Intelligent Voice®

www.intelligentvoice.com

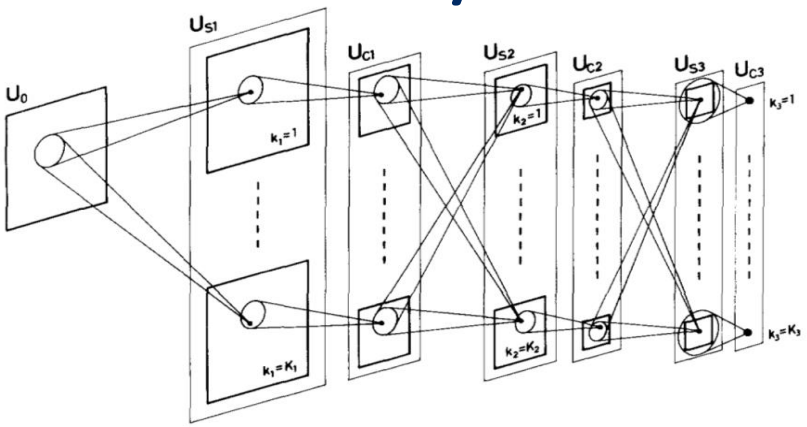
Deep Convolution Neural Networks for Dialect Classification of Spectrogram Images

Nigel Cannings

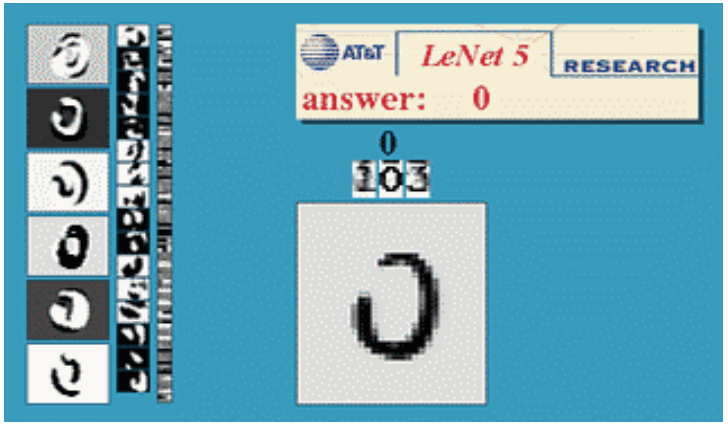
Chase Information Technology Services Limited

Convolution Networks: Brief History

- Inspired from receptive fields in the visual cortex
- Notable Implementations:
  - Fukushima's NeoCognitron (1980)
  - Explicit parallel implementations (1988)
  - LeCun's LeNet-5 (1998)
  - Ciresan's GPU Implementation (2011)
  - GoogLeNet (2014)



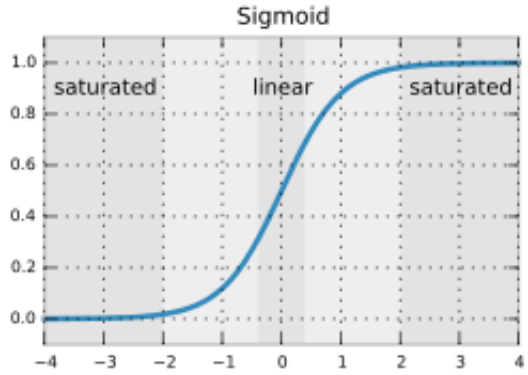
Fukushima, Kunihiko, 'Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position,' Biological Cybernetics **36** (4): 193-202, 1980

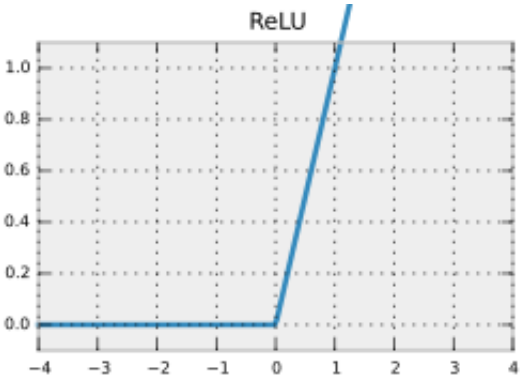


LeNet 5 (1998), image source: http://yann.lecun.com/exdb/lenet/

Deep Learning

- Sigmoidal activation functions have now been largely replaced with rectified linear units (ReLU)
- 'Vanishing error' problem (Hochreiter, 1991) doesn't exist with ReLU
- Now we can do `deep' learning i.e. networks with more than 2 hidden layers
- This discovery and GPU computing has resulted in much recent activity in the Neural Network community



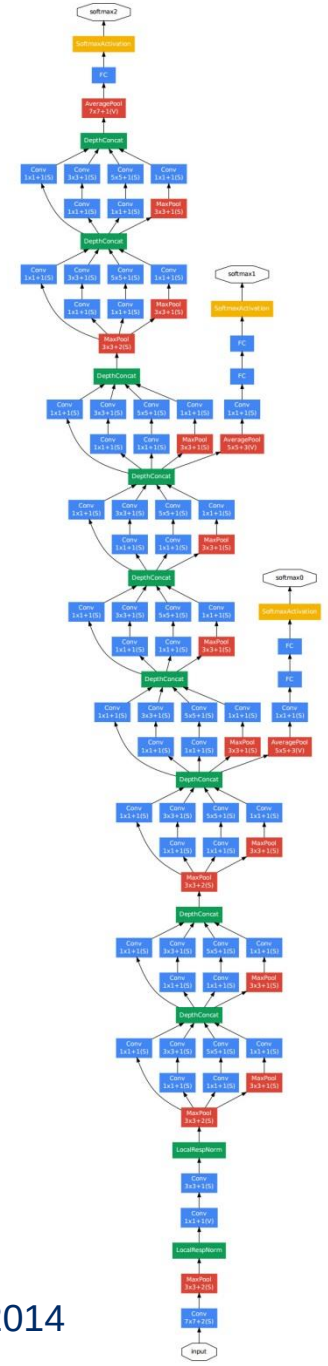


Intelligent Voice®

www.intelligentvoice.com

GoogLeNet

- State of the Art winner of the ImageNet 2014 competition: classifying 1.2M images into 1K classes
- Convolution neural network inspired by LeCun's LeNet-5
- Has 9 'Inception' modules, multiple convolution sizes, and pooling in each module
- Stochastic Gradient Descent used to train the network with 'dropout' which helps prevents overfitting



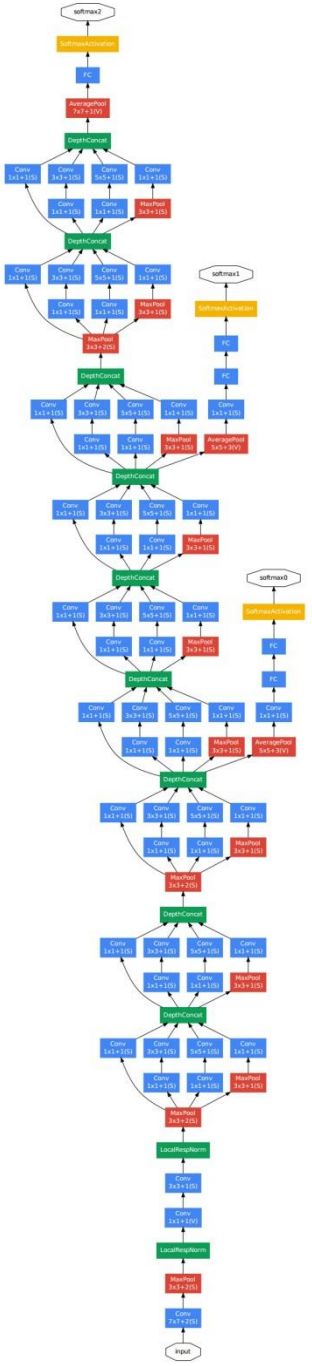
Szegedy, 'Going deeper with convolutions,' arXiv, 2014

GoogLeNet Structure

Topology consists of 'Inception' modules consisting of:

- Convolutions – Filters for extracting features, filter size tends to be small in the early layers, bigger in later layers
- Pooling – dimensionality reduction
- Softmax loss for predicting classes at 3 progressive stages of the network
- Other – concatenations for combining convolutions

'Rinse and Repeat' 9 times

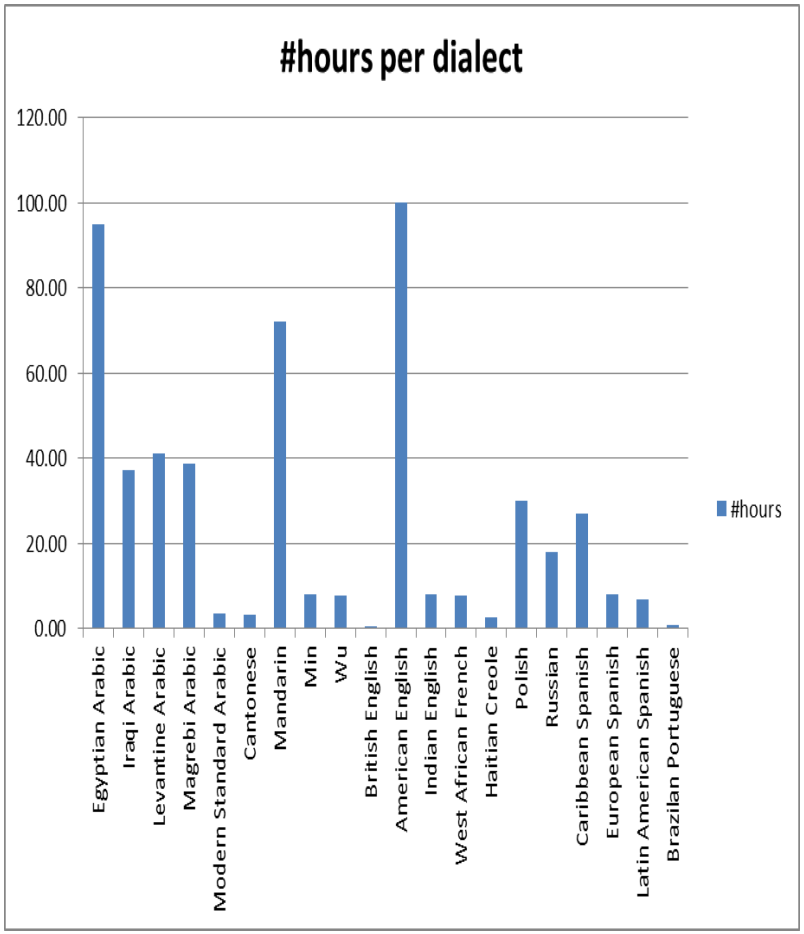


Intelligent Voice®

www.intelligentvoice.com

NIST LRE Competition

- 6 Language clusters, 20 dialects:
  - **Arabic** (Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard)
  - **Chinese** (Cantonese, Mandarin, Min, Wu)
  - **English** (British, General American, Indian)
  - **French** (West African, Haitian Creole)
  - **Iberian** (Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese)
  - **Slavic** (Polish, Russian)
- 500+ hours of speech data
- Data set very unbalanced



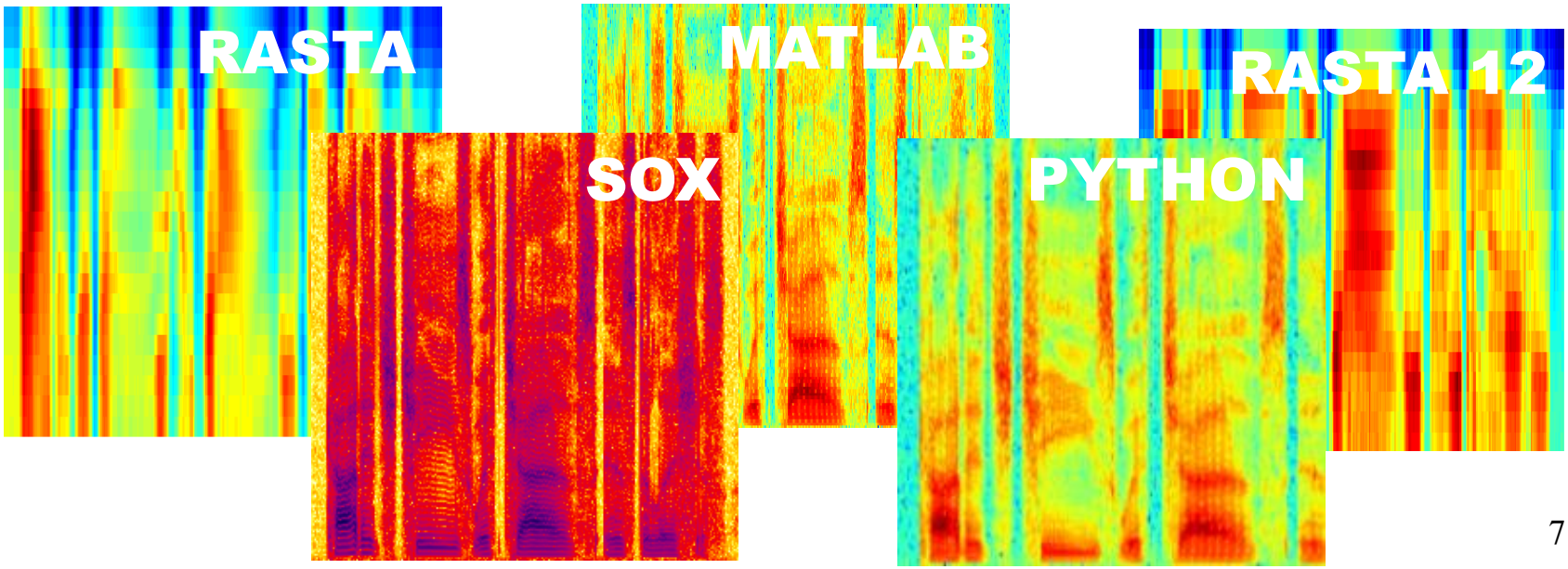
2015 NIST Language Recognition Evaluation, http://www.nist.gov/itl/iad/lre15.cfm

Intelligent Voice®

www.intelligentvoice.com

Spectrogram Convolution Network

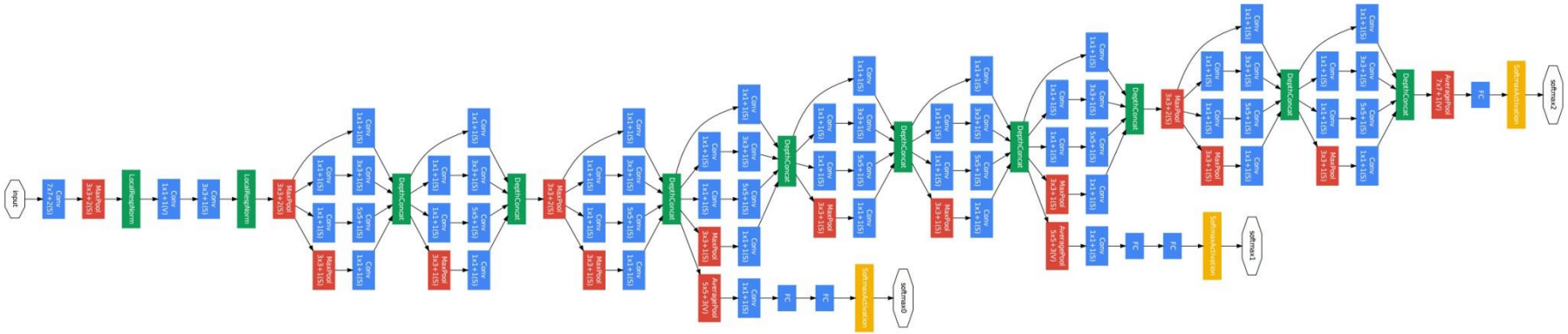
- Based on Nvidia's Digits implementation of GoogLeNet
- Converted speech to 256x256 pixel spectrograms
- Tried different spectral representations and coding…



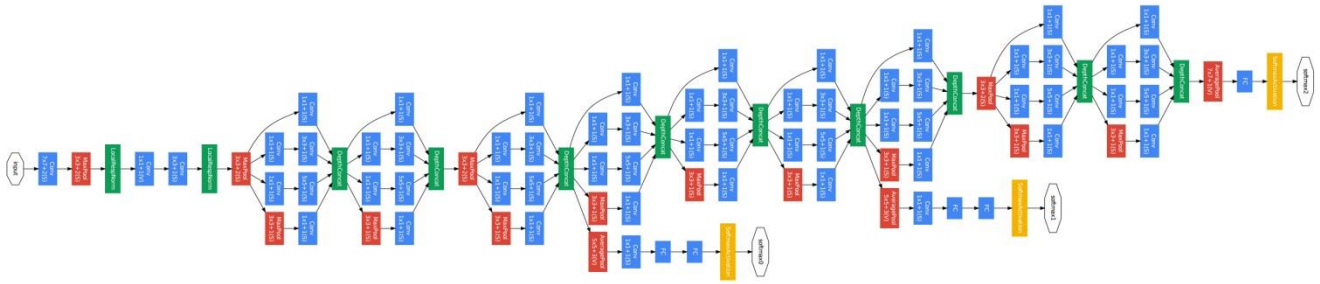
Intelligent Voice®

www.intelligentvoice.com

GoogLeNet Processing



GoogLeNet Processing



Intelligent Voice®

www.intelligentvoice.com

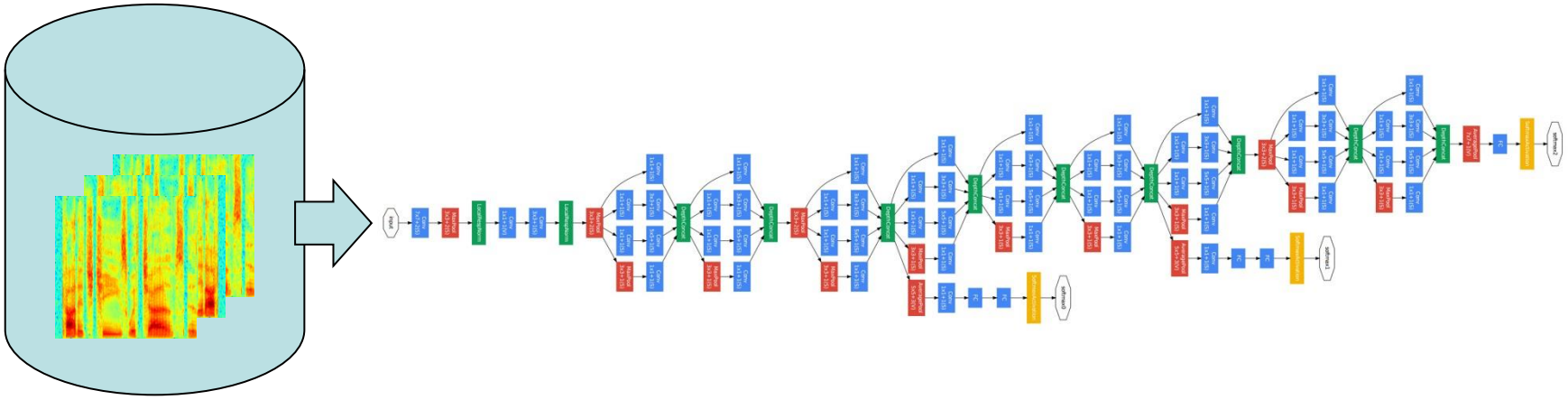
GoogLeNet Processing

Database:

501248 spectrograms for training

24352 spectrograms for validation

51501 spectrograms for testing

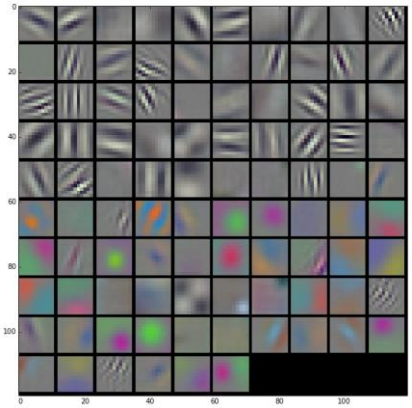


Intelligent Voice®

www.intelligentvoice.com

GoogLeNet Processing

Apply convolutions to extract primitives such as edges

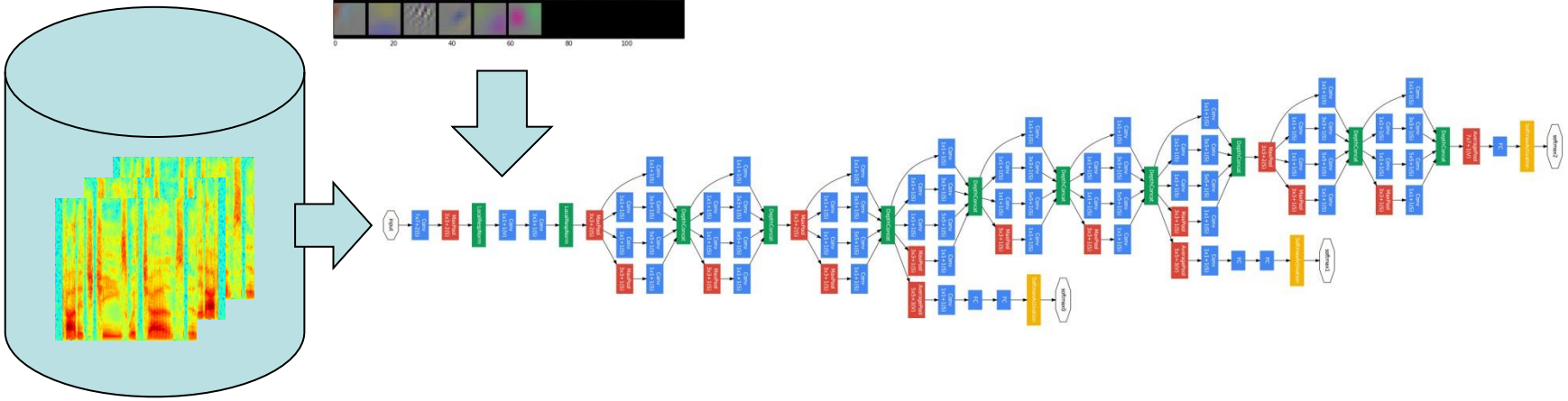


Database:

501248 spectrograms for training

24352 spectrograms for validation

51501 spectrograms for testing

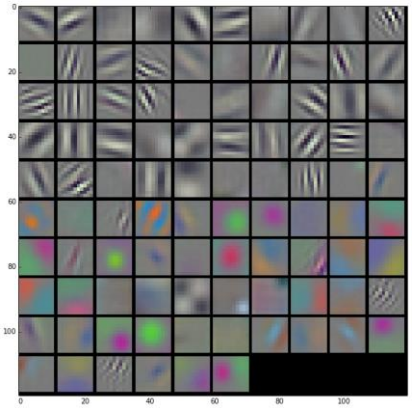


Intelligent Voice®

www.intelligentvoice.com

GoogLeNet Processing

Apply convolutions to extract primitives such as edges



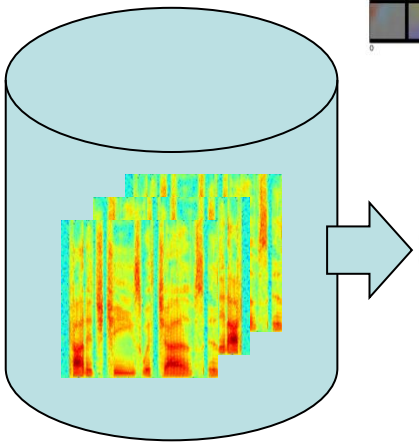
Object parts extracted

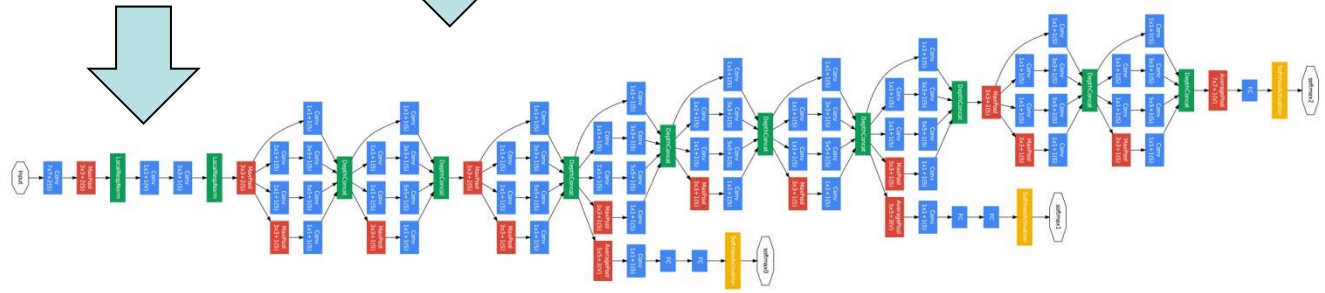
Database:

501248 spectrograms for training

24352 spectrograms for validation

51501 spectrograms for testing





Intelligent Voice®

www.intelligentvoice.com

GoogLeNet Processing

Apply convolutions to extract primitives such as edges

Object parts extracted

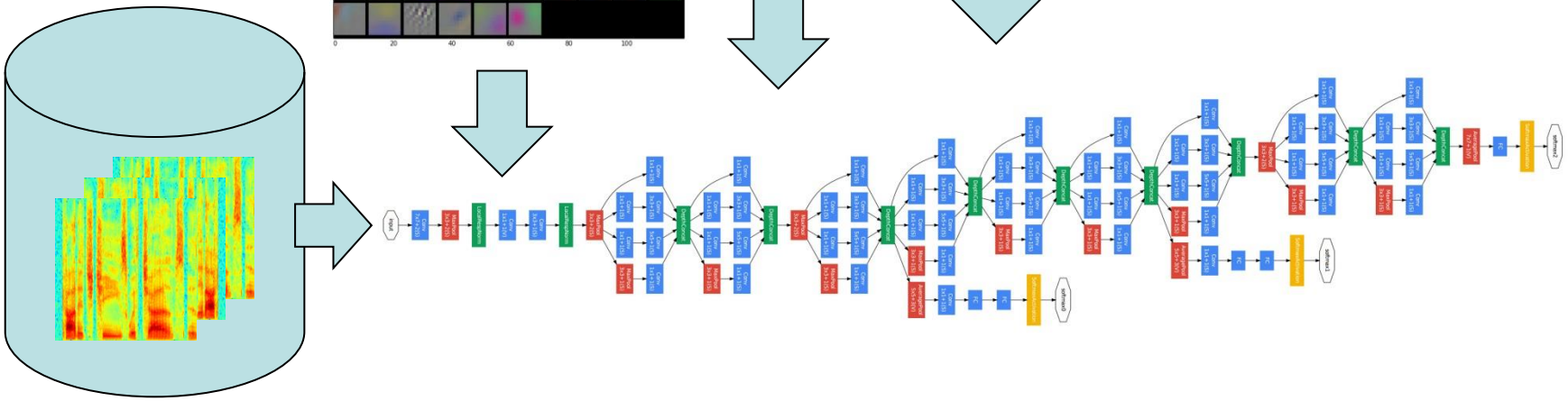
Full Spectral Features, e.g. phones, words

Database:

501248 spectrograms for training

24352 spectrograms for validation

51501 spectrograms for testing



GoogLeNet Processing

Apply convolutions to extract primitives such as edges

Object parts extracted

Full Spectral Features, e.g. phones, words

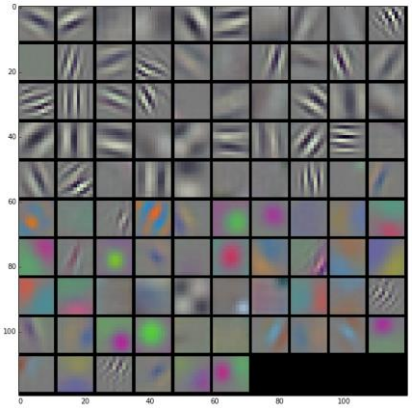
Refinement of accuracy

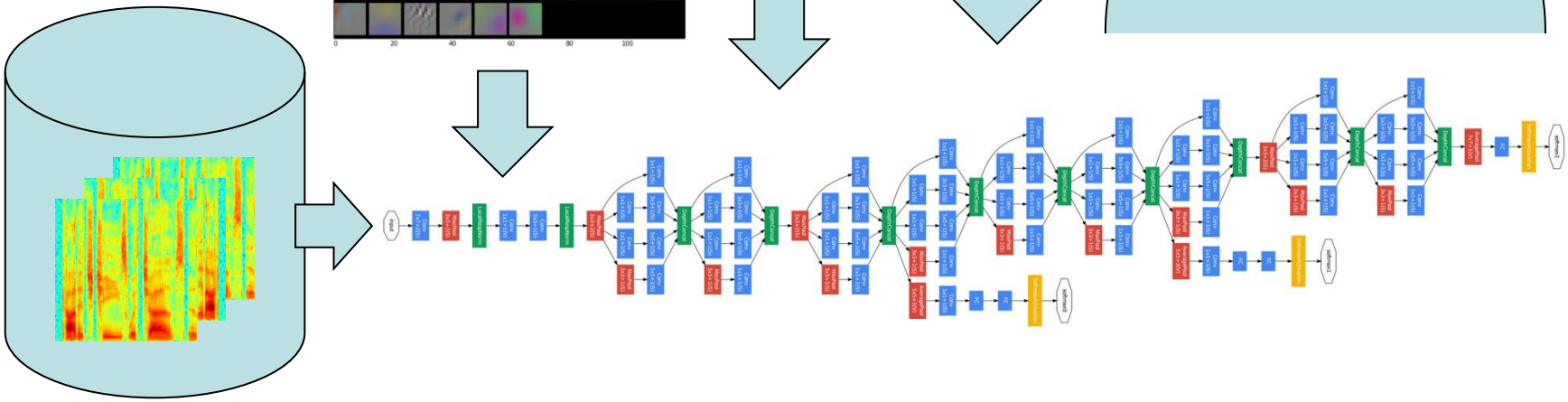
Database:

501248 spectrograms for training

24352 spectrograms for validation

51501 spectrograms for testing





Intelligent Voice®

www.intelligentvoice.com

GoogLeNet Processing

Apply convolutions to extract primitives such as edges

Full Spectral Features, e.g. phones, words

Dialect Classification

Object parts extracted

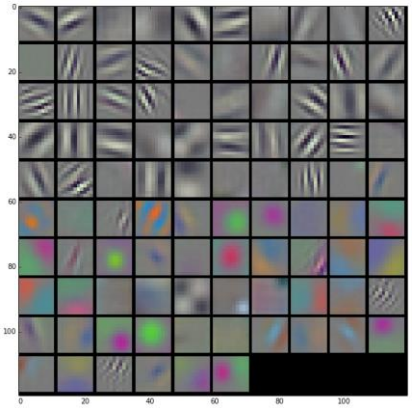
Refinement of accuracy

Database:

501248 spectrograms for training

24352 spectrograms for validation

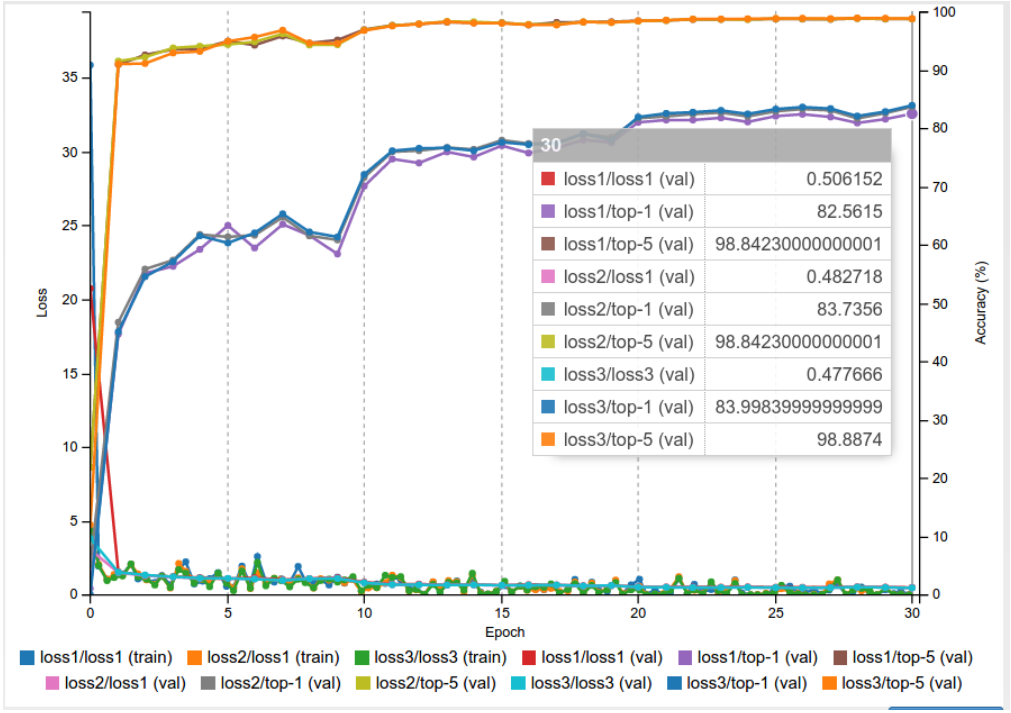
51501 spectrograms for testing



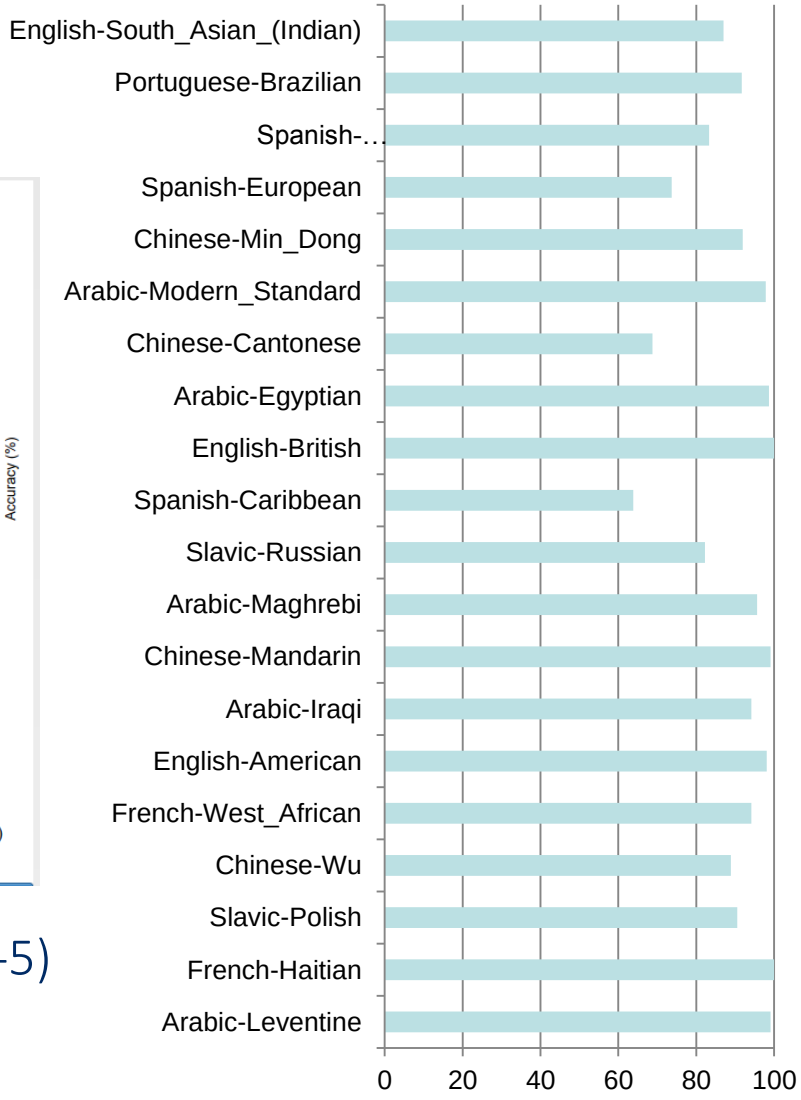
Intelligent Voice®

www.intelligentvoice.com

Preliminary Results



- Accuracy – 83.99 (Top-1), 98.89% (Top-5)



Still to be investigated…

- Many of the scaling, cropping, rotating of images common in image classification to balance data and improve generalisation is not appropriate for spectrograms
- Dynamic frequency warping techniques to balance the data sets and improve generalisation
- Taxonomy of languages investigation of the similarity of classification results across dialects
  - David Cameron – Arabic?

Questions

Thank you