ENVIRONMENTAL SOUND CLASSIFICATION USING A RESIDUAL NETWORK ARCHITECTURE

Arindam Dey

Master’s program in Data Science - Liverpool John Moores University and upGrad

INTRODUCTION

- Background
- Problem Statement

BACKGROUND

- In late nineties, audio data was mainly characterized by name, file-format, sampling rate etc.
- Audio application were primarily limited to archiving, storing and separation of audio sources based on very basic characteristics.
- As many parallel technologies evolved, newer use cases of environmental awareness started emerging.
- Among them , there were use cases that required mimicking human perception based on sight and sound involving object-detection and source classification.
- Machine Learning techniques applied to audio sound classification (ASC) were limited to KNN, SVM, GMM.
- Instances of deep learning techniques started appearing in 2009 with limited datasets. Only from 2014, labelled datasets like ESC-10 and ESC-50 were available for benchmarking different models.
- Sequential Deep Learning Models based on ESC dataset started appearing in 2015 .

PROBLEM STATEMENT

- Since 2015 most of the research on Deep Learning based ASC Task were on centered around Sequential CNN Architectures.
- These models use backpropagation algorithm , that incrementally update the model weights so that the model can 'learn'.
- The algorithm suffers from the problem of vanishing gradients , as the depth of the model increases.
- This is because , as the updates are propagated back to the beginning of a model, they become smaller and smaller with the depth.
- Can building incrementally deeper CNN Architectures and combating Vanishing Gradients by using skip/residual connections lead to better accuracies for audio classification problems ? This leads us to a ResNet Architecture.
- Whether MFCC features alone can lead us to >76% accuracies on the ESC-10 Dataset?

LITERATURE REVIEW

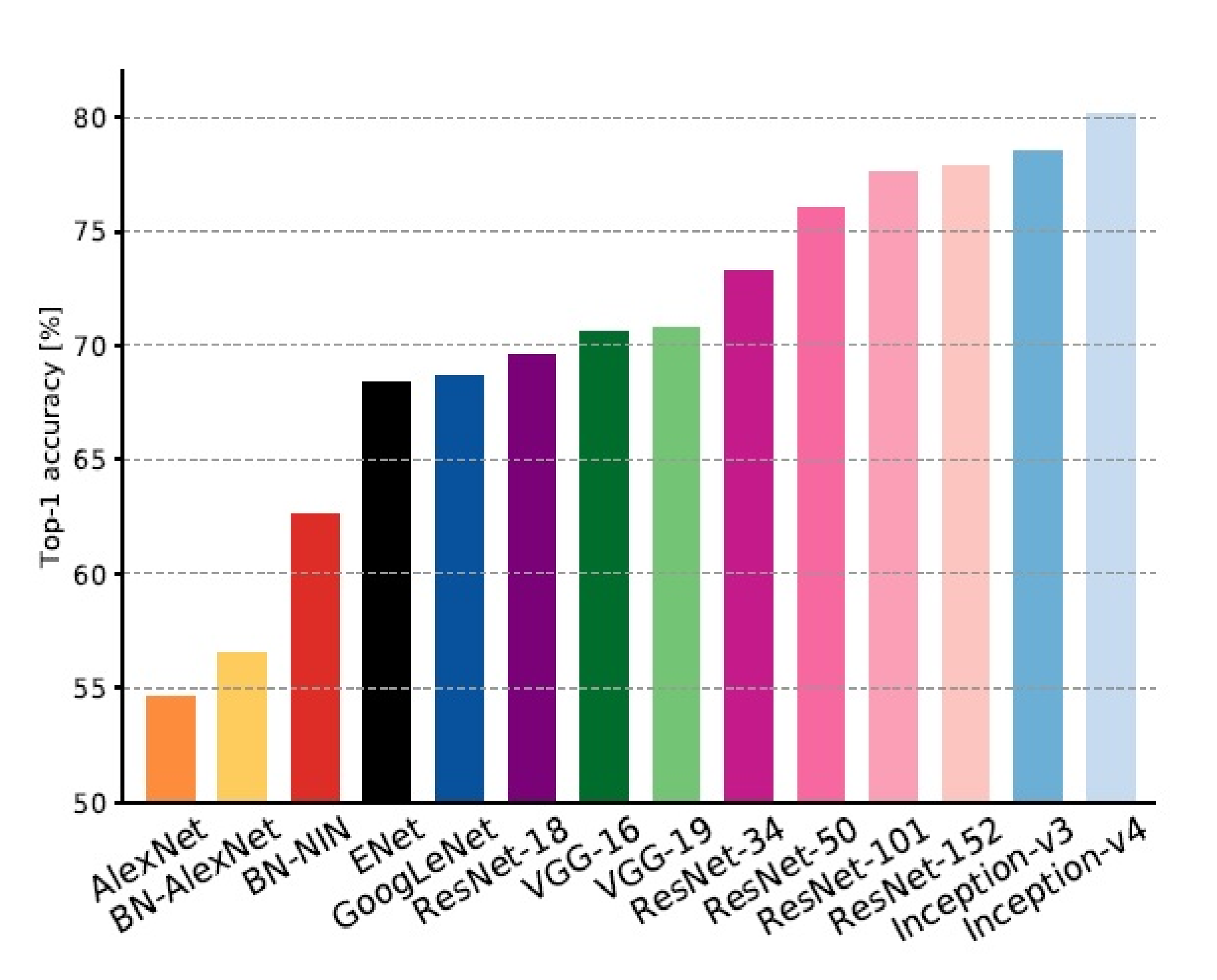
- Early Work on ASC
- Evolution of CNN Architectures
- Residual Connections
- Audio Datasets ESC and Urbansound8K
- CNN for Audio Classification

EARLY WORK ON ASC

- Before the ESC task , most audio classification was related to Speech/Non-Speech, Music/Movie Genre Classification or musical instrument classification.
- Wold et al (1996) , built a Euclidean Distance Based Classifier based on loudness, pitch, brightness and bandwidth. It's purpose was to fetch audio from a database based on acoustical and perceptual features.
- Saunders (1996) reported Zero Crossing Rate (ZCR)Based Speech/Music discrimination as both have distinctly different ZCR characteristics.
- More complex discriminator was built by Scheirer (1997) based on a 13 feature representation and using them on GMM, k-NN and k-d classifiers. They found music was harder to classify than speech.
- Pierangelo (2002) used the findings of Saunders to build a ZB (ZCR Bayesian Classifier) based Speech/Music discriminator. When they compared it with a Neural Network, it outperformed the former by 11% in terms of Total Error Rate.
- There were multiple such efforts based on different datasets , making it difficult to benchmark them against each other.

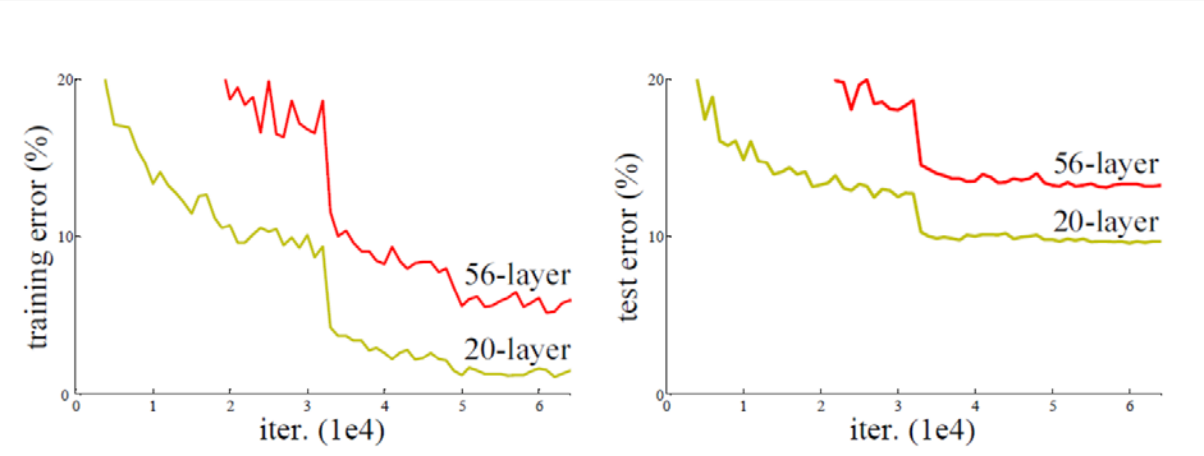
EVOLUTION OF CNN ARCHITECTURES

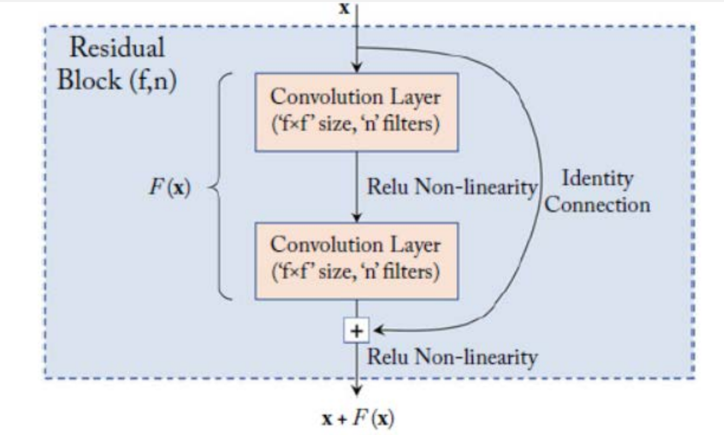
- Unlike auditory problems, deep learning architectures had access to a large repository of labelled imagery datasets.
- As CNN architectures grew deeper and complex , they could be compared against each other , because they were all benchmarked in ILSVRC Challenge.
- Earlier models were based on convolutional layers stacked one after another.
- Szedgedy (2015) made the first departure from a linear architecture by introducing *inception blocks* in *GoogLeNet.*
- He et al (2016) introduced the concept of *residual connection* in their architecture called ResNet.
- Then on, most of the deep learning architectures departed from the linear architecture.



RESIDUAL CONNECTIONS

- Though the winner of the ILSVRC in 2014 was the VGGNET architecture, the deep learning community realized that deeper models do not necessarily mean better performance.
- This happens because of the vanishing-gradient problem , making it harder for weights in the earlier stages of a model to update themselves.
- As He et al pointed out in their seminal paper 2016, the following illustration (left) shows how a 56 layers struggles to achieve the same error rate as a 20-layer model.
- They introduced Residual Blocks (right) , which could combat vanishing gradient, thereby allowing much deeper models.





AUDIO DATASETS

- Piczak et al (2015) noticed that , research on environmental sound classification has been limited due to absence of labelled dataset. This is unlike research on Computer Vision, where there multiple datasets like MNIST, CIFAR and Imagenet.
- The use of the Freesound Project was demonstrated as potential research resource by Font et al in 2013.
- The Freesound Project has a large repository of user uploaded audio samples since 2005.
- Piczak used the Freesound API to build the ESC-10 and ESC-50 Datasets. Salamon presented an even bigger dataset called Urbansound8K in 2018

ESC-10
- 400 Samples of 10 Classes divided into 5 folds.
- Each Class had 40 audio samples
- Each Fold having 8 audio samples

ESC-50
- 2000 Samples of 50 Classes divided into 5 folds.
- Each Class had 40 audio samples
- Each Fold having 8 audio samples

Urbasound8K
- 400 Samples of 10 Classes divided into 5 folds.
- Each Class had 40 audio samples
- Each Fold having 8 audio samples

CNN FOR AUDIO CLASSIFICATION

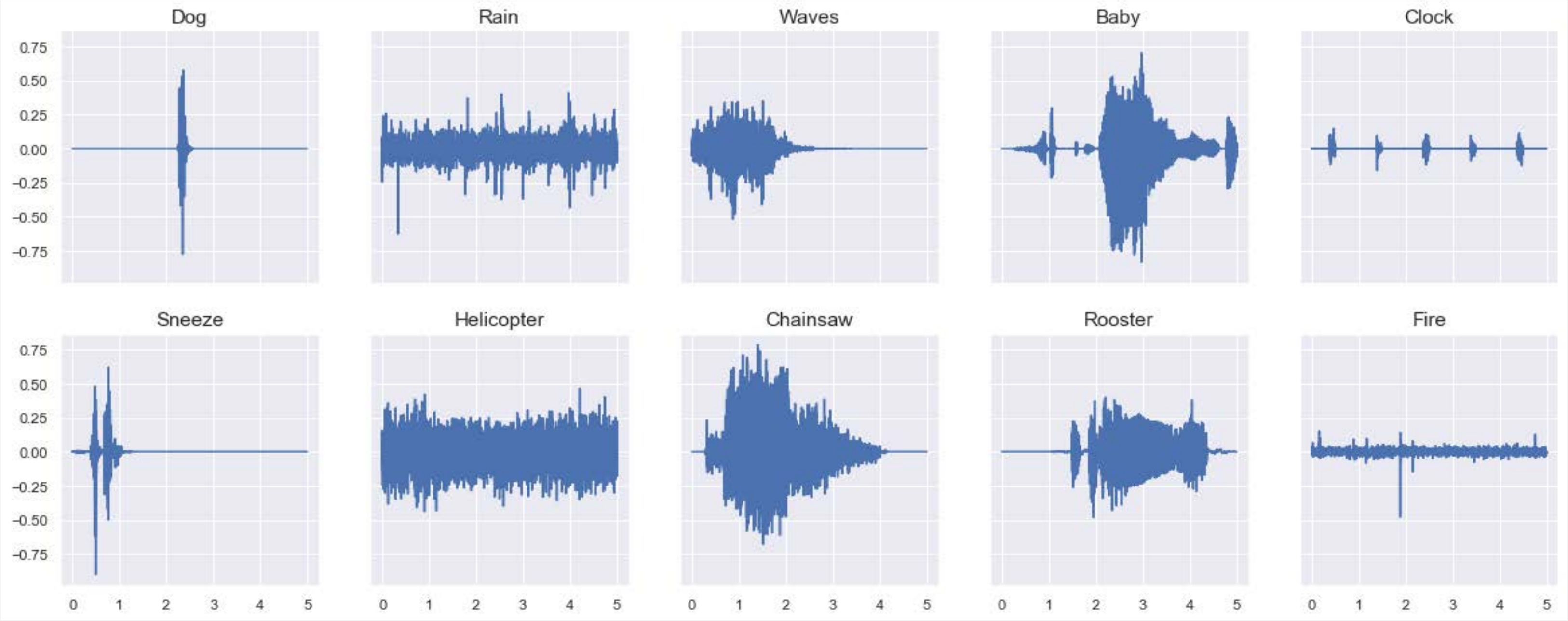
- Prior to 2009, Deep Learning hadn't been used for Auditory problems as reported by Honglak et al.
- They built a CDBN architecture for Speaker , Gender or Phone Classification , but still this wasn't based on ESC dataset.
- We notice CNN architectures applied to ESC datasets from 2015 by Piczak, Tokuzume and Khamaparia in successive years.
- We will use the following as our references and explore the performance of our models with residual connections compared to them.

	ESC-10	ESC-50	Urban 8K	Custom Dataset
Piczak , 2015	85%	77%	65%	-
Tokuzume, 2017	74.10%	-	-	-
Kaustumbh, 2018	-	-	-	85%
Sang, 2018	-	-	79%	-
Khamparia, 2019	77%	49%	-	-

RESEARCH METHODOLOGY

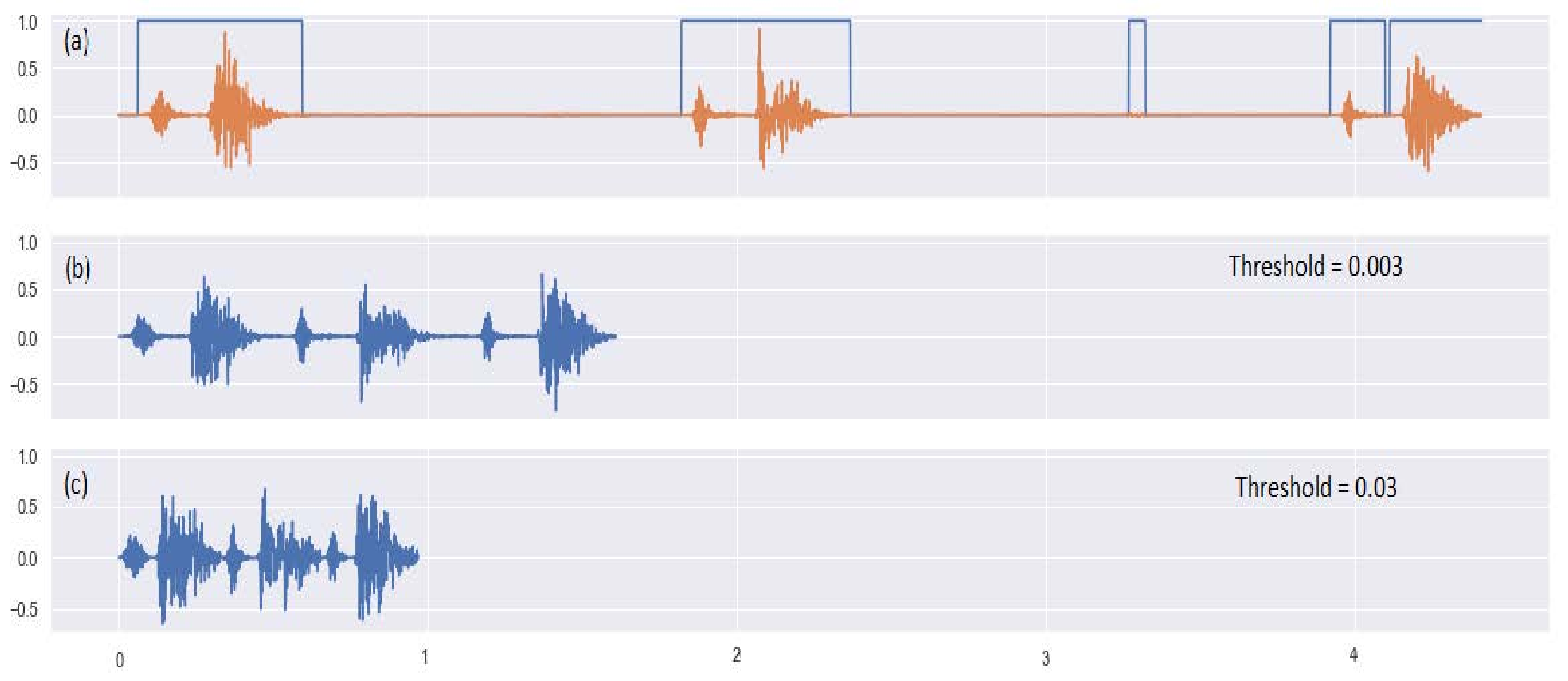
- Pre-processing and Visualization
- Cleaning
- Feature Extraction 1 & 2
- Dataset for Modelling

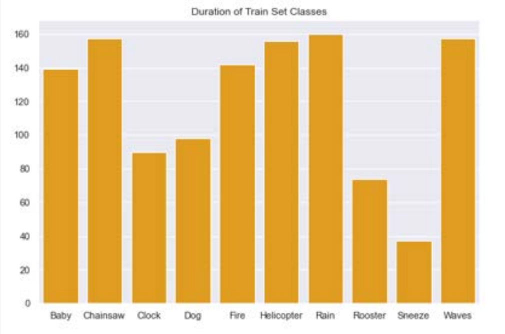
PRE-PROCESSING AND VISUALIZATION

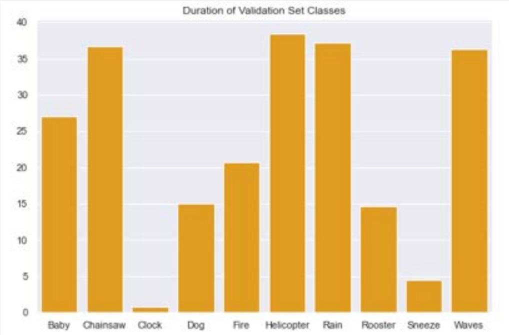


- The ESC-10 Dataset has 400 files with 10 classes. We show here one sample from each class sampled at 44100samples/second.
- We divide the 400 files into 320 and 80 for respectively our training and validation sets.
- Some samples are dominated by silent periods.
- We clean the silent periods from each file and re-write them onto the disk

CLEANING



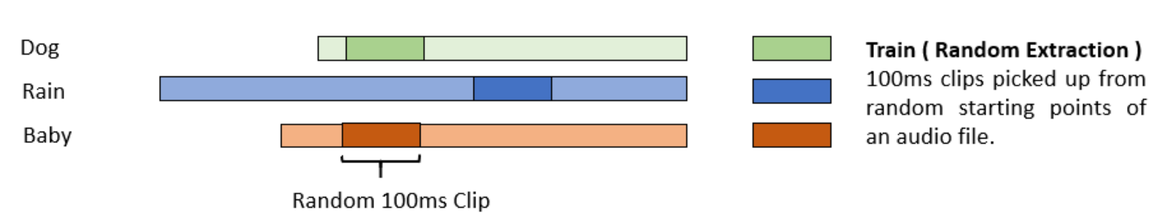




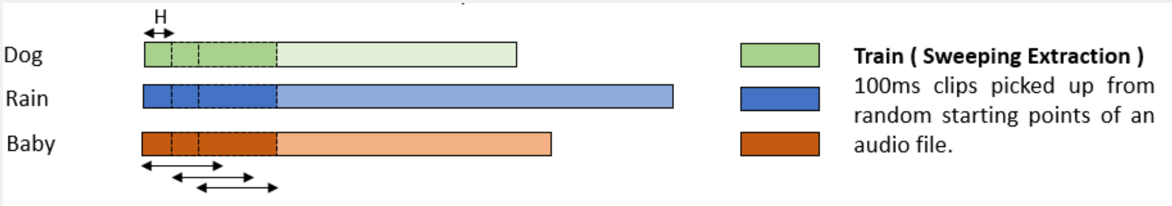
- We eliminate the silent portions of an audio file by taking a rolling average over 100ms.
- We eliminate the portions , wherever the average is less than 0.003 for training and 0.03 for validation set.
- The cleaned files are re-written onto the disk at a rate of 16000 samples/second.
- Note, how we are left with very limited duration of audio for few classes (clock and sneeze) in the validation set

FEATURE EXTRACTION 1/2

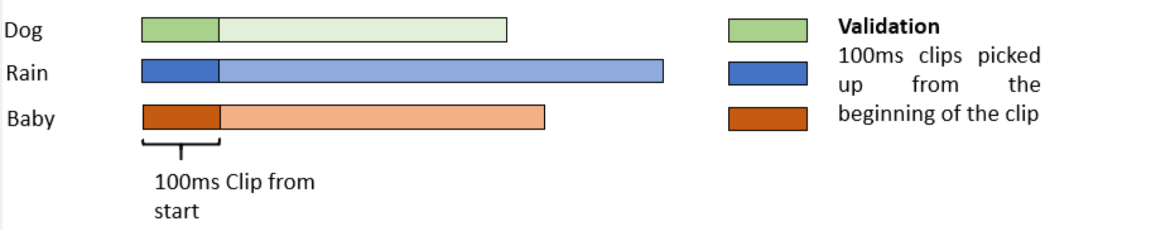
- For building the training set, we extract 100ms clips (1600 samples) at random from anywhere within the audio file. We call this random extraction



- For sweeping extraction , we use a gradually moving window with hop length H

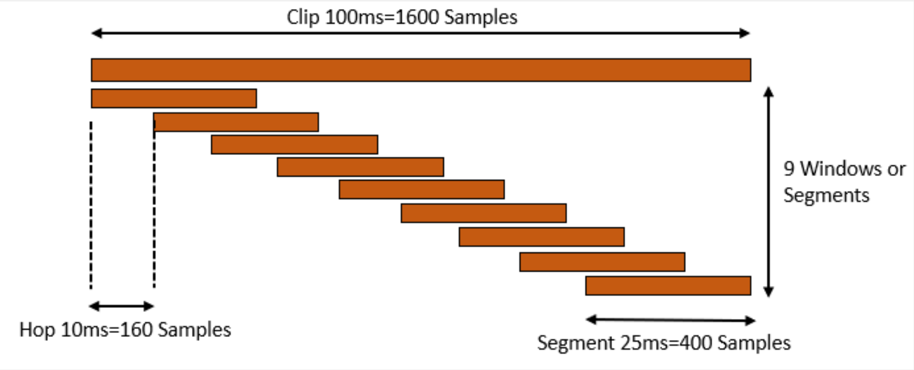


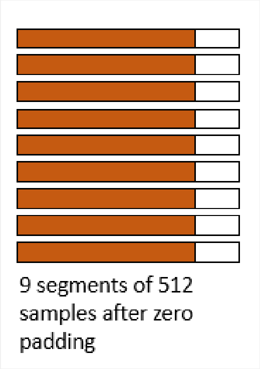
- For validation set however, we always extract the first 100ms as a clip.

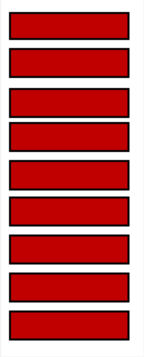


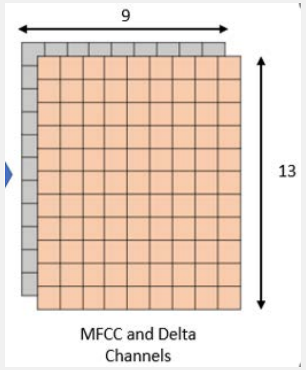
- This process leaves us with a large number of 100ms clips, that we'll use to build our training data.

FEATURE EXTRACTION 2/2









- Make overlapping segments of 25ms (M = 400samples), using a moving window of step size 10ms (L=160Samples).

- This produces 9 segments of 25ms (400 samples) each, which we zero-pad to make the length 512 to calculate FFT.

- The corresponding power spectra is given as $|Xi\ (k)|^2$.

- Each of the 9 power spectra is filtered through a mel-filter bank followed by a discreet cosine transform to generate the MFCC and Delta Features

DATASET FOR MODELLING

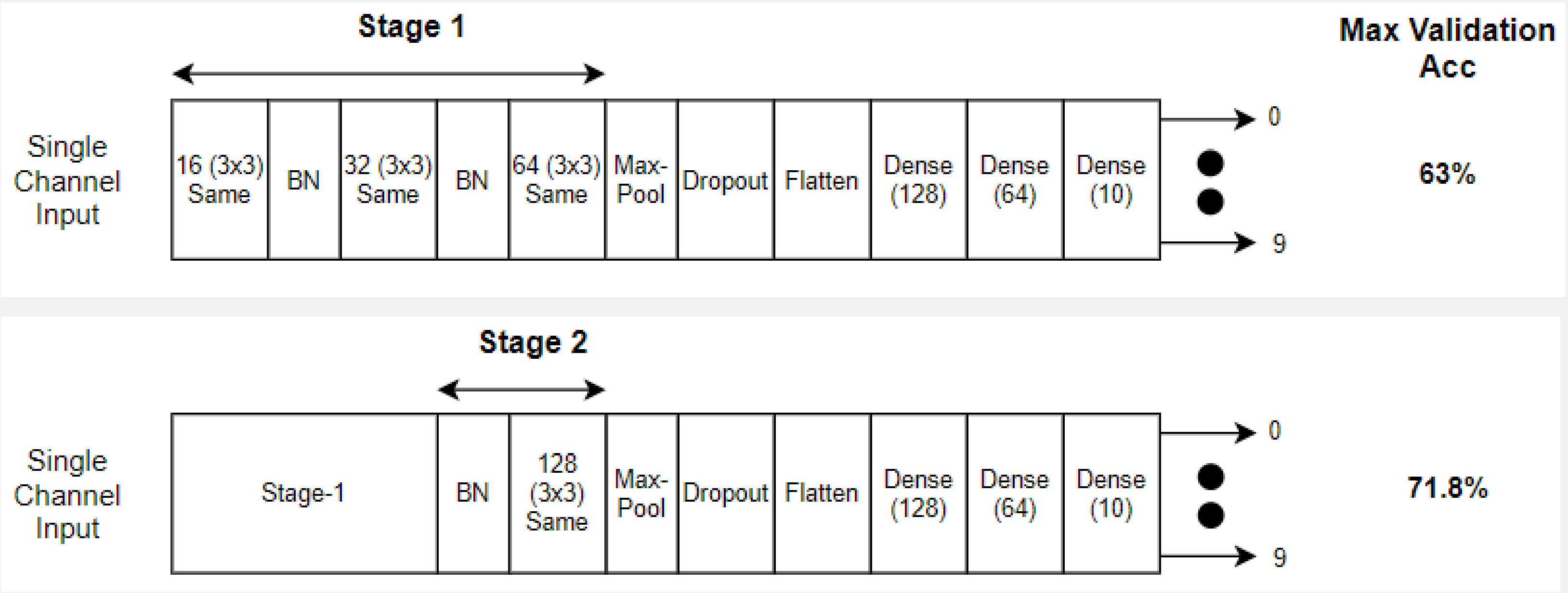
	Training Set Size	Validation Set Size
ESC_10 One Channel Input Random Extraction	24200 x 13 x 9	71 x 13 x 9
ESC_10 Two Channel Input Random Extraction	24200 x 13 x 9 x 2	71 x 13 x 9 x 2
ESC_10 One Channel Input Sweeping Extraction	92274 x 13 x 9	71 x 13 x 9
Usound_1000 Sweeping Extraction	263633 x 13 x 9 x 1	554 x 13 x 9 x 1
Usound_800 Sweeping Extraction	327185 x 13 x 9 x 1	554 x 13 x 9 x 1

- The feature extraction highlighted earlier gives us these datasets that were used to train models.
- For the Urbansound8K dataset, we used the sweeping extraction method exclusively
- During training, we built a generator function to extract random batches from these arrays.

MODELLING

- Baseline Models
- Deeper Models
- Introducing Skip Connections
- Implementing ResNet50

BASELINE MODELS

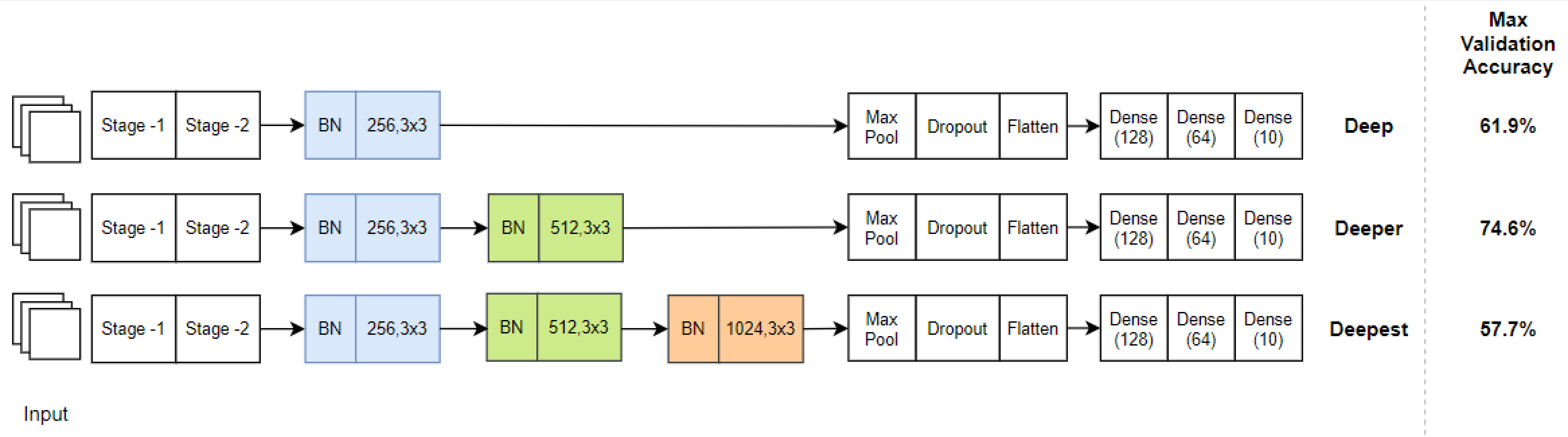


- We built a shallow model with 11 stages and it could reach a maximum validation accuracy of 63%

- When we introduce two additional layers , we reach a validation accuracy of 71.8%.

- Does this mean, adding more layers can generate better accuracies ?

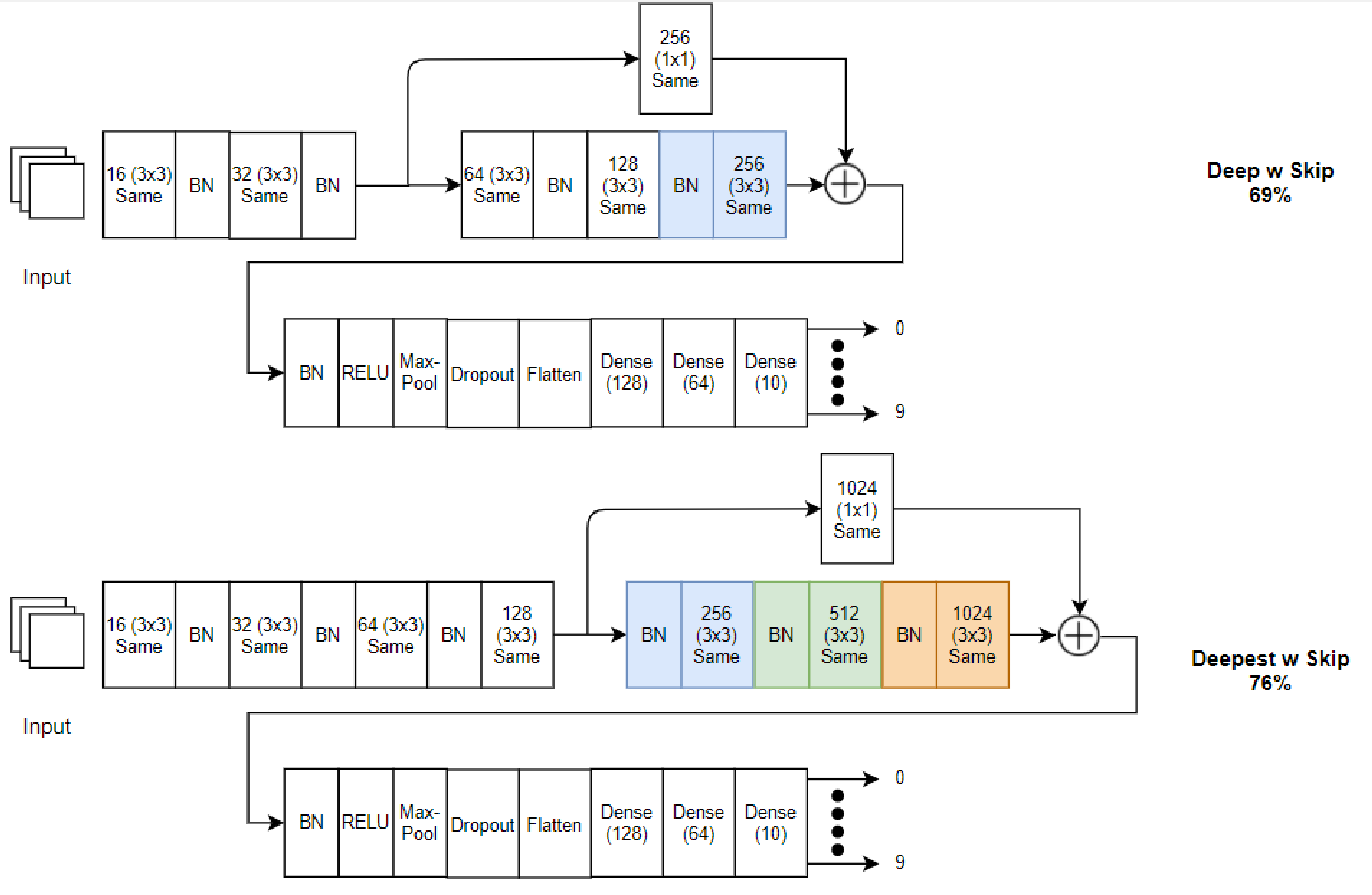
DEEPER MODELS



- As we increase the depth of the models, we notice that the *Deep* and *Deepest* models struggle to keep up with our baseline performance.

- At this point, if we introduce skip connection in the *Deep* and *Deepest* models, can we recover the lost accuracy ?

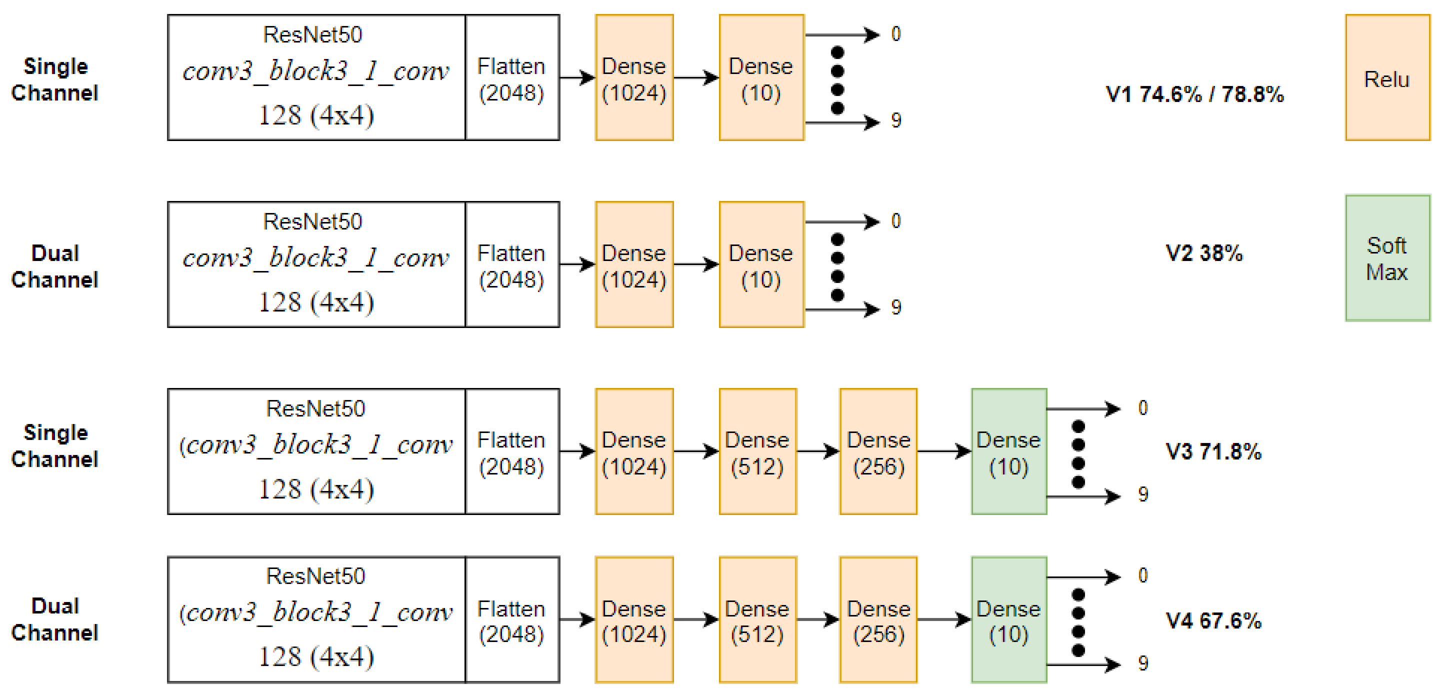
INTRODUCING SKIP CONNECTIONS



- The choice of the position of the residual connection is very important.
- These configurations lead to accuracies that are close to or better than the baseline models.

IMPLEMENTING RESNET ARCHITECTURE

- We used a section of the ResNet50 architecture to see how they perform.
- We only use the first 62 layers of the ResNet50 model and follow it with custom flattening layers.
- The dual channel input causes performance to degrade.
- The single channel input performance are very close to or better than the baseline model.
- Finally , when we try the V1 variant with sweeping extraction input, we are able to exceed 76% (our problem statement)



RESULTS

- Final Results
- Further Studies

FINAL RESULTS

Architecture	ESC -10 Dataset (Single/Dual Channel Input)	Max Validation Accuracy	Train Accuracy at Max Val	Number of Stages	Reference
Shallow Model	Random Single	63.0%	61.2%	11	Piczak - 85% Tokuzume -74.10% Khamparia - 77%
Best Baseline Single		**71.8%**	73.5%	13	
Best Baseline Multi	Random Dual	**66.2%**	75.5%	13	
Deep	Random Single	**61.9%**	58.0%	15	
Deep w Skip		**69.0%**	71.5%	15	
Deeper		74.6%	77.7%	17	
Deepest		**57.7%**	73.9%	19	
Deepest w Skip		**76.0%**	79.9%	19	
Resnet V1	Random Single	**74.6%**	99.4%	62	
Resnet V1 Sweep	Sweep Single	**78.8%**	99.4%	62	
Resnet V2 Multi	Random Dual	38.0%	34.3%	62	
Resnet V3	Random Single	71.8%	96.0%	62	
Resnet V4 Multi	Random Dual	67.6%	97.9%	62	

FURTHER STUDIES

- The impact of residual connection on classification provides motivation for training the variants of Resnet architecture proposed by He et al.
- We were able to reach 78.8% validation accuracy based on the MFCC features alone. Adding a second or third channel with feature selection techniques can provide potentially better results.
- The performance of some of the classes like *Clock* and *Sneeze,* warrant more samples from Freesound to get better results.
- The performance of a reduced version of the Resnet50 also paves the way of other alternatives like ResNext or Inception.
- We did not use any data augmentation techniques in particular. This can be explored with the full ResNet or Inception architectures to get even better results.

THANK YOU