Распознавание речи обзор задачи: CTC подход на примере QuartzNet

На основе статьи:
QUARTZNET: DEEP AUTOMATIC SPEECH RECOGNITION WITH 1D TIME-CHANNEL SEPARABLE CONVOLUTIONS, 2019 г.

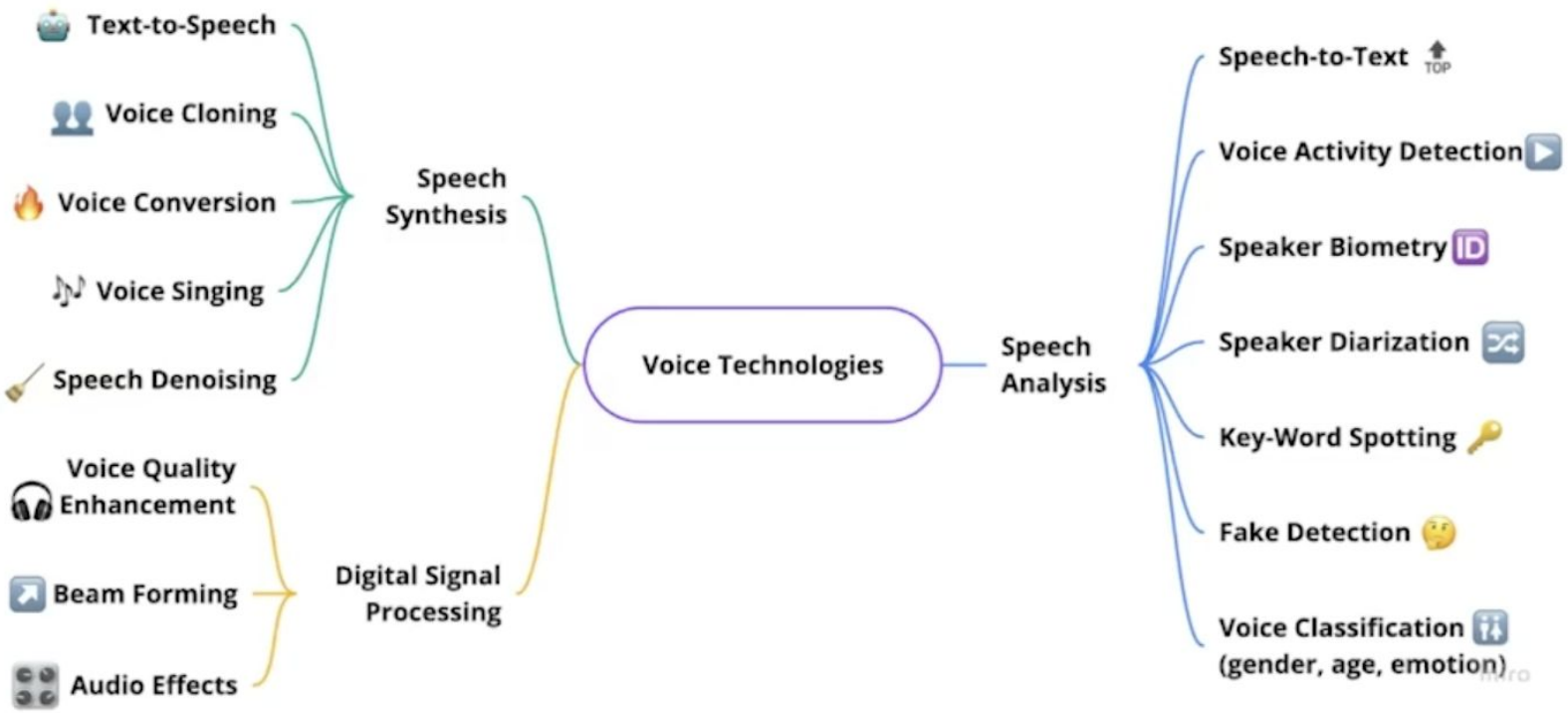
Щербакова Светлана, ММОВС21_3

Высшая школа экономики, 2022

План

Voice technologies
История
ASR
Преобразование сигнала
Преобразование Фурье
CTC
QuartzNet и сепарабельные свертки

Voice technologies



https://github.com/yandexdataschool/speech_course

HISTORY

1784

Wolfgang von Kempelen creates the **Acoustic-Mechanical Speech Machine** in Vienna

Thomas Edison invents the first dictation machine

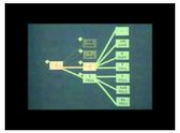
1879

1952

Bell Labs releases **Audrey**, capable of recognizing spoken digits with **90% accuracy** - but only when spoken by its inventor

IBM Shoebox can understand 16 English words

1962



1971

Harpy, created at Carnegie Mellon University, can comprehend **1,011 words** - and some phrases

IBM Tangora, using the Hidden Markov Model, predicts upcoming phonemes in speech

1986

2006

The National Security Agency (NSA) starts using speech recognition to isolate key words in recorded speech

Google launches a voice search app, bringing speech recognition to mobile devices

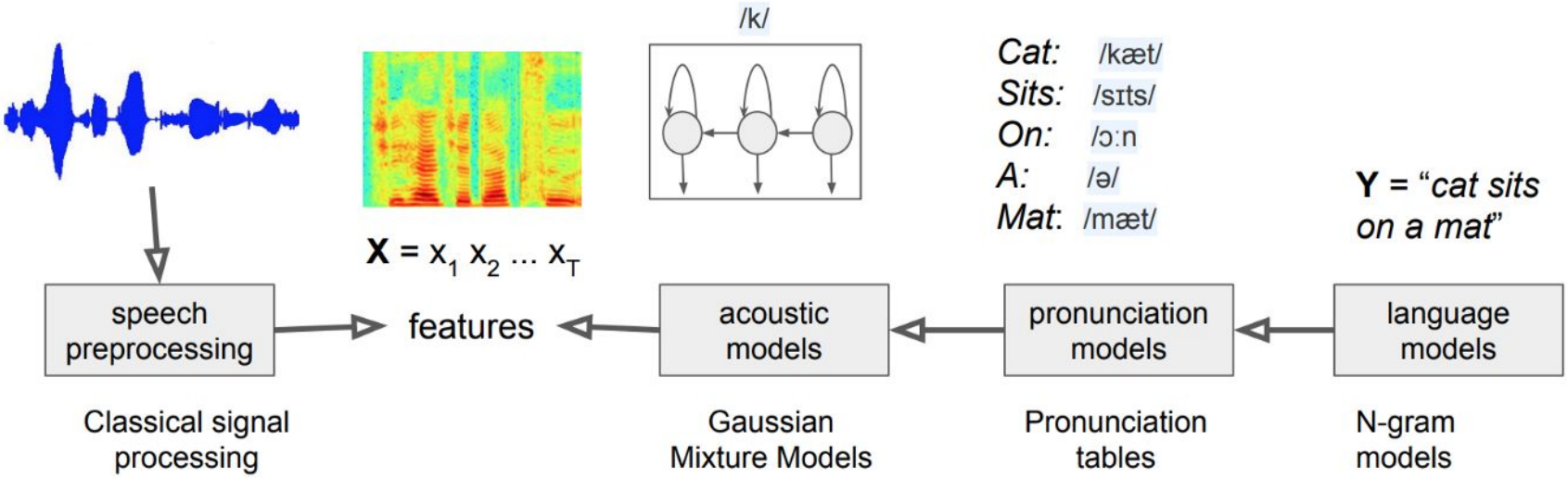
2008

2011

Apple announces Siri, ushering in the age of the voice-enabled digital assistant

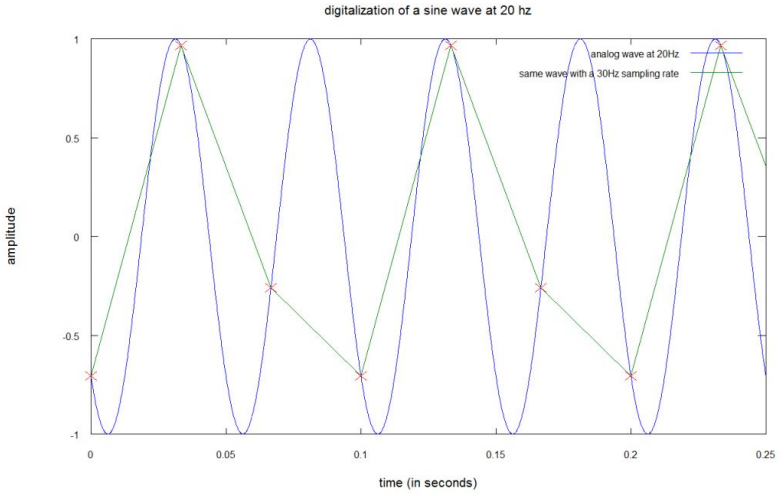
ClickZ

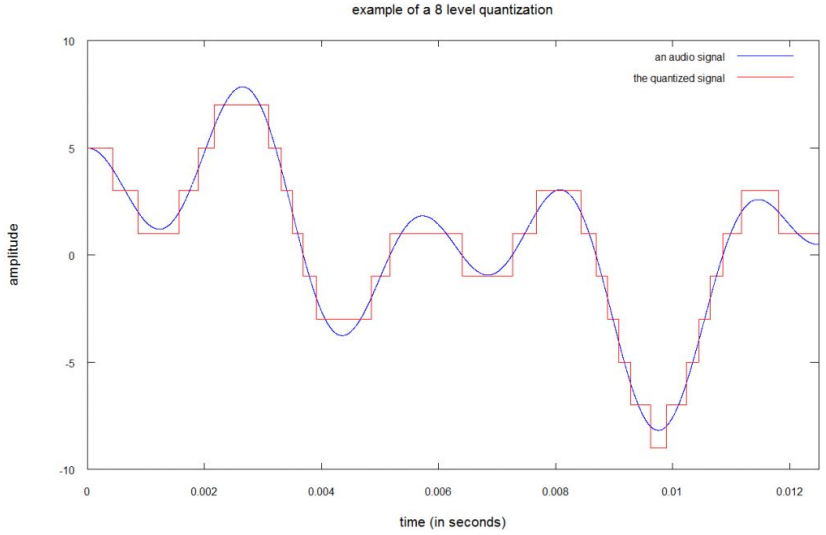
ASR



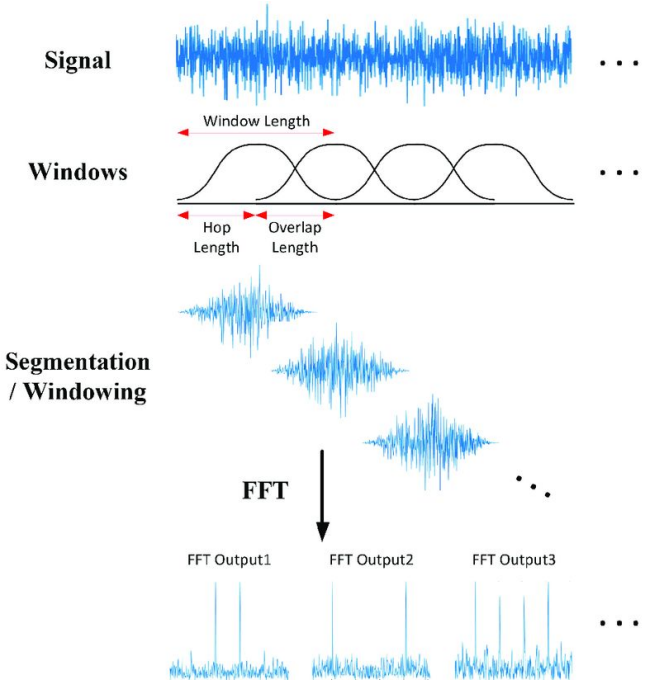
$$P(W_n \mid W_1, W_2, \ldots, W_{n-1}) \approx P(W_n \mid W_{n-1}) \qquad b_j(\mathbf{x}) = p(\mathbf{x} \mid S = j) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$$

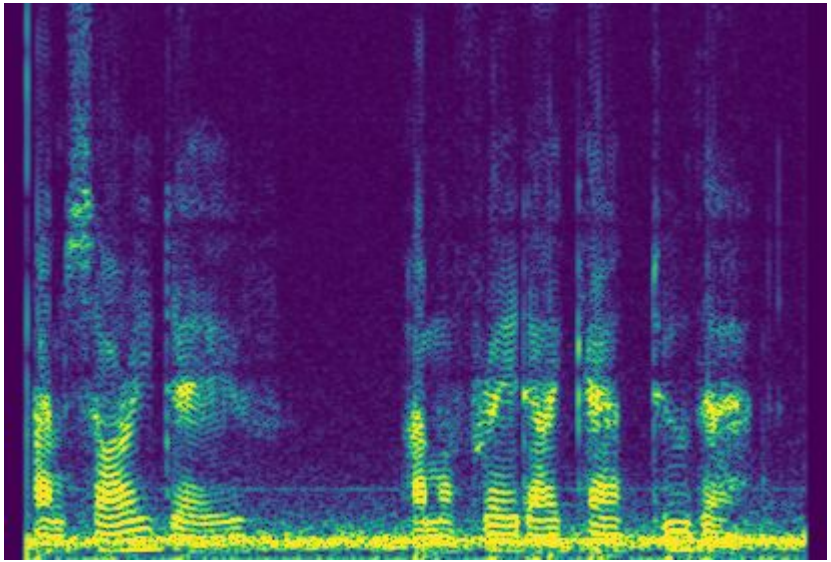
Преобразование сигнала



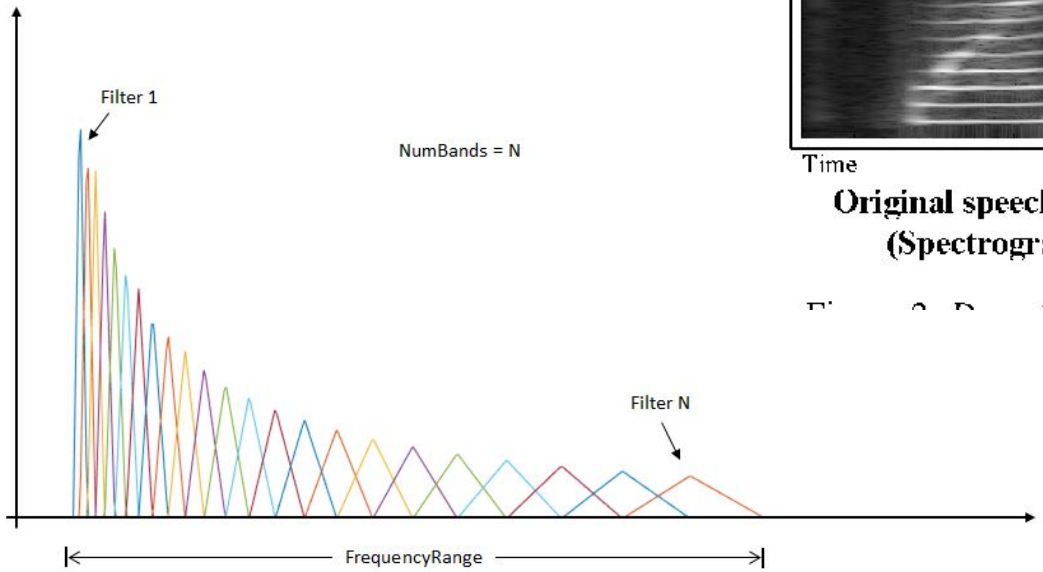


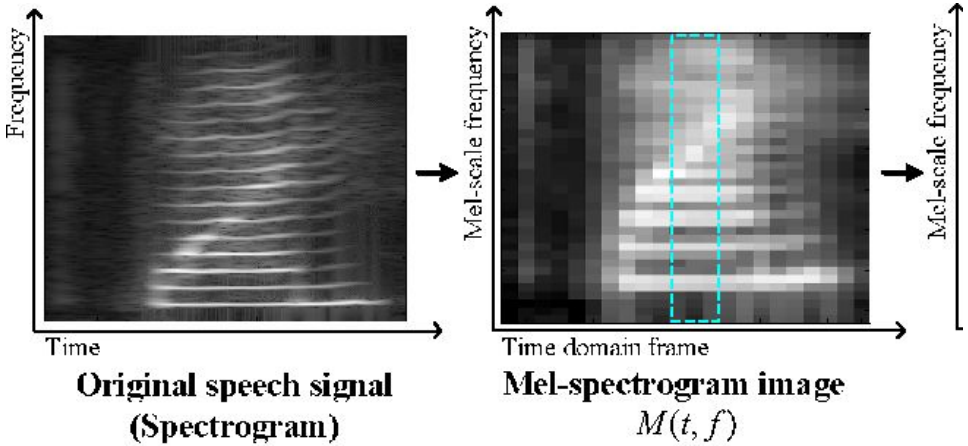
Преобразование Фурье





МЕЛ Спектрограмма

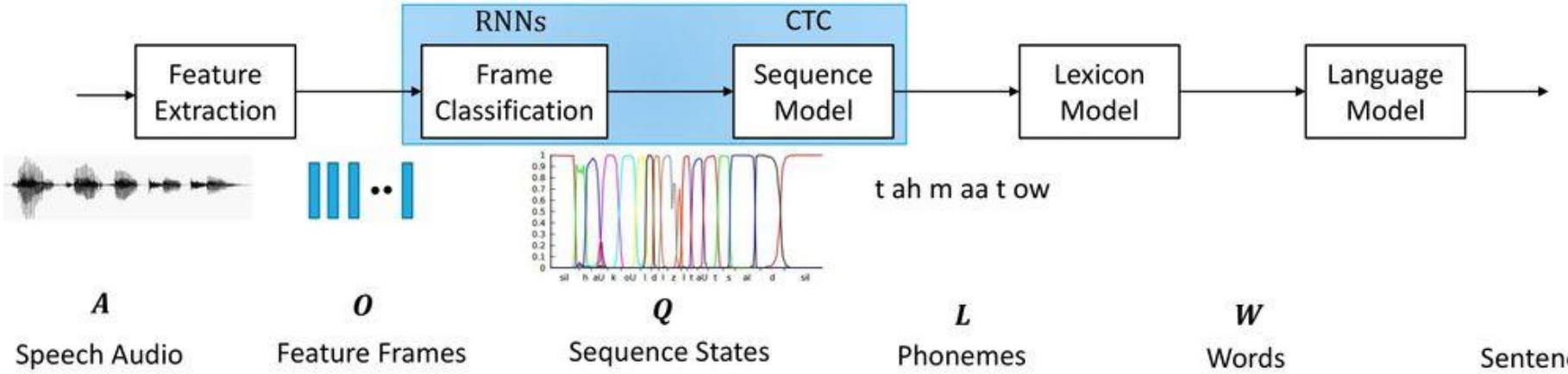




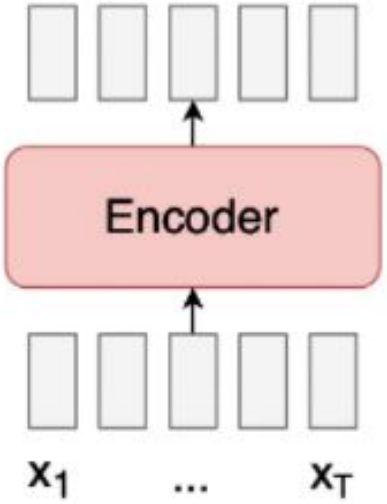
CTC

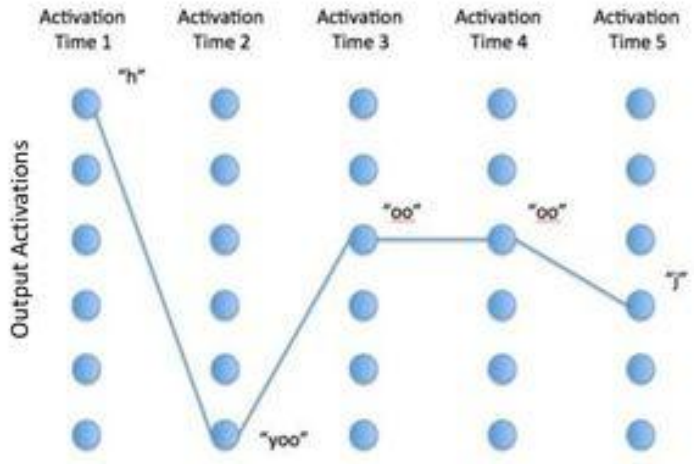
RNN: Recurrent Neural Networks

CTC: Connectionist Temporal Classification



CTC





h h e ϵ ϵ l l l ϵ l l o

h e ϵ l ϵ l o

h e l l o

First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.

LOSS CTC

the ground truth of the word sequence — acoustic frames

$$\mathcal{L}_{CTC} = -\log P(\boldsymbol{S}|\boldsymbol{X})$$

$$P(\boldsymbol{S}|\boldsymbol{X}) = \sum_{\boldsymbol{c}\in A(\boldsymbol{S})} P(\boldsymbol{C}|\boldsymbol{X})$$

sum over all possible paths

(e.g. ccɛaaɛtt, cccɛaɛtt, cɛaɛtttt, ...)

$$P(\boldsymbol{C}|\boldsymbol{X}) = \prod_{t=1}^{T} y(c_t, t)$$

joint probability of a path

(e.g. ccɛaaɛtt)

LOSS CTC

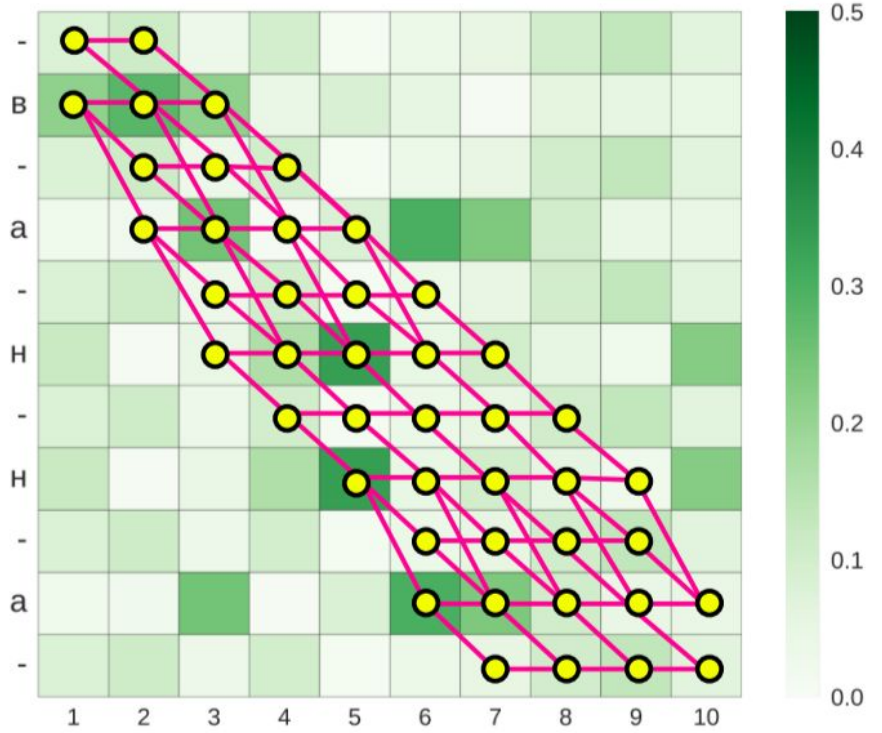
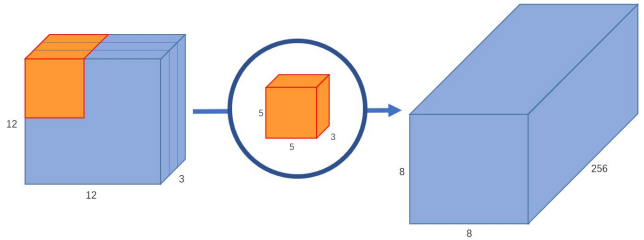
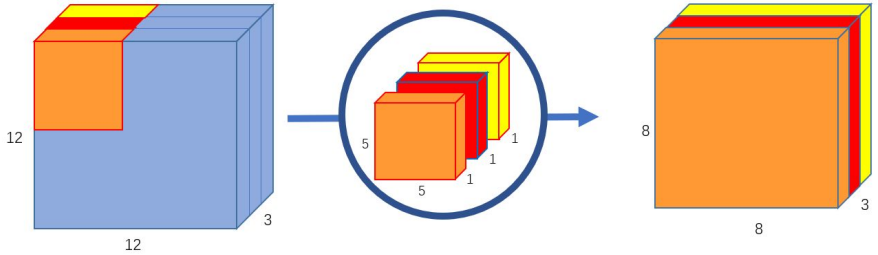


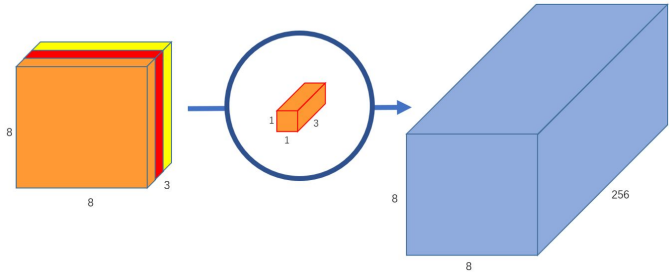
Табл. 2.3 — Всевозможные пути, приводящие за время T к результату l при преобразовании их функцией B.

Сепарабельные свертки

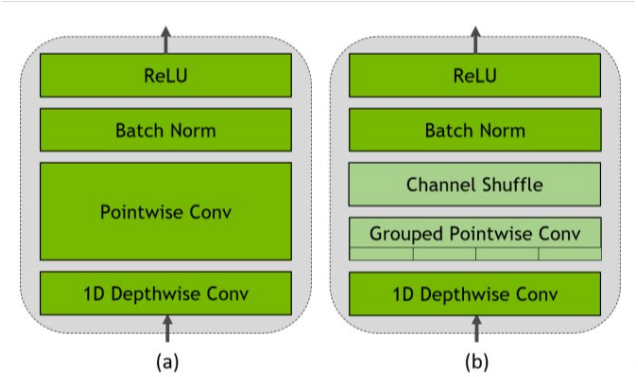
$$\begin{bmatrix} 3 & 6 & 9 \\ 4 & 8 & 12 \\ 5 & 10 & 15 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$







QuartzNet



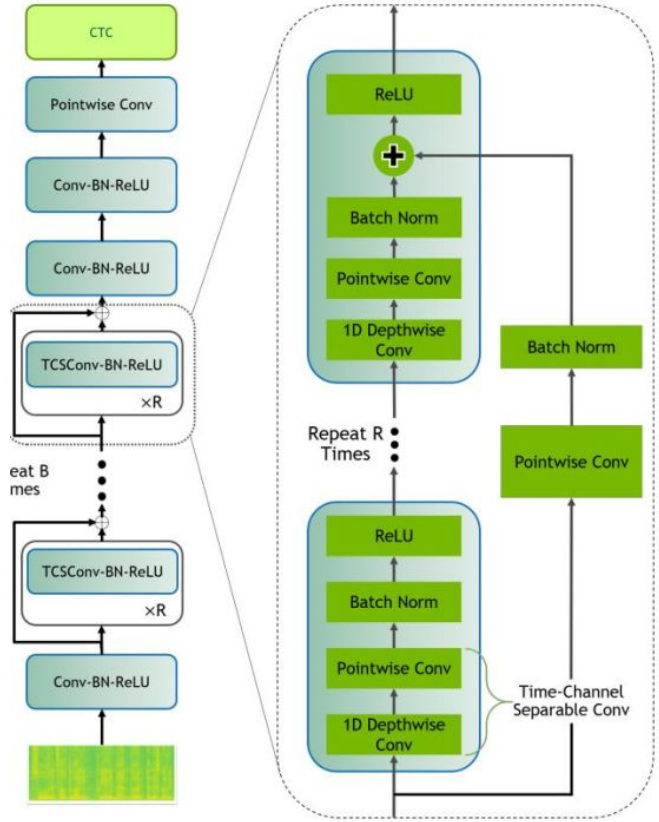


Рис. 1. Архитектура QuartzNet BxR

Вопросы

Что такое ASR

Чем полезна CTC в архитектуре модели распознавания речи

Зачем нужно все остальное если есть архитектуры основанные на Attention