telecommunications and Multimedia Applications. In 2011, he joined the Universitat de València, where he is currently an Associate Professor. His work is focused on the area of digital signal processing and machine learning for audio and multimedia applications, where he has authored/coauthored more than 100 technical papers in international journals and conferences. He is a member of the Audio Signal Processing Technical Committee of the European Acoustics Association and serves as Associate Editor for IEEE Signal Processing Letters.



10 VOLUME 4, 2016

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.