

A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network

Ievgenii Redchyts

Agenda

- Introduction
 - Basics of Neural Network (NN)
 - Common Speech Recognition (SR) with NN
 - Alternative method of SR with NN
- Summary

Introduction

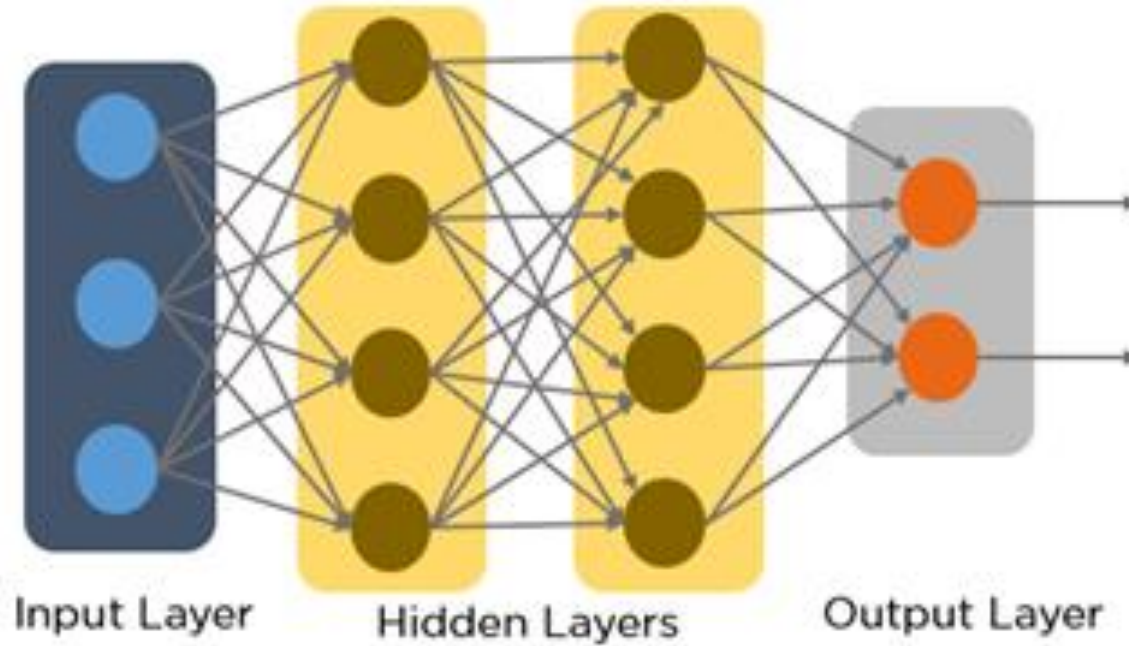
- MarketsandMarkets forecasts the global Artificial Neural Network Market size to grow from USD 117 million in 2019 to USD 296 million by 2024, at a Compound Annual Growth Rate (CAGR) of 20.5% during the forecast period



1

Basic NN

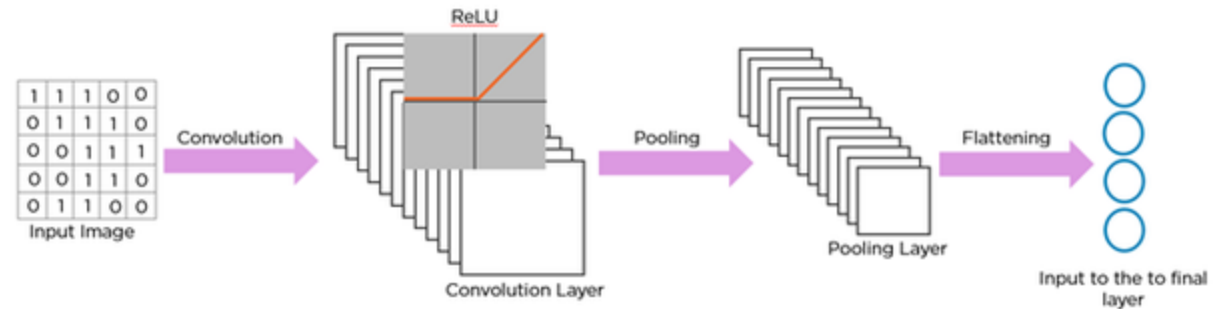
- A typical neural network consist of 3 layers - input layer, hidden layers and output layer



1

Basic NN

- A Convolutional NN (CNN) accepts arrays of pixel values as input to the network



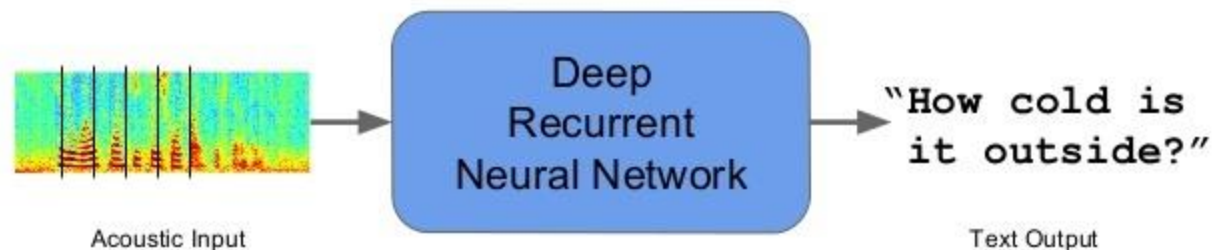
- These are Convolution layer, ReLU layer, Pooling layer and Fully Connected Layer
- **Convolution layer** - uses a filter matrix to obtain a convolved feature map
- **ReLU layer** - introduces non-linearity to the network (1/0)
- **Pooling layer** - reduces the dimensionality of the feature map max(4, 3, 2, 1)
- **Fully Connected Layer** - flattening

2

Common SS with NN

- Traditionally speech recognition models relied on classification algorithms to reach a conclusion about the distribution of possible sounds (phonemes) for a frame

Speech Recognition



Reduced word errors by more than 30%

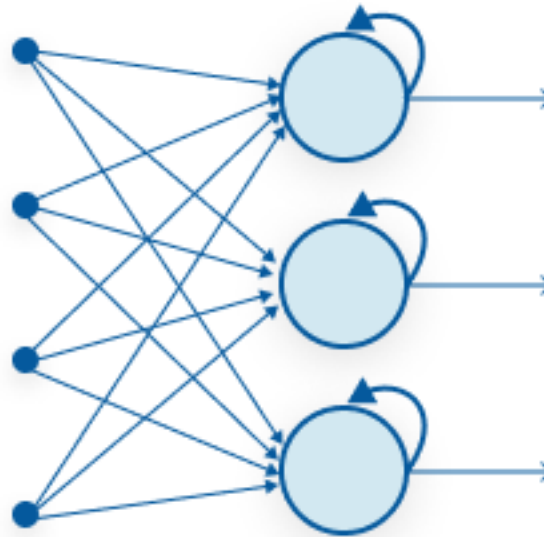
Google Research Blog - August 2012, August 2015

2

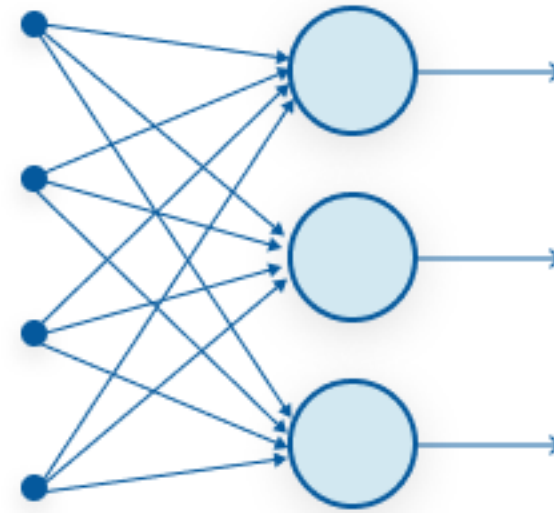
Common SS with NN

- The simplest form of RNN is similar to a regular neural network, only it contains a loop that allows the model to carry forward results from previous neuron layers

Recurrent Neural Network structure



Recurrent Neural Network

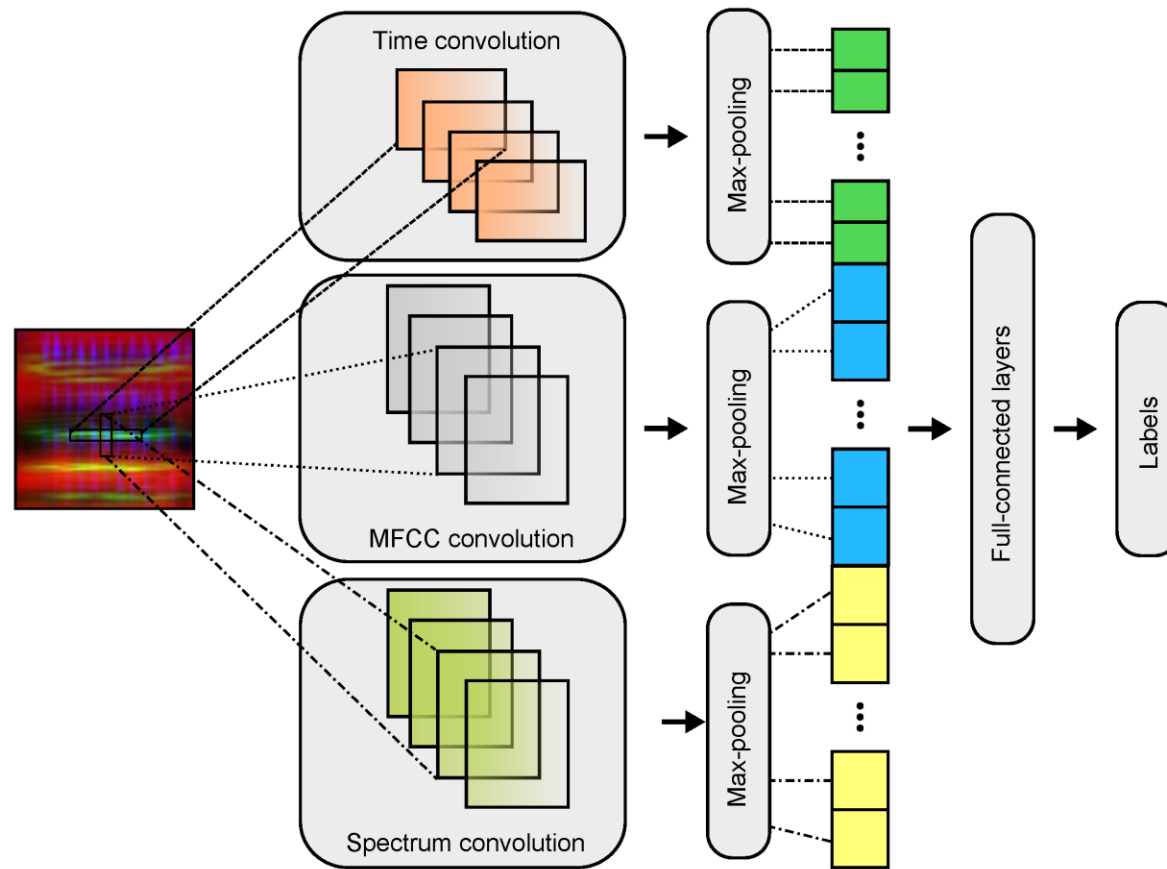


Feed-Forward Neural Network

3

Alternative method of SR with NN

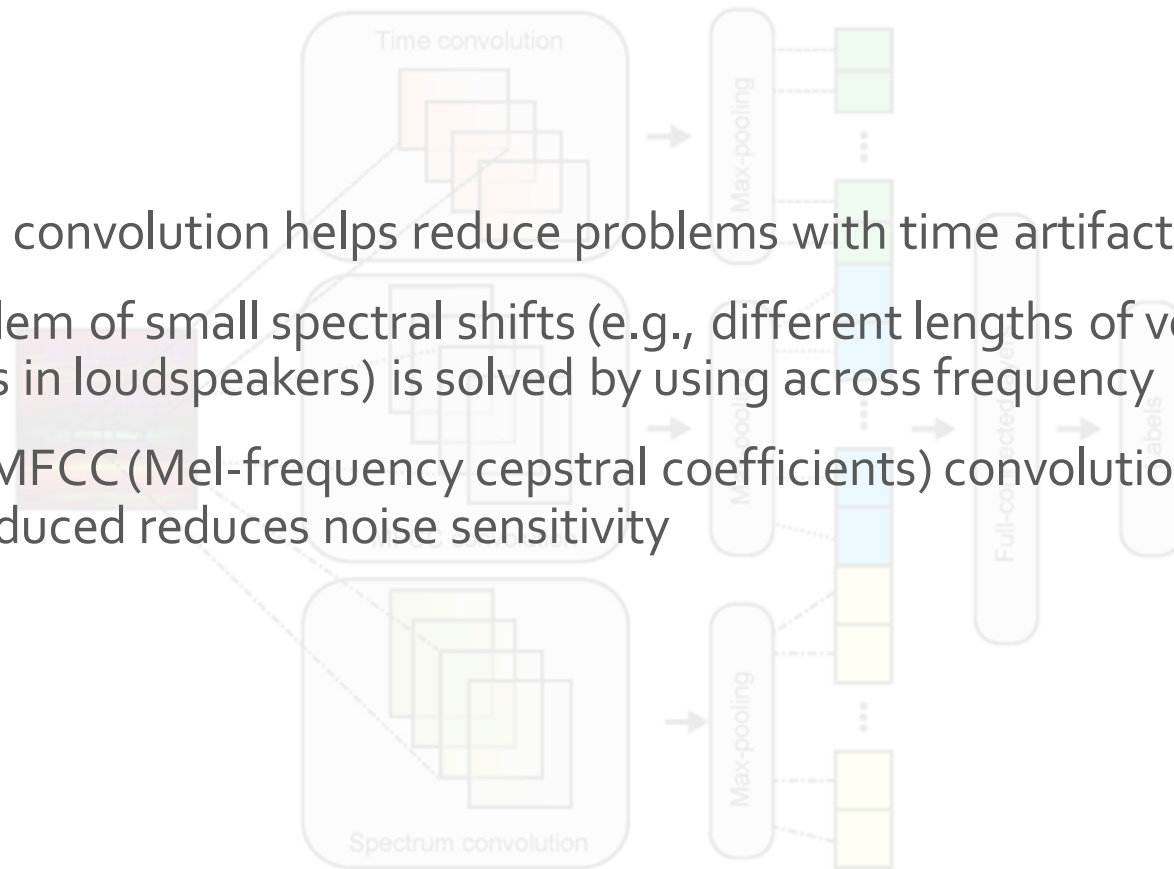
- The method uses sound encoding using images



3

Alternative method of SR with NN

- Time convolution helps reduce problems with time artifacts
- Problem of small spectral shifts (e.g., different lengths of vocal paths in loudspeakers) is solved by using across frequency
- The MFCC (Mel-frequency cepstral coefficients) convolution introduced reduces noise sensitivity



3

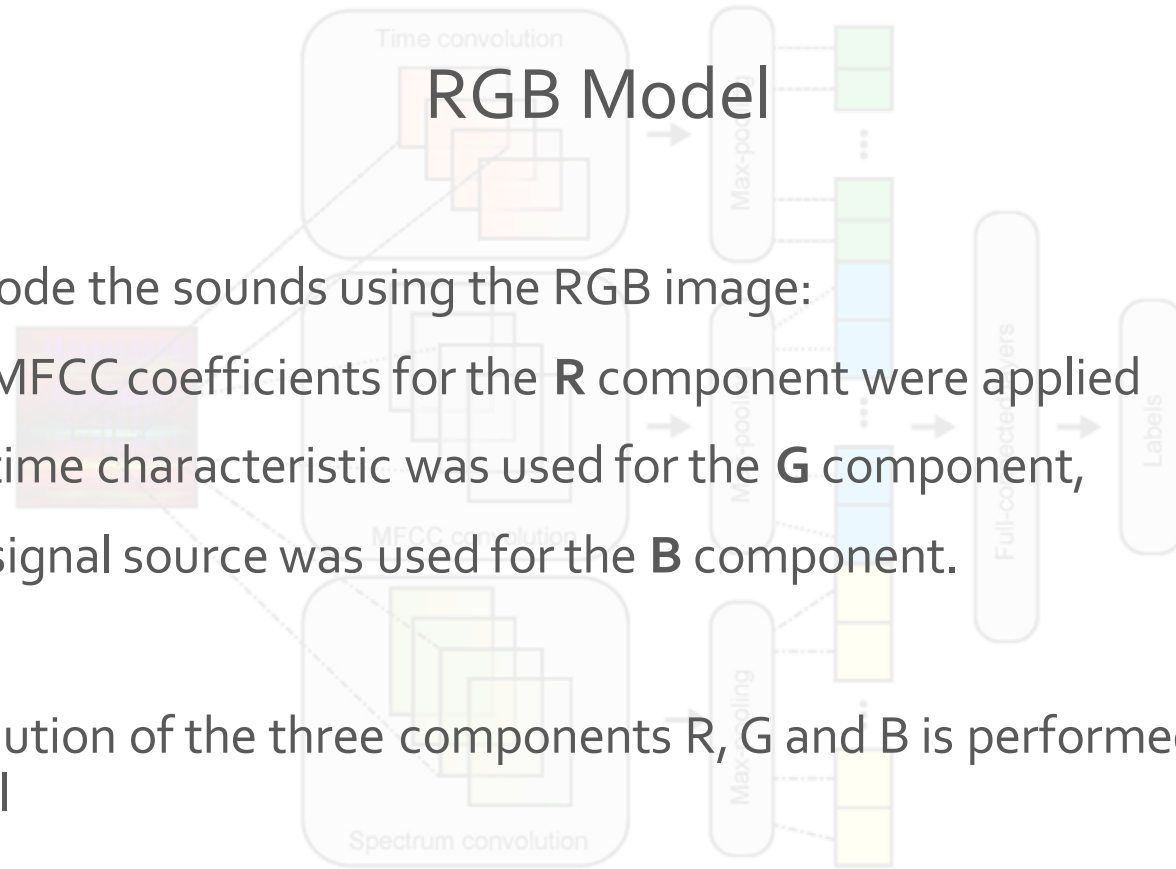
Alternative method of SR with NN

RGB Model

To encode the sounds using the RGB image:

- The MFCC coefficients for the **R** component were applied
- The time characteristic was used for the **G** component,
- The signal source was used for the **B** component.

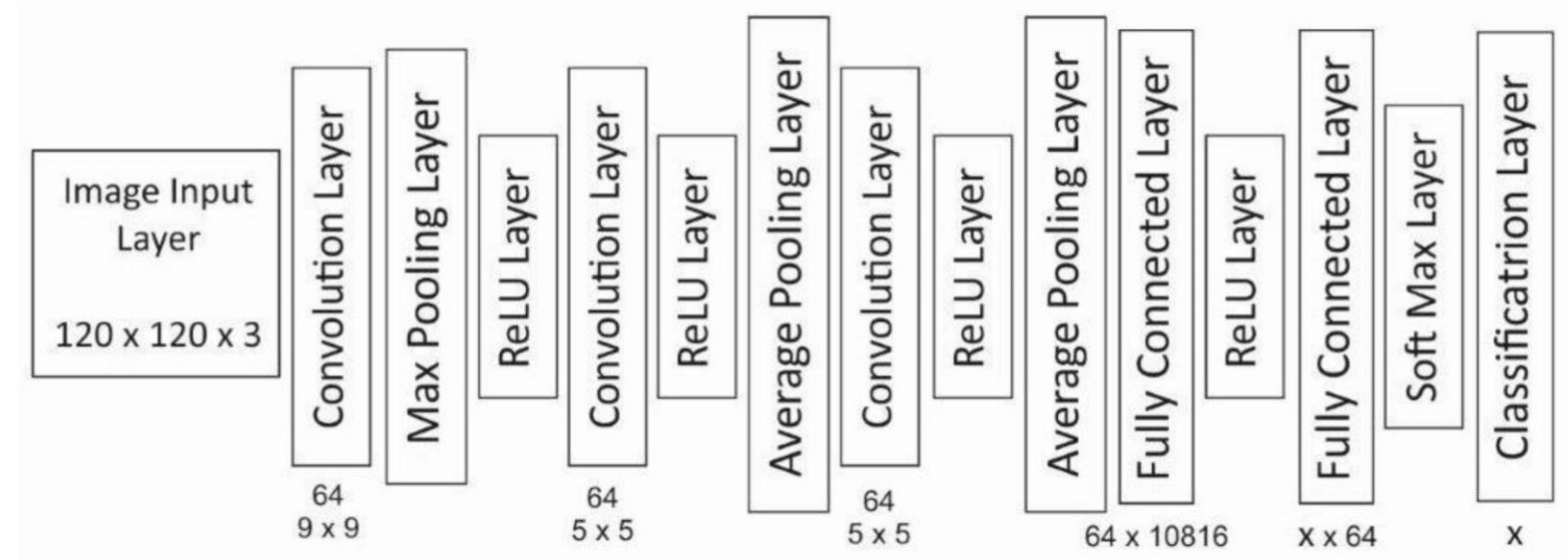
Convolution of the three components R, G and B is performed in parallel



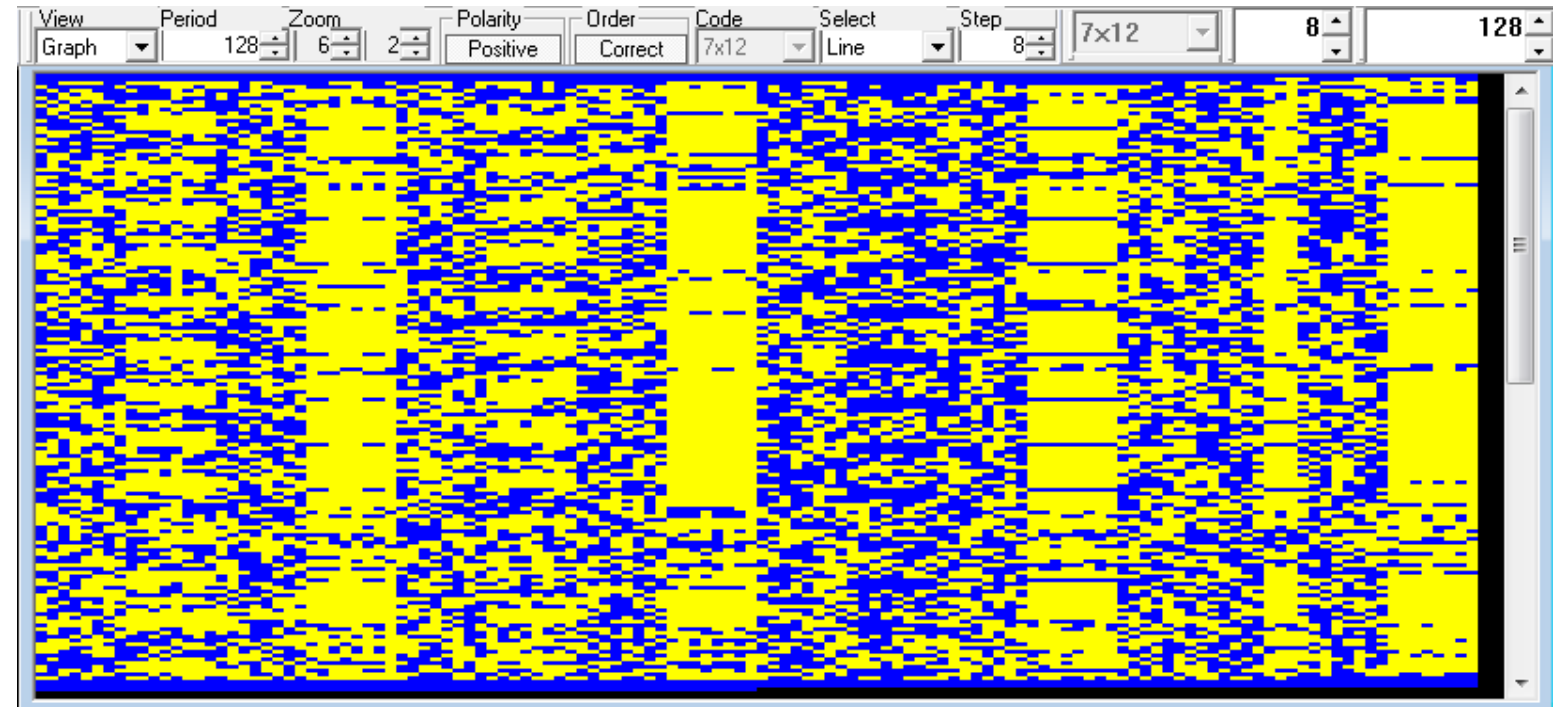
3

Alternative method of SR with NN

Neural Network Structure



Summary



Summary

- Appropriate speech coding by means of images allows use in CNNs
- The proposed method of speech coding is an interesting alternative to the classic approach

Summary

- Thank you!

Sources

- <https://www.mdpi.com/2073-8994/11/9/1185/htm>
- <https://www.marketsandmarkets.com/Market-Reports/artificial-neural-network-market-21937475.html>
- <https://missinglink.ai/guides/tensorflow/tensorflow-speech-recognition-two-quick-tutorials/>
- <https://cs231n.github.io/>
- <https://www.quora.com/Is-there-a-difference-between-neural-networks-and-convolutional-neural-networks>
- <https://missinglink.ai/guides/neural-network-concepts/recurrent-neural-network-glossary-uses-types-basic-structure/>
- <https://www.tensorflow.org/tutorials/images/cnn>
- <http://www.pcz.pl/>