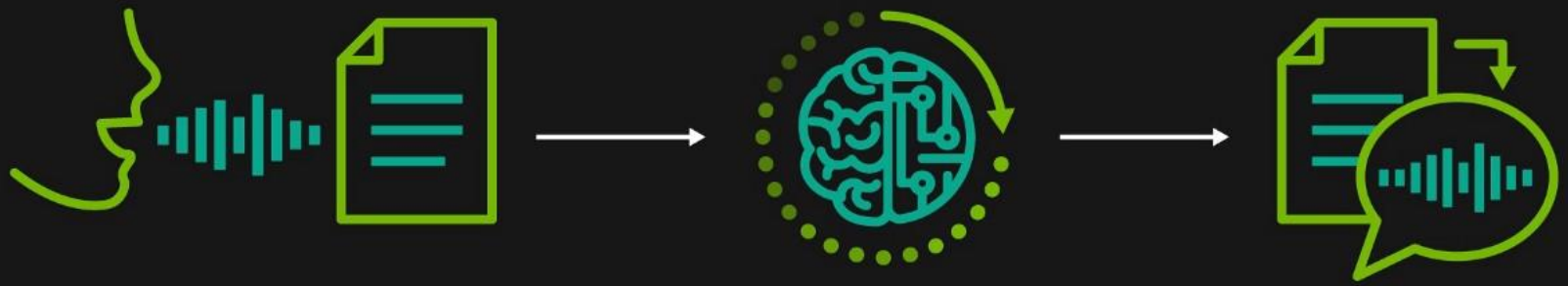
Nemo and Rasa Application in a Chatbot Environment

Natural Language Processing Project

Andrea Gurioli, Giovanni Pietrucci e
Mario Sessa

Objectives



Automatic Speech Recognition

Natural Language Processing

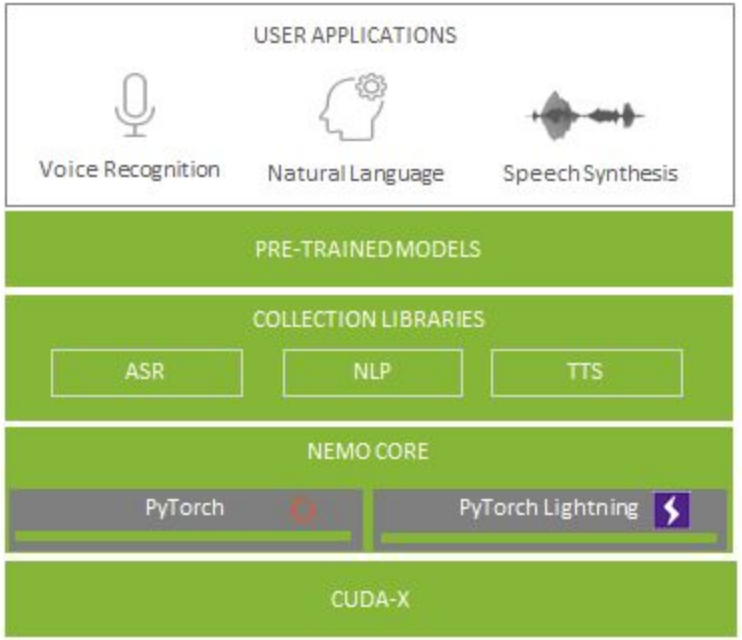
Text to Speech

NeMo Application Stack

NeMo provides different modules to build pre-trained models using collections libraries.

NeMo models are built on PyTorch and PyTorch Lightning libraries

It is possible to improves performances with the GPU-based library CUDA-X.



Automatic Speech Recognition Pipeline

audio

Preprocessing

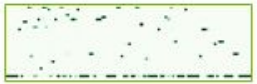
- Data augmentation
  - speed perturbation (resampling)
  - additive Gaussian noise
  - time / frequency masks (cutout)
- Windowing 20-25 ms, stride 10 ms
- FFT, log, mel scale
- Normalization

spectrogram

Neural Acoustic Model

- Jasper
- Wav2Letter+
- DeepSpeech 2

$P_t(c)$



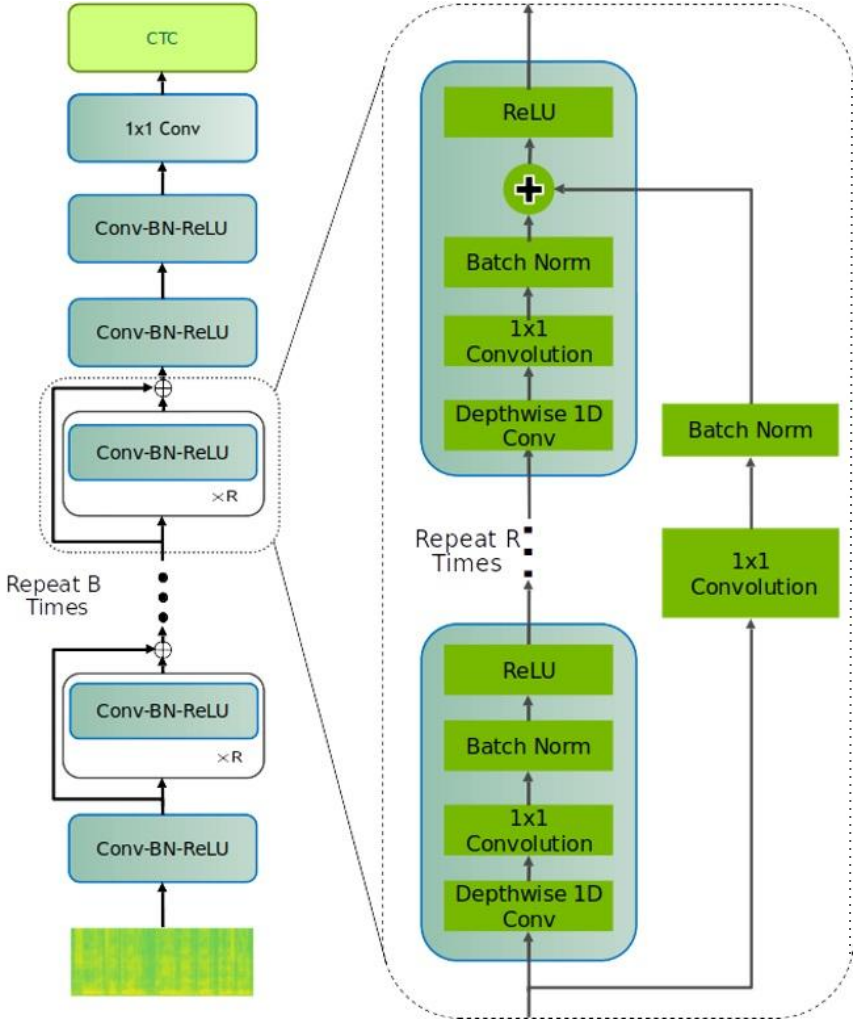
training

CTC Loss

inference

Decoder

- greedy (argmax)
- beam search
- beam search with language model rescoring



Quartznet

Small WER and high transcription speed. It is based on Jasper model.

Its composition consists in 1 Convolutional layer followed by a sequence on N blocks repeated Z times

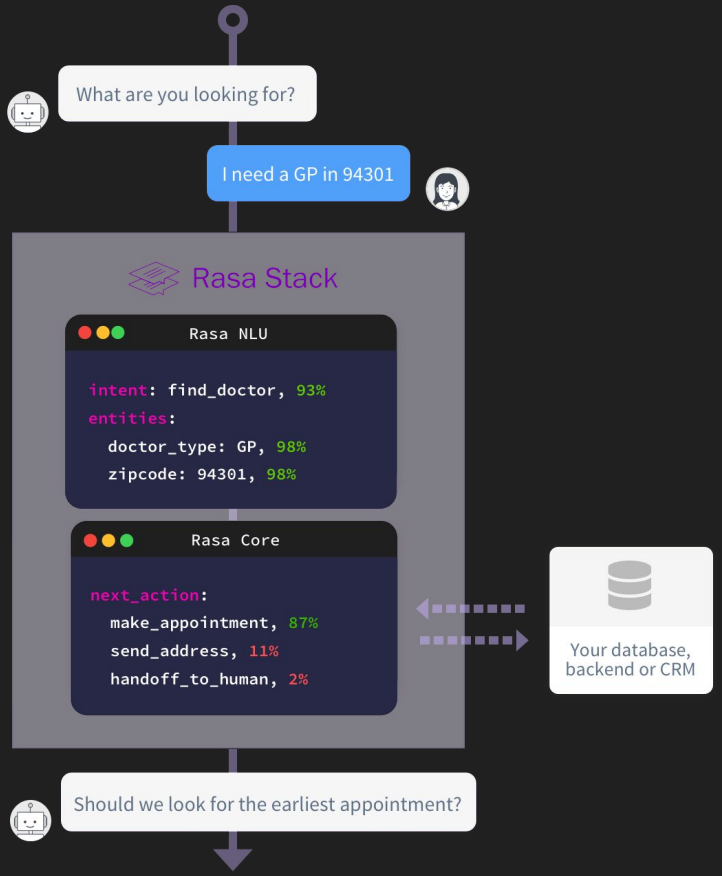
Each block has a depthwise 1D Convolutional, Pointwise Convolutional and a Normalization layer with a ReLu activation function

Rasa Framework

Rasa is an open-source conversational AI tools formed by two modules:

1. **Rasa NLU:** It consists in the intent classification and entity extrapolation

2. **Rasa Core:** ML-based dialogue management to predict appropriate actions from an intent trigger.

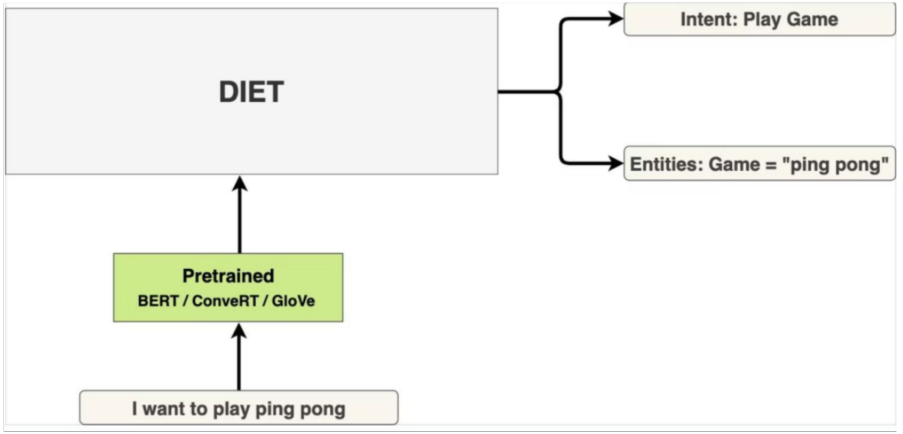


NLP Pipeline

During intent predictions and entity extractions, Rasa NLU executes a pipeline based on question tokenization, tokens featurizing and output prediction.

We tested a Sparse (BoW) and Dense (Word embedding) vector on a DIET Classifier which uses an inner CRF Entities Extractor.

During the analysis, we saw better performances on the chosen Sparse vector usage.

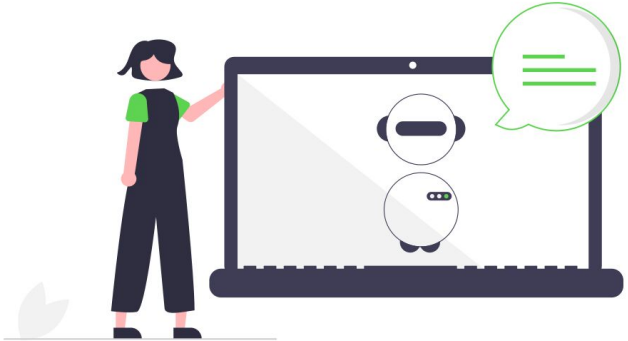


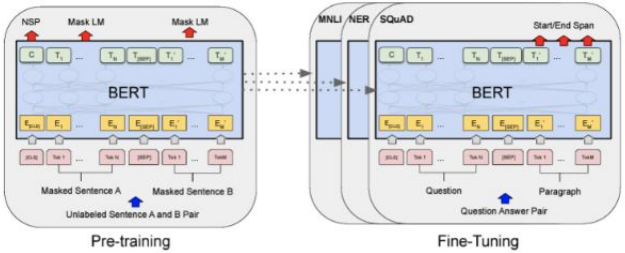
Vectors	F1	Recall	Precision
Sparse vectors	0.92 (±0.05)	**0.90**(±0.03)	**0.90**(±0.06)
Dense vectors	0.89(±0.04)	**0.91**(±0.03)	0.90(±0.05)

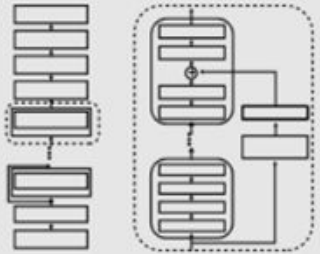
Alysia Tasks

- To-Do List Management
- Weather System
- Jokes System
- Sending Emails
- Time Service
- Wikipedia Search

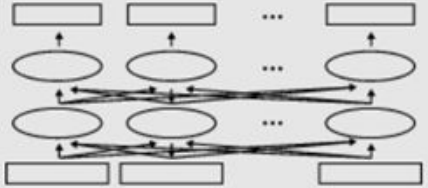
Wikipedia Search System uses the QAModel provided by NeMo NLP based on a BERT-based configuration and trained on the SQuAD dataset.



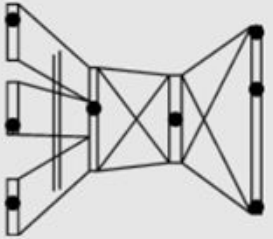




Acoustic models

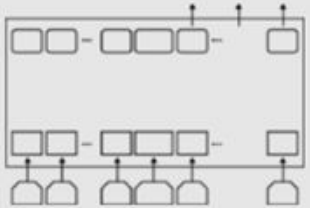


Decoders

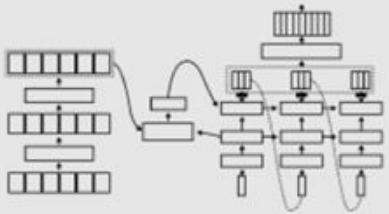


Language models

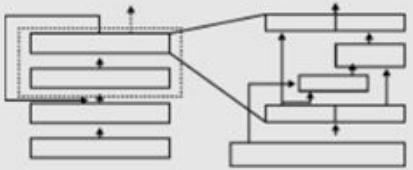
TTS Modules



BERT models

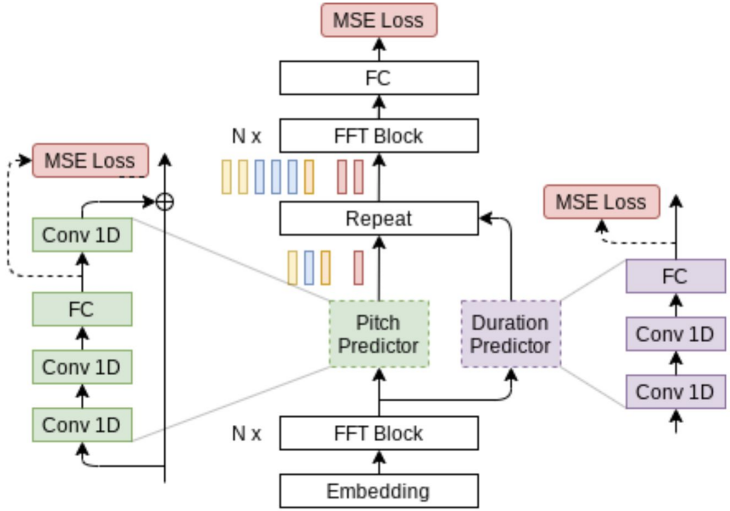


Speech Synthesis models



Voice encoder models

Spectrogram Generator



FastPitch

It's the chosen model for the spectrogram generator.

FastPitch is a fully-parallel text-to-speech model based on FastSpeech with higher real-time factor than Tacotron2 for the mel-spectrogram synthesis of an utterance.

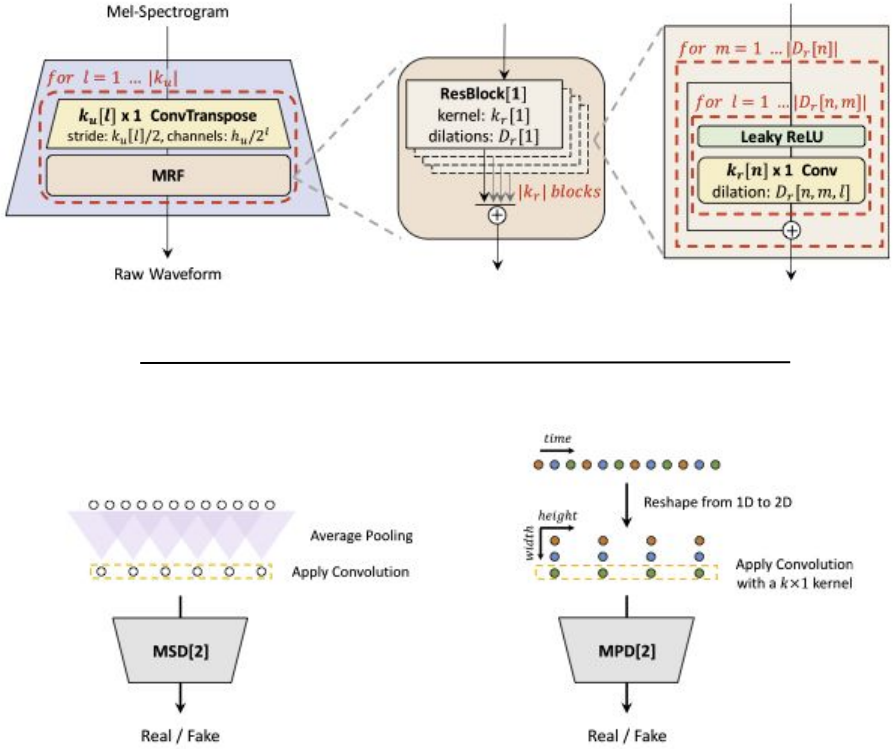
It processes letters and predicts pitches and durations of mel-spectrogram elements using a parallel approach.

HiFiGAN

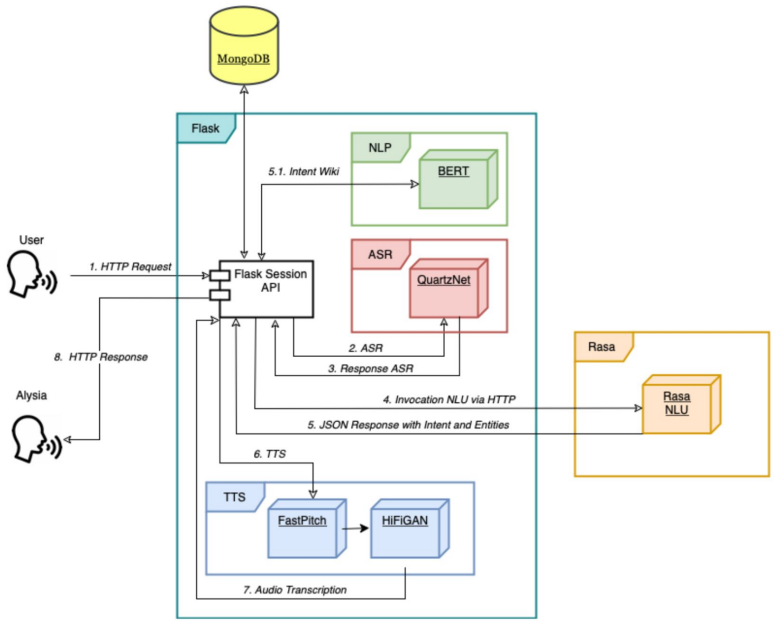
GAN-based model for audio generation from a mel-spectrogram.

It is composed by a generator and two discriminators: multi-scale and multi-period discrimnators.

HiFiGAN defines one of the state-of-art speech synthesis model with optimal performances in synthesis speed suitable for real-time applications like Alysia



Alysia Deployment Diagram

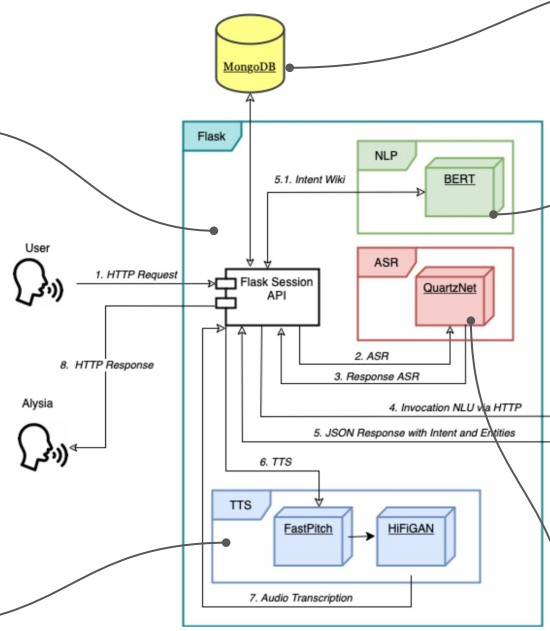


Alysia Deployment Diagram

Flask Server for Models Bridging and Actions Management

MongoDB for Entity Persistance

NeMo NLP Model for Wikipedia Search Service

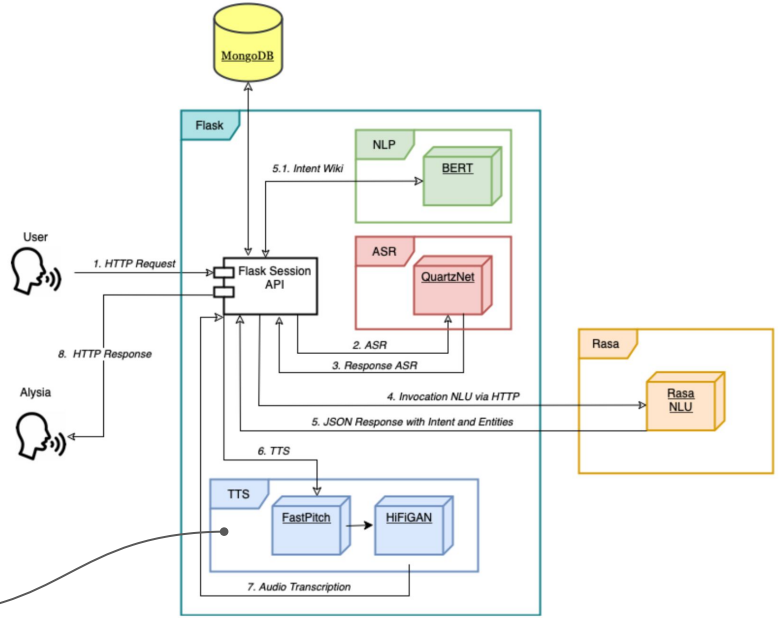


FastPitch and HiFiGAN for Speech Synthesis from an input text

Quartznet 15x5 for Audio transcription

Rasa Server for Rasa NLU and Rasa Core Deployment

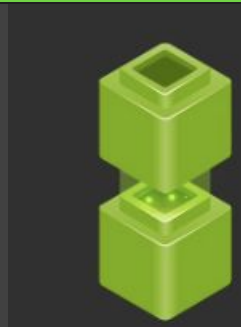
Alysia Deployment Diagram



Important Note:

For Performances problems, we replace it with the Speech Synthesis Model from Web Speech API

Conclusions



- NVIDIA NeMo is adaptive, easy to use and with high abstraction for many applications.
- It guarantee state-of-art models in pre-trained form available in its collection libraries.
- It can integrates external services like MongoDB or other databases for entity persistence and intent predictions or entity extractions like Rasa and Google Dialogflow.
- NeMo framework obtains the high performances in TPU or GPU-based running environments but it works on CPU-only host too.

NVIDIA

References

- Automatic Speech Recognition: A Deep Learning Approach, Dong Yu, Li Deng, Springer Edition, 2015
- Audio Augmentation for Speech Recognition, Tom Ko, Vijayaditya Peddint, Daniel Povey, Sanjeev Khudanpur, 2015
- Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Graves, Alex et al., 2006
- QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions, Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Yang Zhang, 2019
- Rasa: Open Source Language Understanding and Dialogue Management, Tom Bocklisch, Joey Faulkner, Nick Pawlowski, Alan Nichol, 2017
- Pointwise Convolutional Neural Networks, Binh-Son Hua and Minh-Khoi Tran and Sai-Kit Yeung, 2018
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2019
- FastPitch: Parallel Text-to-speech with Pitch Prediction, Adrian Łáncucki, 2021
- HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, Jungil Kong, Jaehyeon Kim, Jaekyoung Bae, 2020
- MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, Aaron Courville, 2019
- NeMo: a toolkit for building AI applications using Neural Modules, Oleksii Kuchaiev, Jason Li, Huyen Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, Stanislav Beliaev, V. Lavrukhin, J. Cook, P. Castonguay, M. Popova, J. Huang, J. M. Cohen. Santa Clara, NVIDIA, 2019