Università degli Studi di Padova

Classification on ESC-50

the role of attention mechanisms in Environmental Sound Classification

Anthony Palmieri, 2038503

Human Data Analytics,
A.A. 2022/2023

Overview

In this presentation you'll find

1. The Dataset
2. Our Contributions
3. Preprocessing
4. CNN raw audio
5. CRNN mel features
6. CNN mel features
7. CNN Ensemble
8. Conclusions

The Dataset

- Audio recognition three main areas:
  a. Music Information Retrieval (MIR)
  b. Automatic Speech Recognition (ASR)
  c. Environment Sound Classification (ESC)
- ESC: categorizing environmental audio signals
- Greater challenges compared to MIR and ASR
  - unstructured nature and
  - lower Signal to Noise Ratio (SNR)

ESC-50

- 2000 audio clips
- 5 seconds long
- 44.1 kHz
- 50 classes
- perfectly balanced
- 5 folds

Our contributions

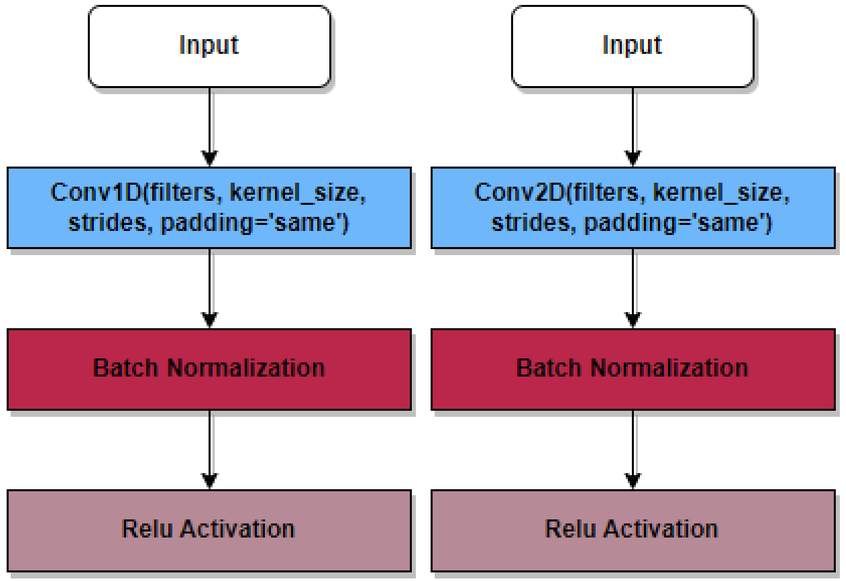
- CNN architecture to handle raw audio. Relatively simple, many pooling layers for dimensionality reduction.
- CRNN model for mel-based features. Initial inception layer followed by a bidirectional GRU layer, enhanced with trainable alignment.
- CNN model that incorporates both attention and skip connections. Skip connection implemented via convex combination with trainable weights.
- CNN ensemble that combines all the aforementioned models.
- Final accuracy of 76% on test set

Preprocessing

- Resampling from 44100 Hz to 22050 Hz
- Audio segmentation
  - 4 overlapping segments
  - 2 seconds long
  - 50% overlap
  - predictions by probability voting scheme
- Mixup Augmentation (discarded)
- Delay Addition and Pitch Shift

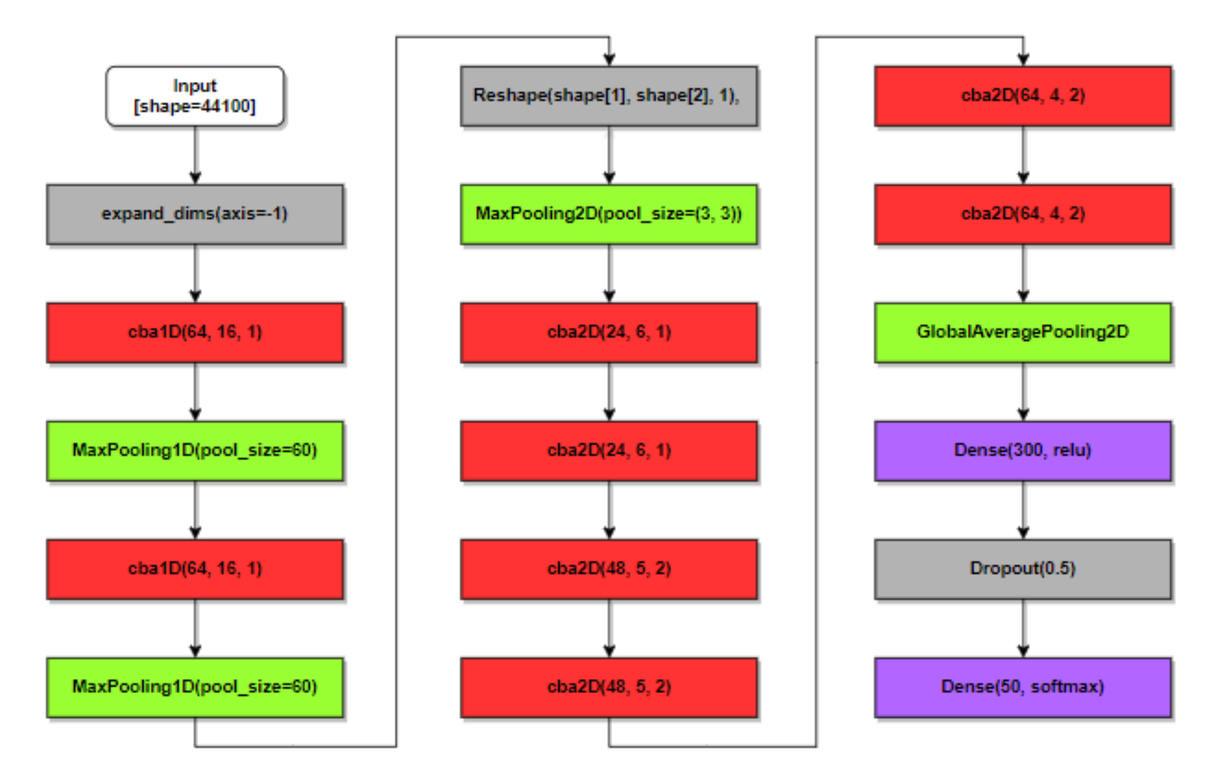
- Raw model
  - input shape (1, 2*22050)
- Mel models:
  - 60 mel-bands with first and second derivatives
  - 20 MFCC with first and second derivatives
  - input shape (60, 87, 4)
- Hold-out approach: 80%, 20%, 20%

CBA blocks



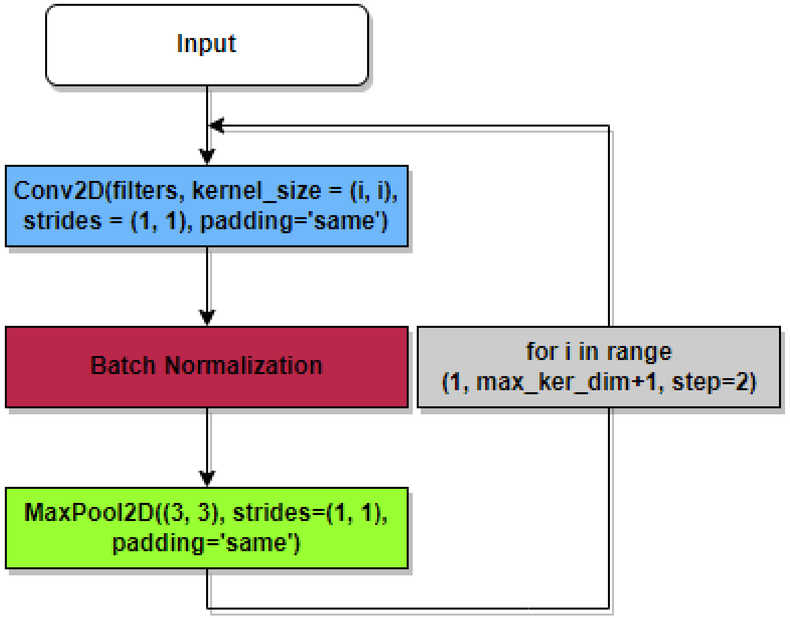
- Building blocks for both raw model and mel models
- Convolution layer extracts meaningful features;
- Batch normalization accounts for the internal covariance shifting, accelerating training and improving performances;
- Relu activation introduces non-linearity

Raw Model



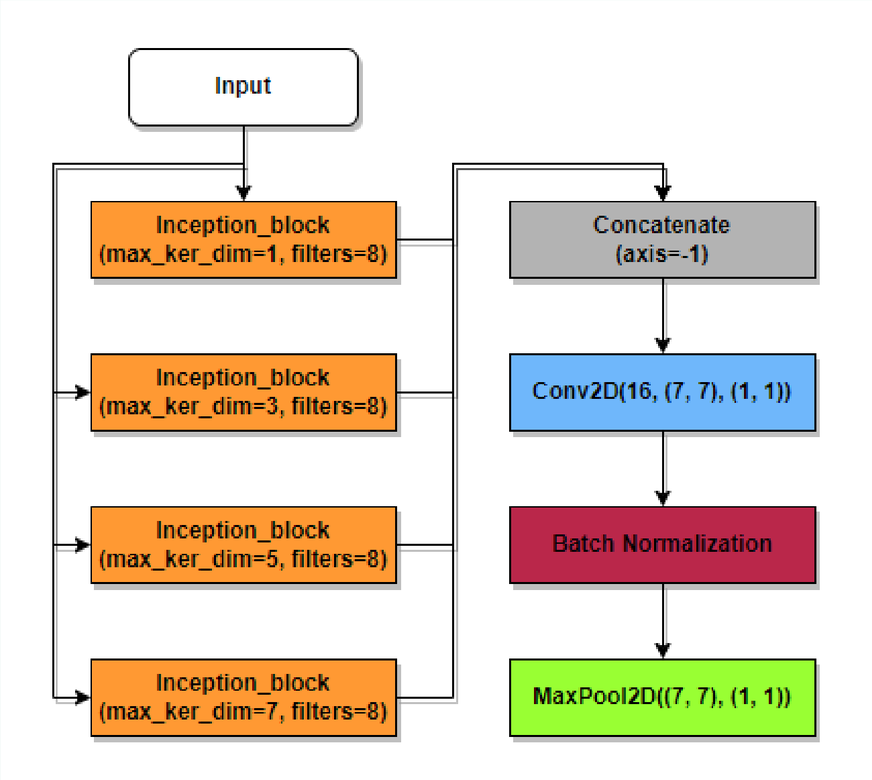
- Combination of 1D and 2D convolutional layers for feature extraction
- Pooling layers for spatial dimensionality reduction
- Dropout to mitigate overfitting
- Output layer with softmax activation produces probabilities for each class
- Final accuracy: 60.5%

Inception Block



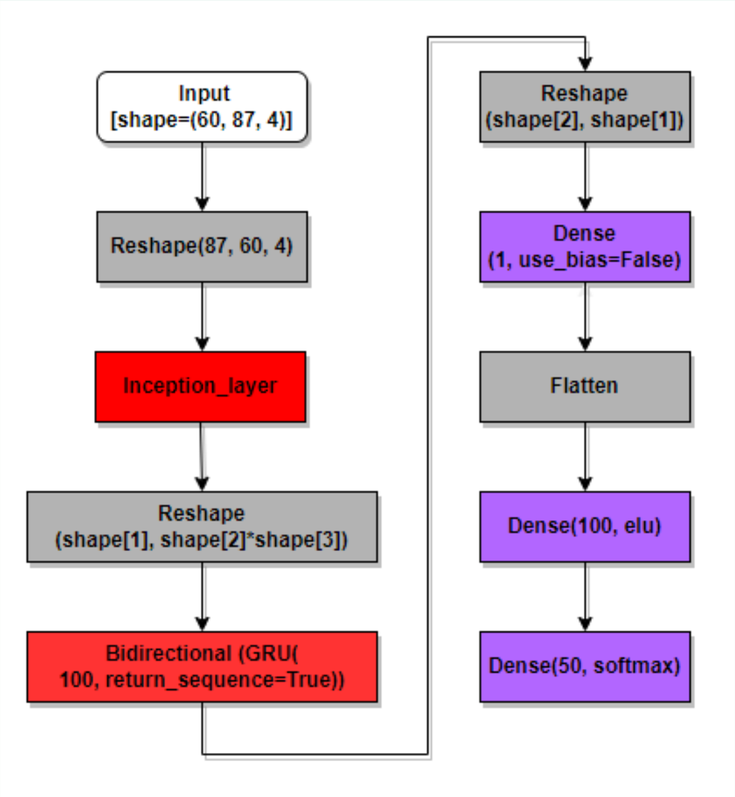
- Building block of the Inception layers
- Sequentially deploy layers with increasing receptive fields
- This allow the models to capture features at different scales and resolutions.
- Conv2D layers extract local features
- BatchNormalization ensure stable and normalized activations
- MaxPool2D layers aggregate local information

Inception Layer



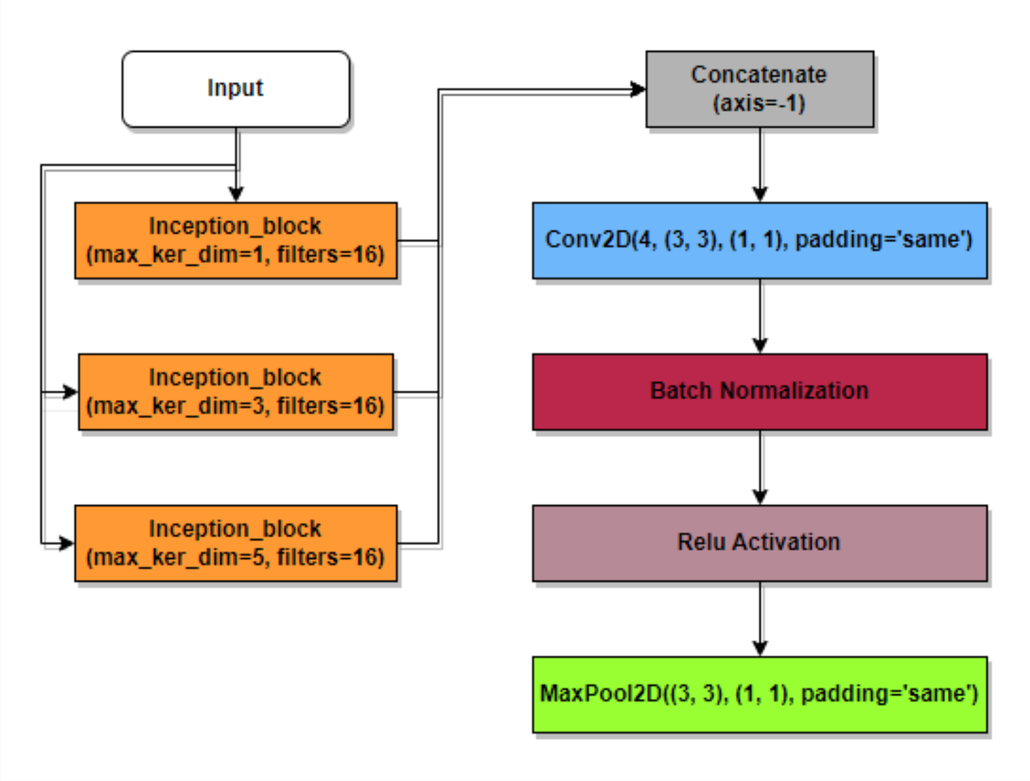
- Obtained by concatenating several inception blocks working in parallel
- Increasing kernel sizes allow to capture information at multiple scales
- Powerful feature extraction capabilities
- More effective than basic Conv2D layers

Mel-Model-1



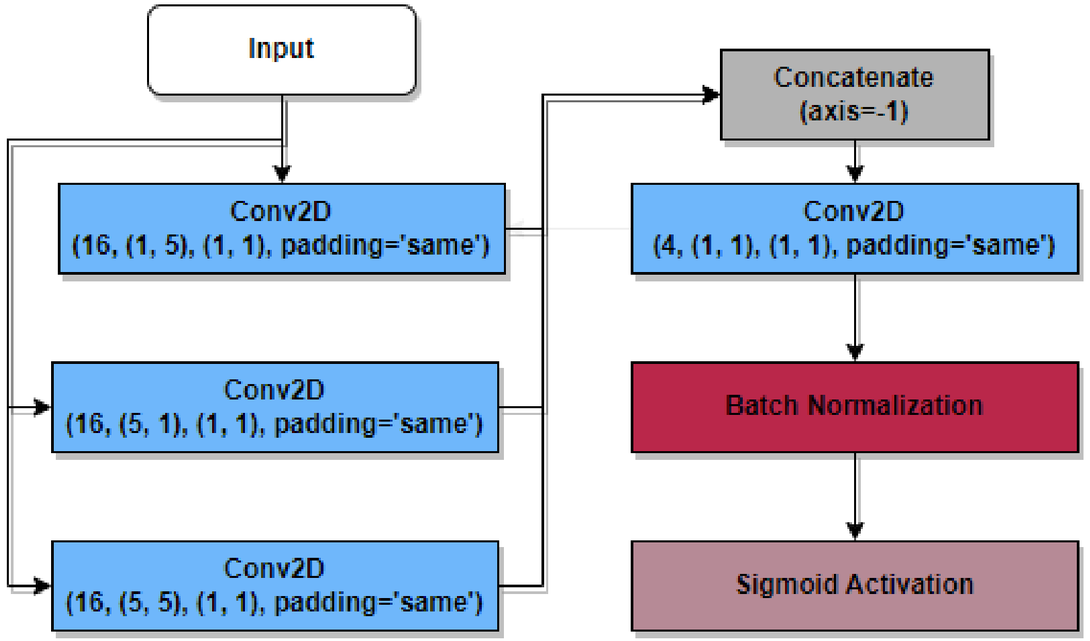
- Inception layer for feature extraction
- RNN to capture temporal dependencies across frames.
- Bidirectional GRU layer to capture dependencies both forward and backward
- Dense layer implements a weighted alignment mechanism
- Weighted alignment helps the network focus on the most relevant frames.
- Final accuracy: 70.5%

Main Block



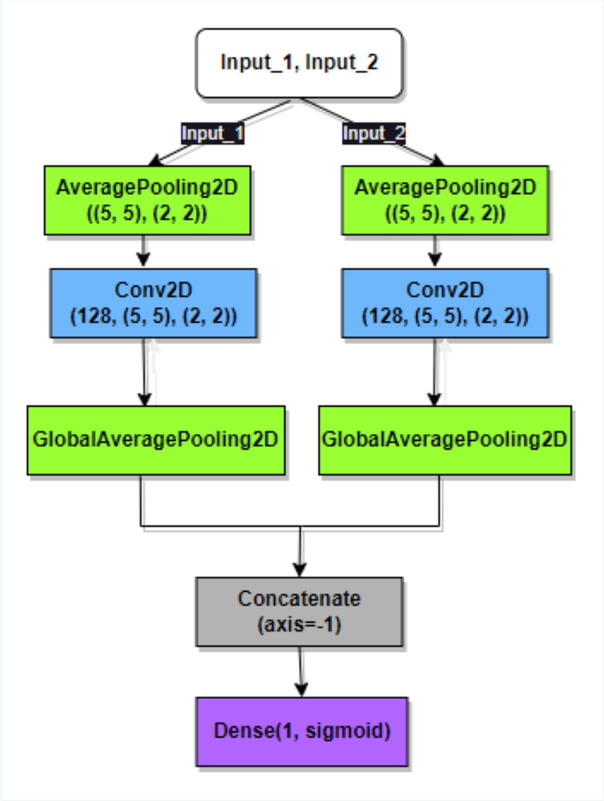
- Second mel-model consists of 3 building blocks:
  - Main
  - Attention
  - sigma_CNN
- Main block concatenates several inception blocks
- It acts as an inception layer for effective feature extraction

Attention Block



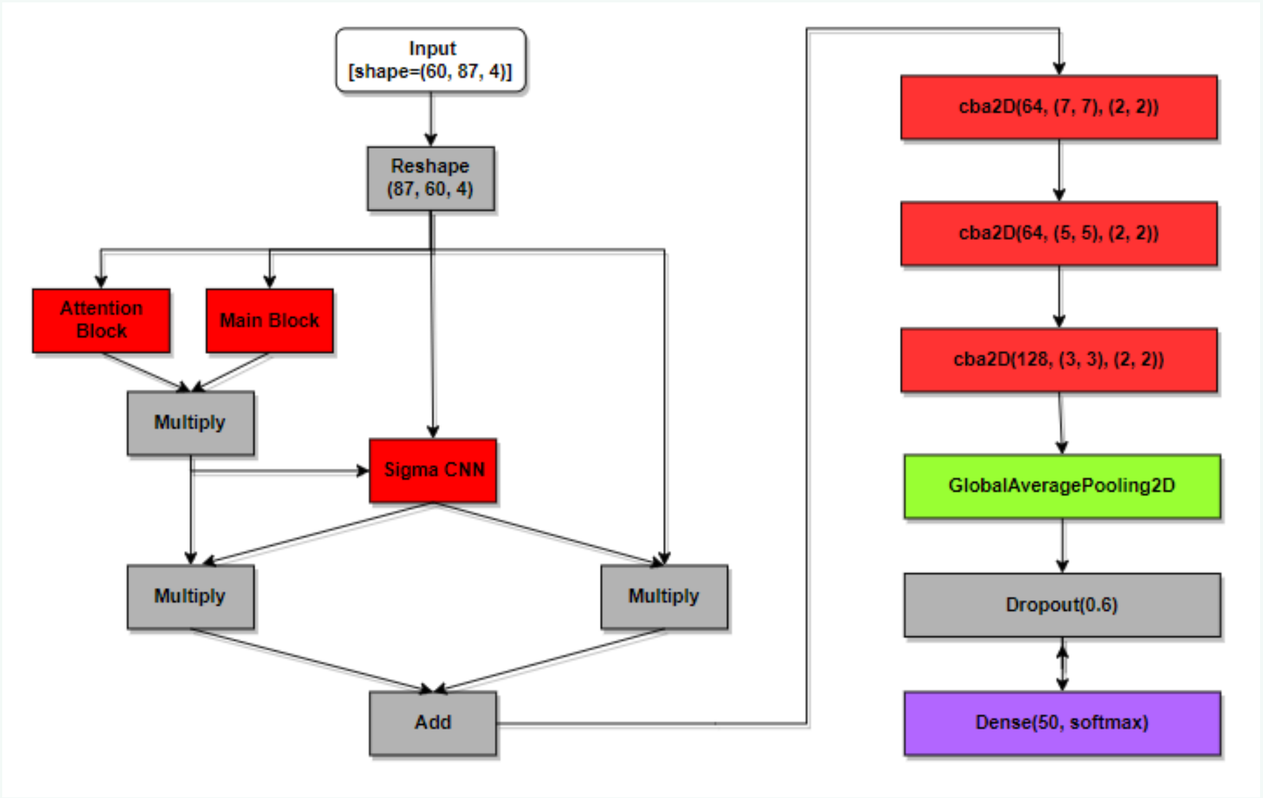
- Various types of attention are computed
- kernel (1, 5) for spatial attention on the frequency domain
- kernel (5, 1) for temporal attention on the time domain
- kernel (5, 5) for time-space attention on the time-frequency domain
- These attention scores are concatenated to weigh the output of Main block

sigma_CNN block



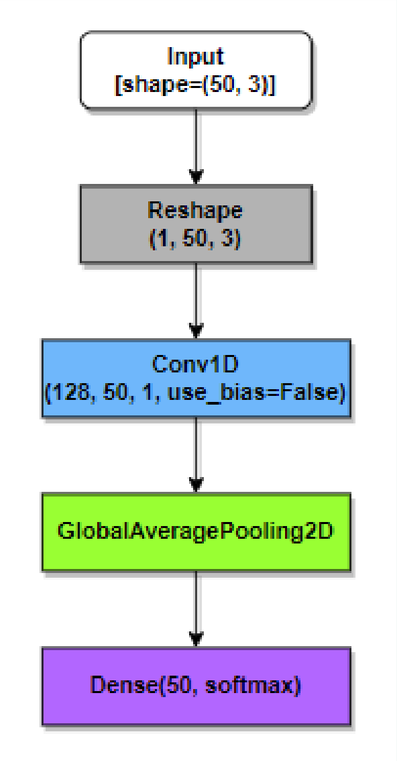
- To control overfitting and avoid gradient problems we introduce skip connection
- Skip connection slows down the fitting too much
- Sigma_CNN block to guide the skip connection and accelerate training
- It takes as input both
  - original input (i.e. X_input) and
  - Main(input)⊙Attention(input) (i.e. X_new)
- Final layer with sigmoid activation computes a convex multiplier σ in [0, 1]
- Skip connection: σ*X_new + (1 - σ)*X_input

Mel-Model-2



- Main block an inception layer for feature extraction
- Attention block to weigh Main block output
- Trainable skip connection to escape suboptimal solutions and accelerate training
- Many CBA blocks for dimensionality reduction
- Final accuracy: 71.5%

Ensemble



- Designed to leverage the predictions from multiple models
- More robust classification
- Architecture very simple to avoid overfitting
- Stride of 50 introduces a trainable weight for each probability score from the three models.
- Flexible weighting mechanism to dynamically adjusting the contribution of each model's output
- Final accuracy: 76%

Conclusions

Final observations

- Increasing filter counts did not improve performances.
- Audio segmentation crucial across all models.
- Delay and pitch shift useful; Segmentation+Mixup not so much.
- Inception layers way better than simple conv layers.
- Trainable alignment in CRNN completely changed fitting, making it extremely faster.
- Skip connections useful to escape suboptimal solutions but slowed convergence.
- Trainable skip connection for accelerated training and improved accuracy.
- Future directions: transfer learning, autoencoders for feature extraction, additional data augmentation like time-stretch and noise injection.

Thanks for the attention!