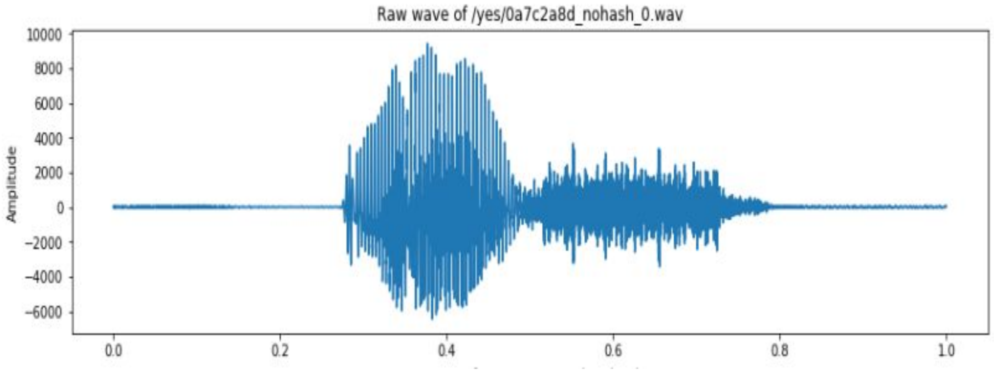
MGU Project 3

TensorFlow Speech Recognition Challenge

Data representation

Possible techniques:

Simple

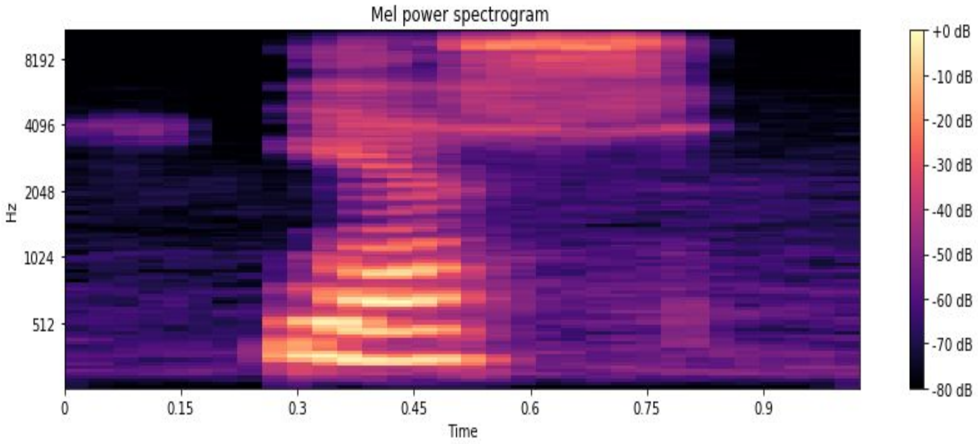


1D Conv

Dense

RNN

More informative



2D Conv

Dense

RNN

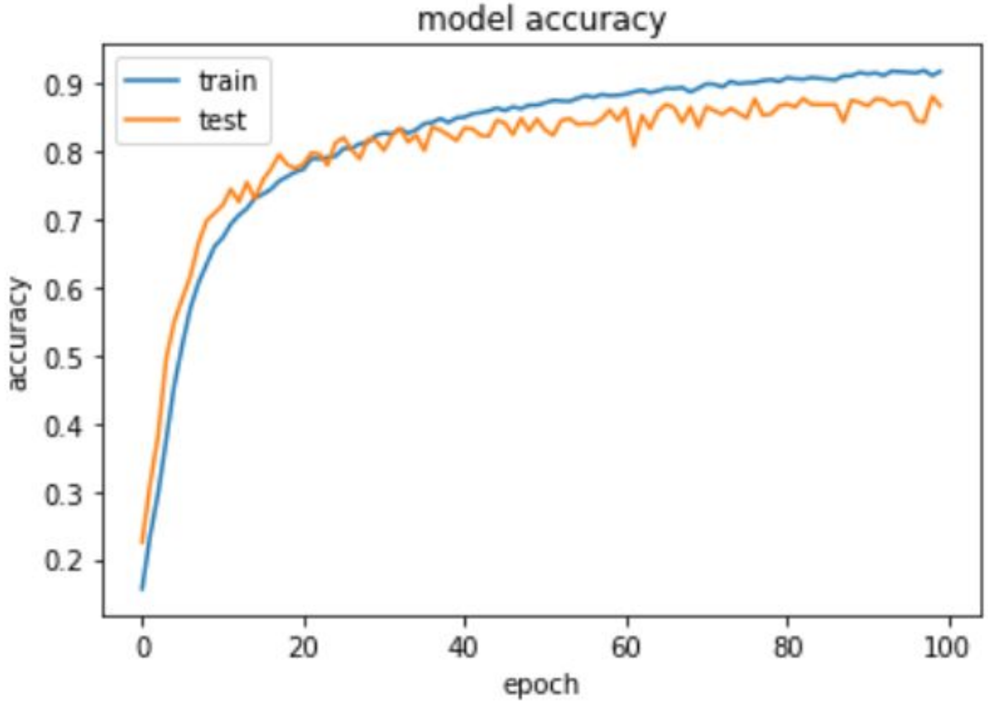
Baseline model

1D Convolutions + 2 x LSTM layer

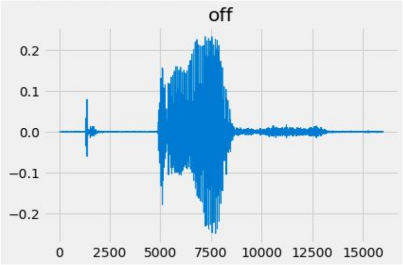
Validation accuracy almost 90%
Kaggle accuracy < 70% !?

Reasons:

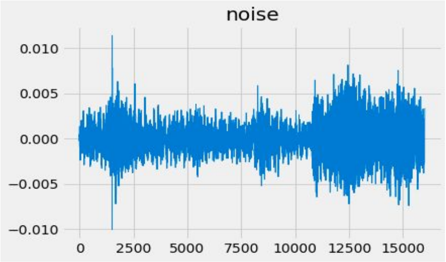
1. Omitting frequency dimension.
2. No data augmentation.
3. Not good enough model?



Data augmentation techniques

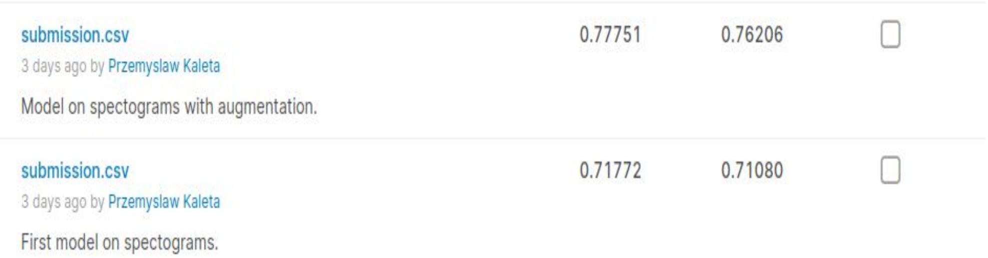


+ r



=

r - random noise level (between 0 and 1)
noise - randomly sampled from background noise



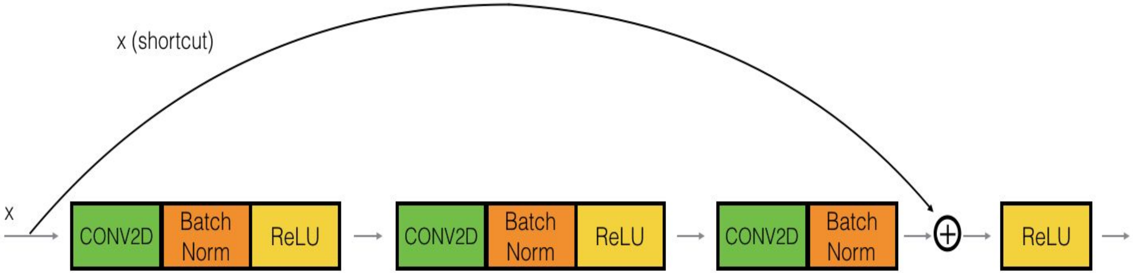
submission.csv
3 days ago by Przemyslaw Kaleta
Model on spectograms with augmentation.
0.77751 0.76206 □

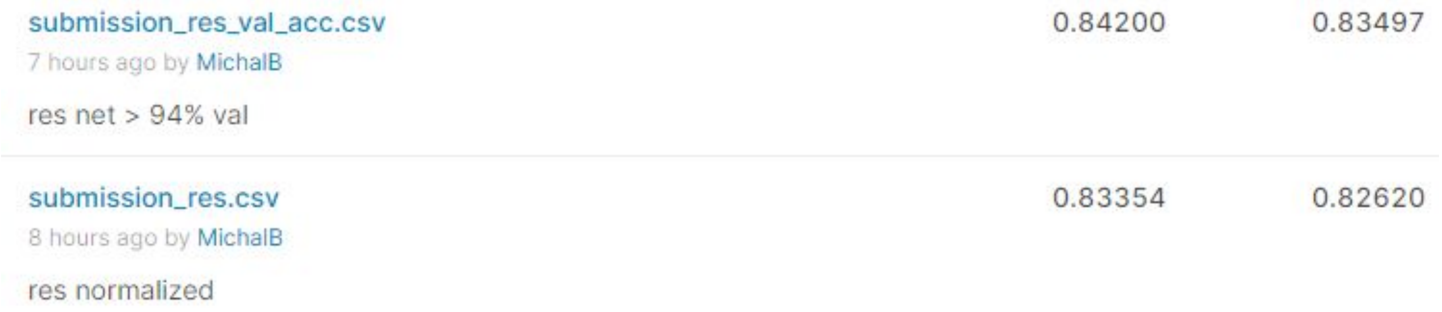
submission.csv
3 days ago by Przemyslaw Kaleta
First model on spectograms.
0.71772 0.71080 □

Convolutions only

ResNet:

1. Extremely good at features extraction.
2. No explicit “time” dimension



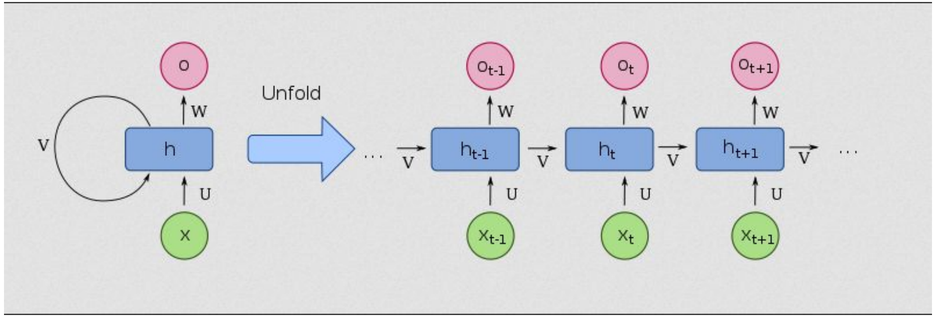


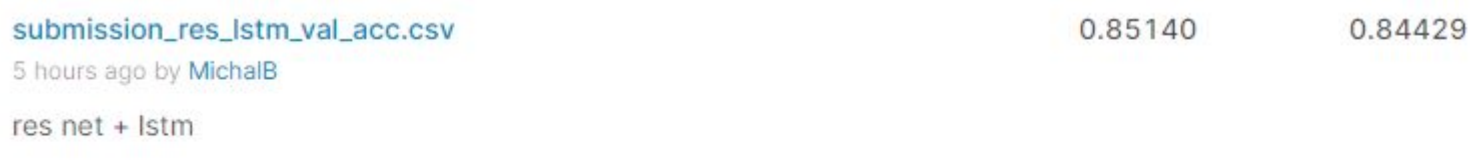
submission_res_val_acc.csv 7 hours ago by MichalB res net > 94% val	0.84200	0.83497
submission_res.csv 8 hours ago by MichalB res normalized	0.83354	0.82620

Recurrent architectures

ResNet + LSTM

Using recurrent network helps to capture sequential nature of speech data.





submission_res_lstm_val_acc.csv 0.85140 0.84429

5 hours ago by MichalB

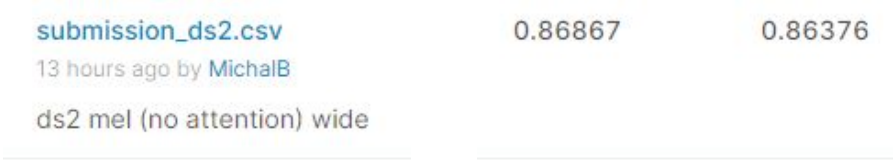
res net + lstm

Deep Speech 2 alike architecture

State of the Art model for speech recognition on the podium in most competitions from this field.

Our modification:

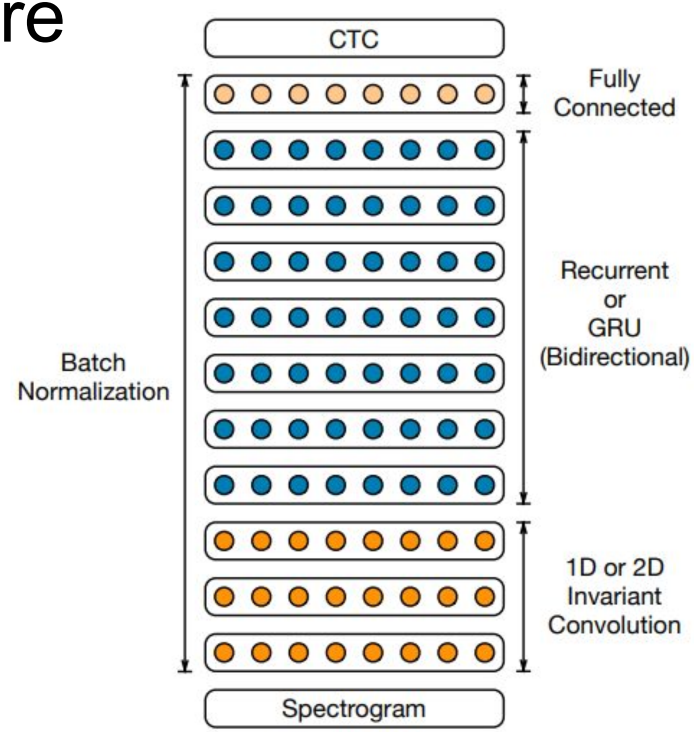
1. 2x Conv block
2. 2x Bidirectional wide (512 units) LSTM
3. Dense + dropout + softmax



submission_ds2.csv 0.86867 0.86376

13 hours ago by MichalB

ds2 mel (no attention) wide

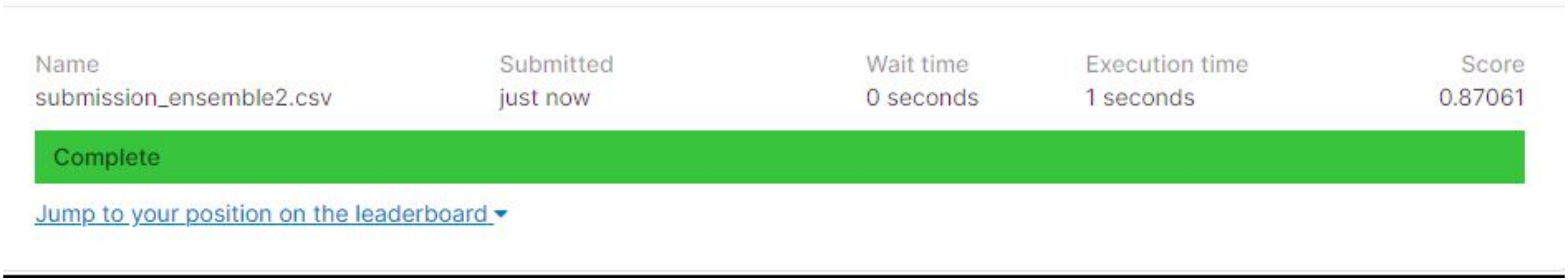

CTC
Fully Connected
Recurrent or GRU (Bidirectional)
Batch Normalization
1D or 2D Invariant Convolution
Spectrogram

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, Baidu Research – Silicon Valley AI Lab

Ensemble

Meta learning :
train upper level model how to interpret predictions from lover level models

We trained a lot of models. Why not use them all?



Name	Submitted	Wait time	Execution time	Score
submission_ensemble2.csv	just now	0 seconds	1 seconds	0.87061

Complete

Jump to your position on the leaderboard ▾

Conclusions

Data processing matters a lot.

Augmentation helps making models more robust.

Spectograms are helpful for speech representation.

Convolutional Neural Networks are useful for dimensionality reduction.

Recurrent Neural Networks can leverage CNN in tasks like this one.

Thank you :)