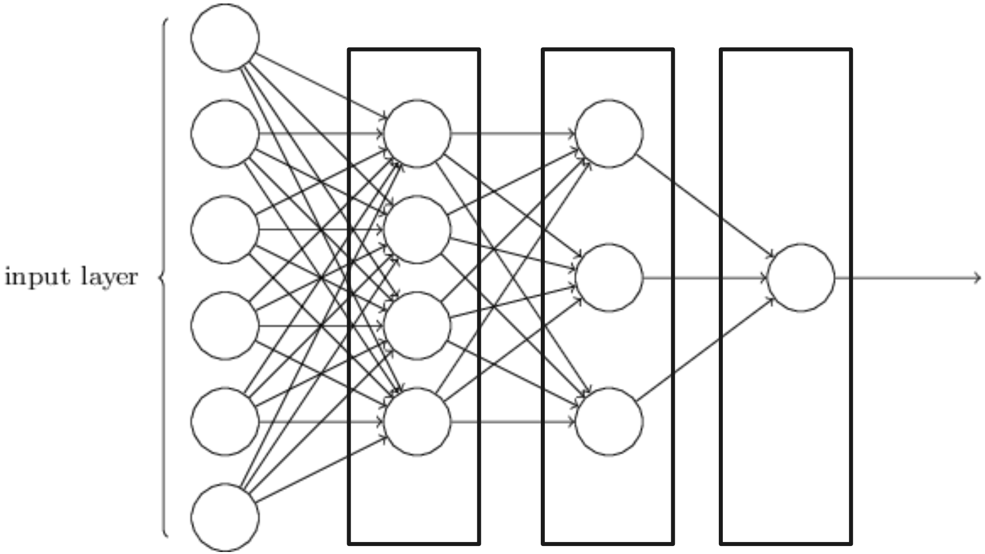
Minutes: Speaker Diarization and Tech Talk

Some of Our Learnings on Transfer Learning

Dense Layer (Fully connected layer)

Each neuron in this layer is connected with every neuron in the last

Most basic form of a neural network

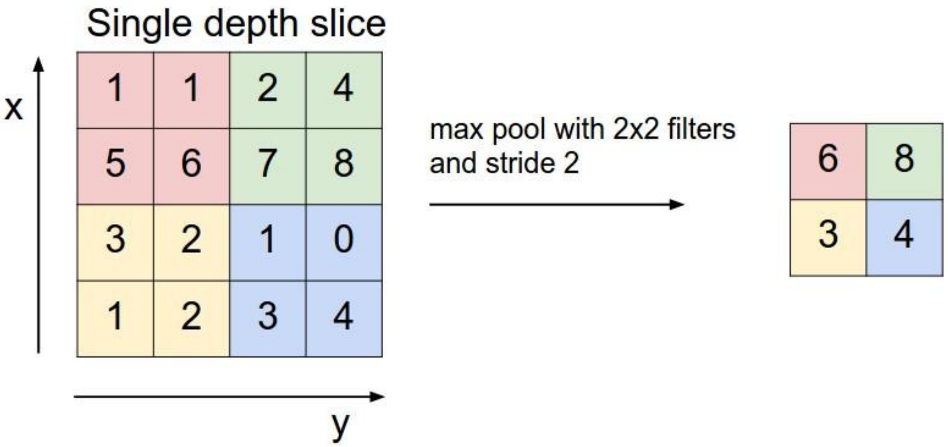


Pooling Layer

Each set of neurons is averaged out to form the new neuron

Reduces complexity

Fast to compute



Single depth slice

1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

max pool with 2x2 filters
and stride 2

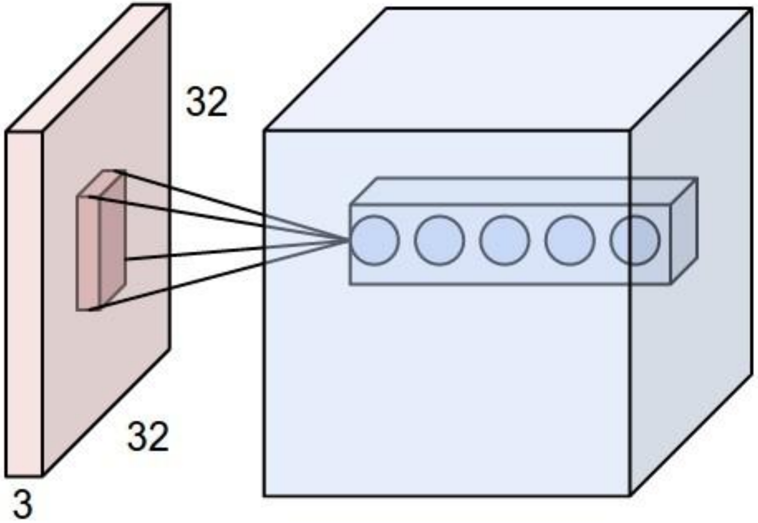
6	8
3	4

Convolution Layer

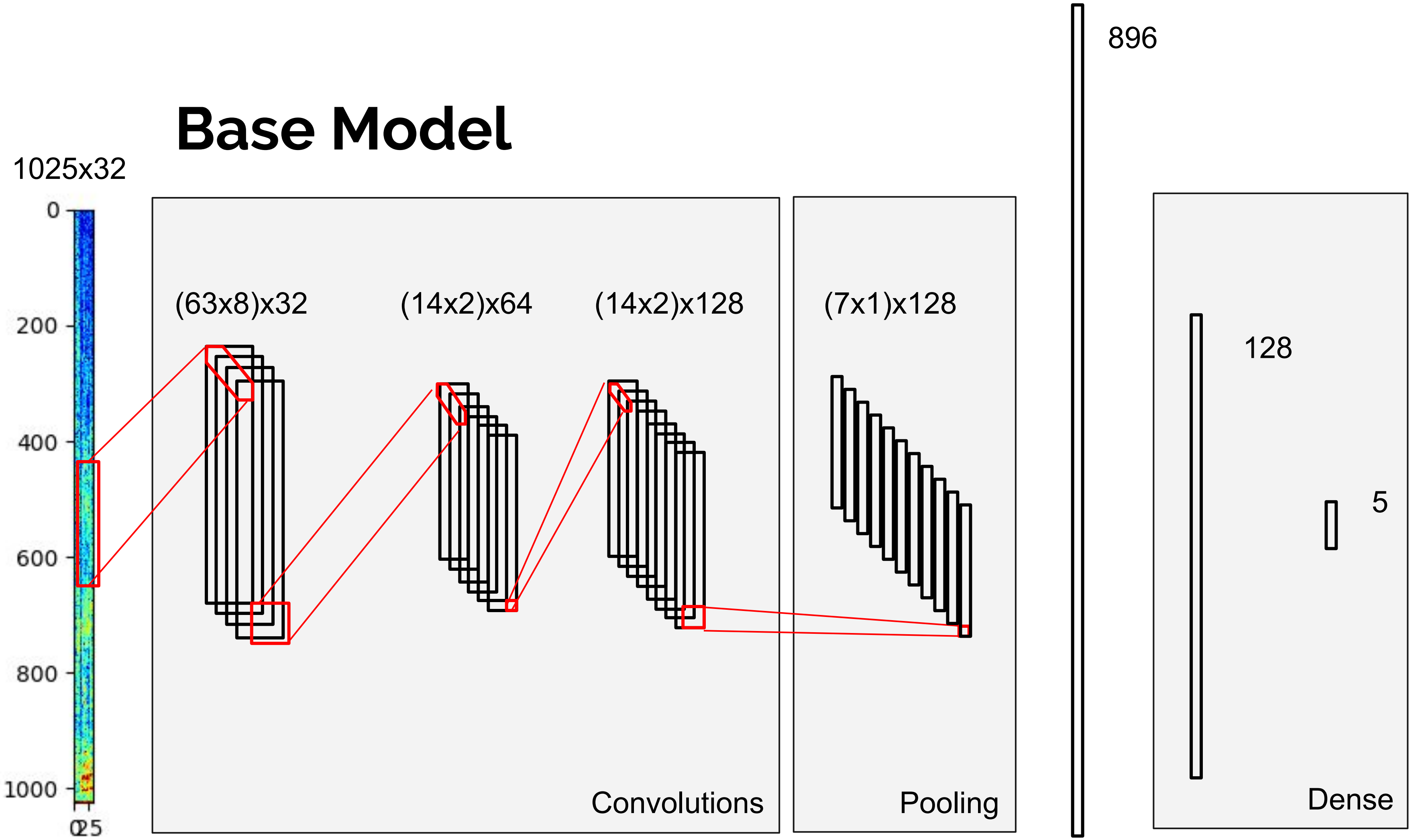
Similar to pooling

Instead of applying "average" onto the neurons, it applies a dense layer.

Good for identifying patterns



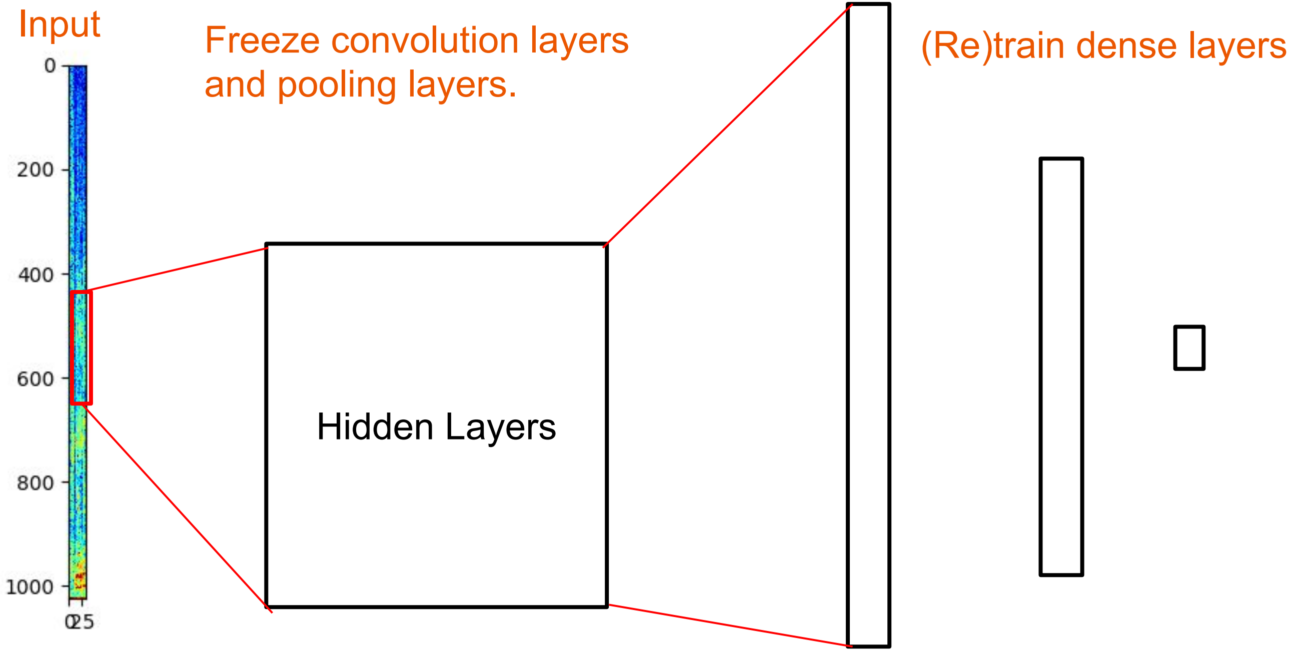
Base Model


1025x32
0
200
400
600
800
1000
0
25
(63x8)x32
(14x2)x64
(14x2)x128
Convolutions
(7x1)x128
Pooling
896
128
5
Dense

Transfer Learning Basics

- Train your model on a big, general dataset, then pop off the last few layers, freeze the early layers, and retrain on a very specific dataset.
- Useful if you don’t have a large enough dataset or you want to train your model faster.
- In our case, we wanted to train on pre-existing corpora of audio data (from the “LibriVox” audiobook archive), and then use transfer learning to “learn” features about new speakers (the users of our API).

Transfer Learning Architecture



Input

Freeze convolution layers and pooling layers.

Hidden Layers

(Re)train dense layers

Important Metrics in Transfer Learning

Base Validation Accuracy

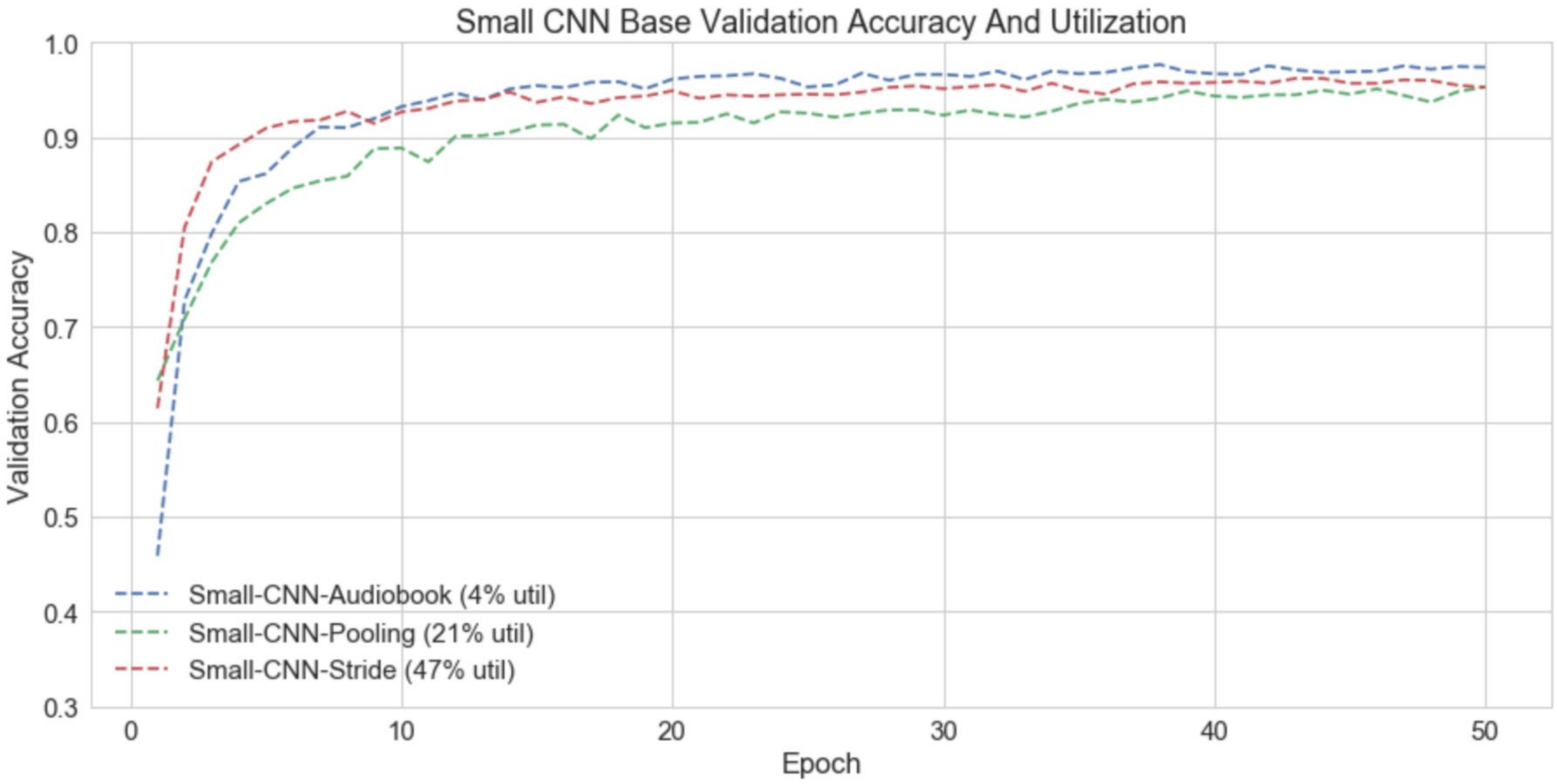
The accuracy of the model when predicting *in-class* on the validation side of the training dataset.

Transfer Validation Accuracy

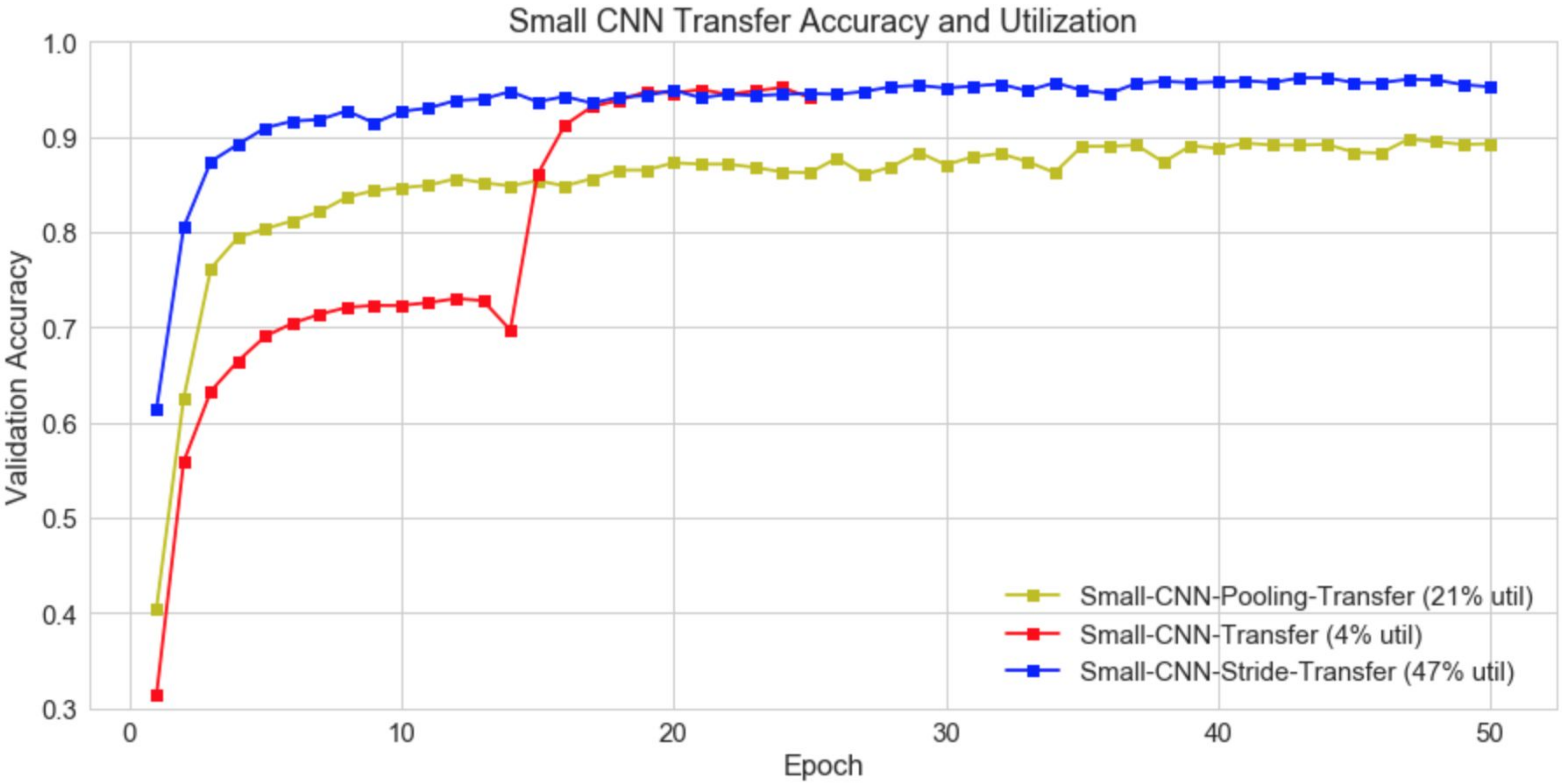
The validation accuracy of the model when predicting *out-of-class* on a new training dataset.

Base Model Utilization

The proportion of the base model *re-used* in generating the transfer model.



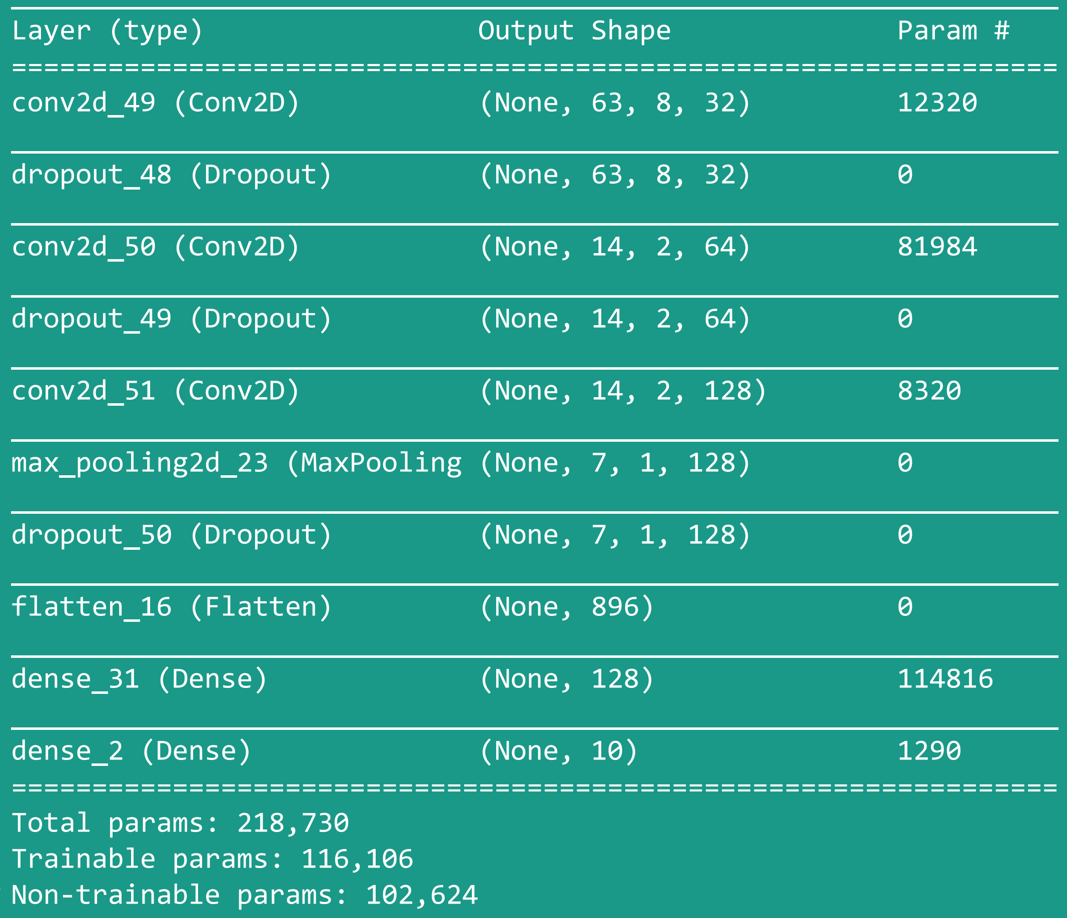
Model Results



Model Results

Results

97.90% Base Validation Accuracy
95.25% Transfer Validation Accuracy
46.91% Base Model Utilization

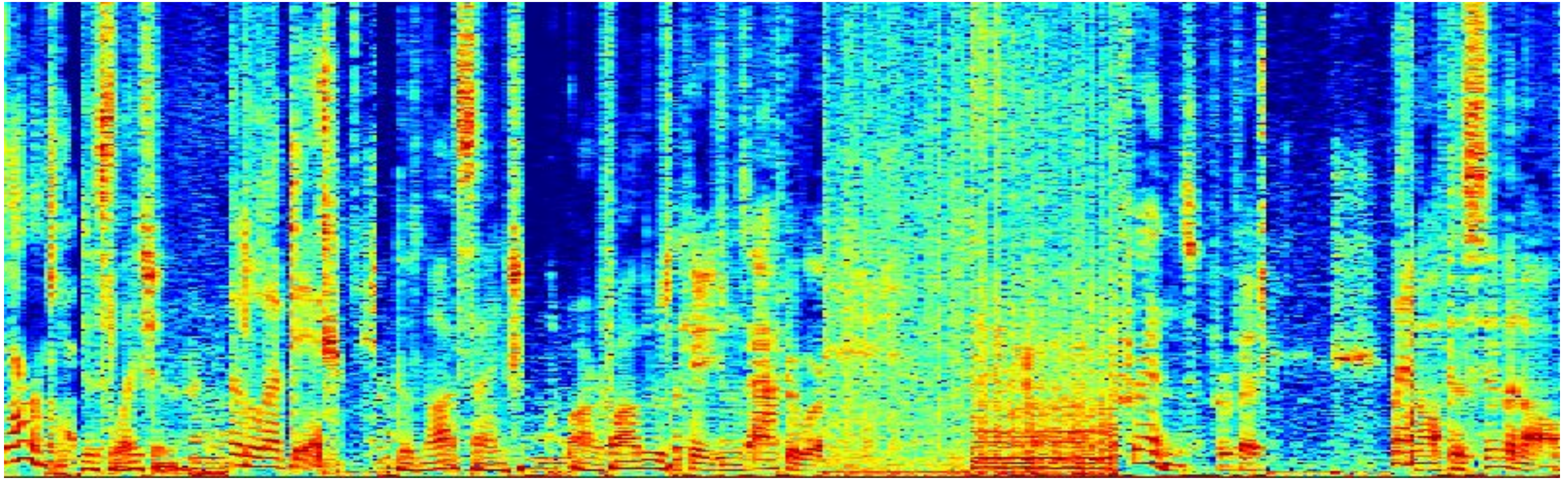


```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_49 (Conv2D)           (None, 63, 8, 32)         12320
_________________________________________________________________
dropout_48 (Dropout)         (None, 63, 8, 32)         0
_________________________________________________________________
conv2d_50 (Conv2D)           (None, 14, 2, 64)         81984
_________________________________________________________________
dropout_49 (Dropout)         (None, 14, 2, 64)         0
_________________________________________________________________
conv2d_51 (Conv2D)           (None, 14, 2, 128)        8320
_________________________________________________________________
max_pooling2d_23 (MaxPooling (None, 7, 1, 128)         0
_________________________________________________________________
dropout_50 (Dropout)         (None, 7, 1, 128)         0
_________________________________________________________________
flatten_16 (Flatten)         (None, 896)               0
_________________________________________________________________
dense_31 (Dense)             (None, 128)               114816
_________________________________________________________________
dense_2 (Dense)              (None, 10)                1290
=================================================================
Total params: 218,730
Trainable params: 116,106
Non-trainable params: 102,624
```

utilization

Dataset

- Librivox Corpus
- Split audiobooks into 1-sec spectrograms



Transfer Learning Summary

🤔 **First iteration:** Retrained **96%** of the model (got 95% accuracy).

😁 **Second Iteration**: Retrained **79%** of the model (got 90% accuracy).

🎉 **Final iteration:** Retrained **53%** of the model (got 95% accuracy)

NOTE: Less retraining means less time to train (and generally less accuracy).