NVIDIA Jarvis
Conversational AI

Ricky and Matan

USE CASES

ACROSS ALL VERTICALS

Online Store

Provide conversational interface for shopping

Industrial

Collaborative robots - Robots and humans collaborate in close proximity

Engineer troubleshooting with the help of AI assistant

Finance

Call center: Sentiment of customers calling

Insurance chatbot: “Add a wedding ring to an insurance policy via an image and receive policy price quote”

Energy / Oil & Gas

Use camera and ask, “what are the safety guidelines for this chemical”?

Loud environment - virtual assistant using lip reading

Consumer Internet

Video diarization - Meeting/conversation transcription per person with timestamps

Content tagging with Image, text, Audio - Recommendation, Ads

In car experience

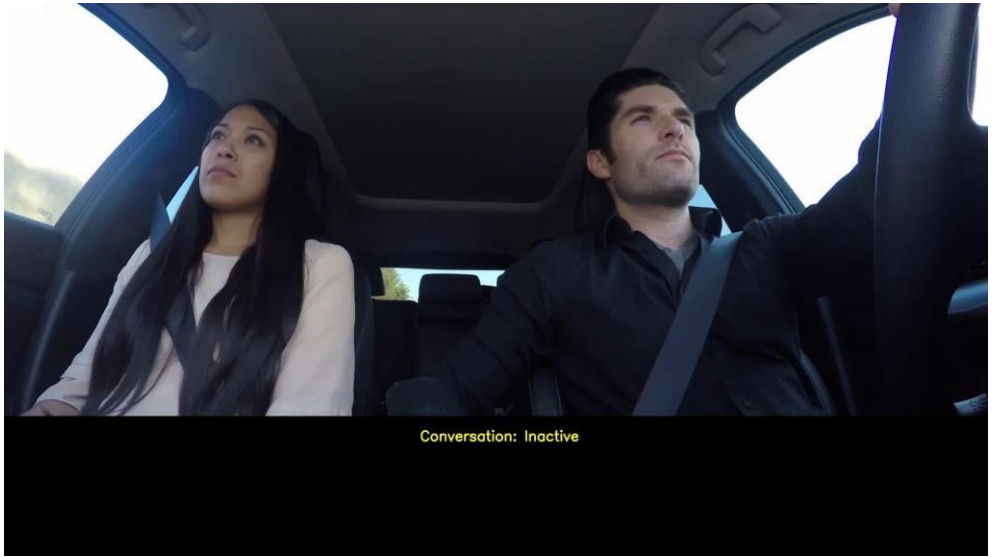
Autonomous Driving: Enhanced In-car experience combining visual inputs with speech

CHALLENGES OF CONVERSATIONAL AI

Custom models	Deployment	Multiple sensors	High accuracy	Real Time
Cloud services not customizable High costs Data Sovereignty	Existing software not designed for modern production environments	Difficult to use multiple sensors efficiently	Need state-of-the-art algorithms and models	Requires low latency for natural interaction

JARVIS

Platform to develop and deploy conversational AI applications
Designed for sensor fusion

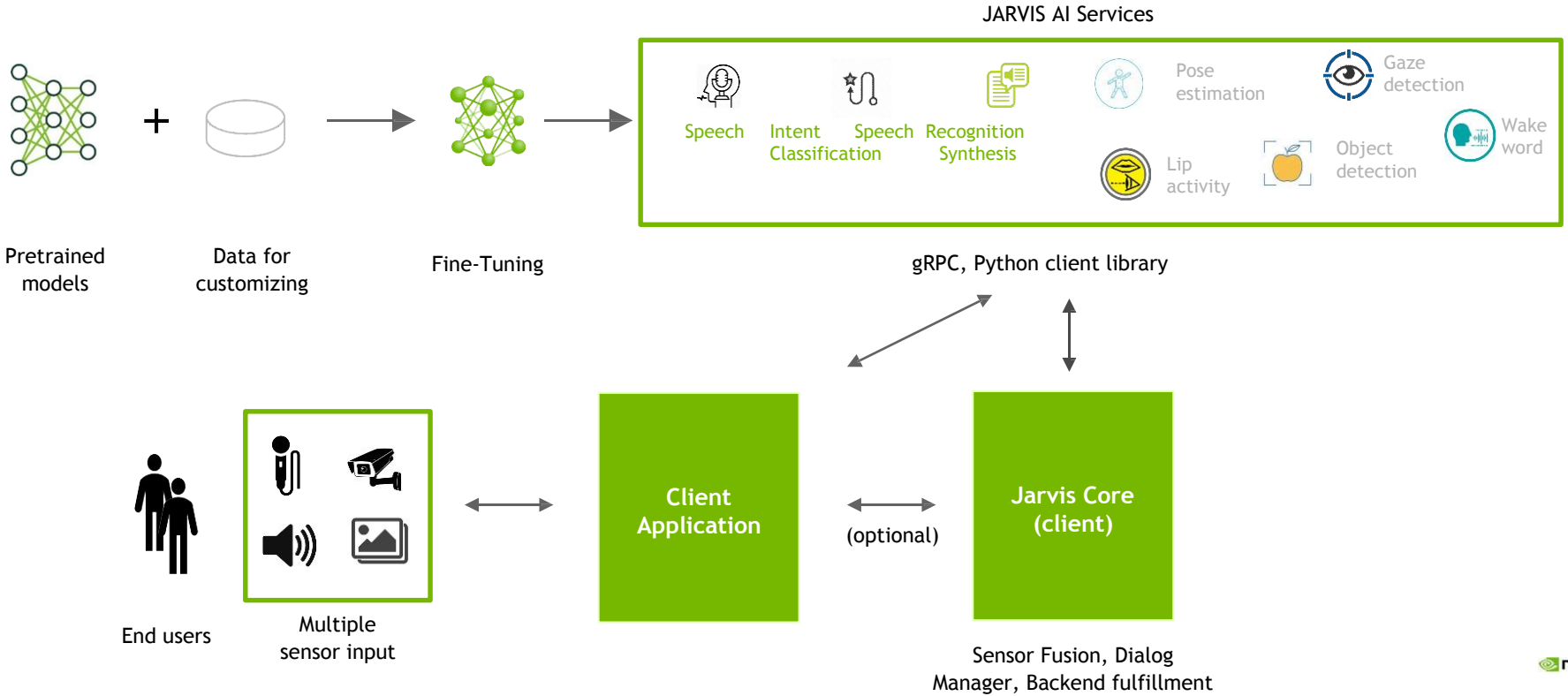


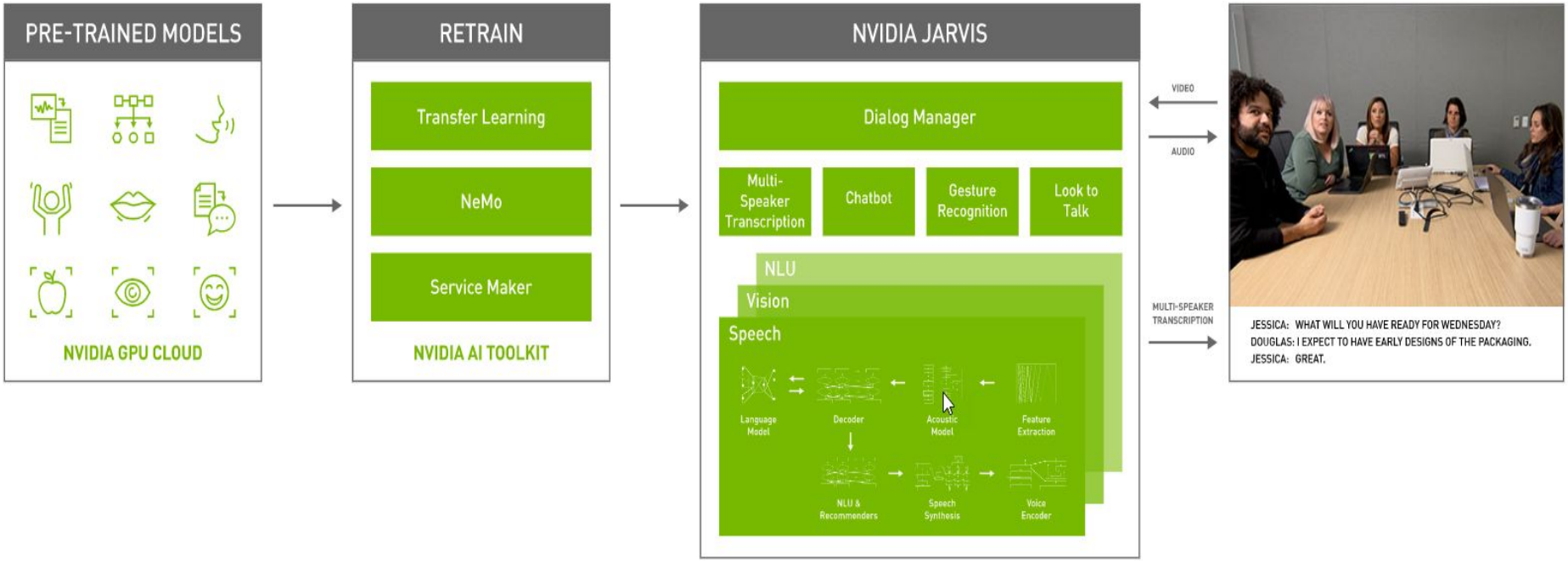
JARVIS BENEFITS

Custom models	Deployment	Multiple sensors	High accuracy	Real Time
Start from base model, train with *your* data on *your* infrastructure	Micro-service approach Designed for K8s Simple APIs, easy to integrate	Framework for training and deploying models across modalities Tools to simplify fusion	Best-in-breed algorithms Direct access to cutting-edge research	End-to-end inference on GPUs optimized to reduce latency

JARVIS WORKFLOW OVERVIEW

More details of Jarvis workflow can be found here




PRE-TRAINED MODELS
NVIDIA GPU CLOUD
RETRAIN
Transfer Learning
NeMo
Service Maker
NVIDIA AI TOOLKIT
NVIDIA JARVIS
Dialog Manager
Multi-Speaker Transcription
Chatbot
Gesture Recognition
Look to Talk
NLU
Vision
Speech
Language Model
Decoder
Acoustic Model
Feature Extraction
NLU & Recommenders
Speech Synthesis
Voice Encoder
VIDEO
AUDIO
MULTI-SPEAKER TRANSCRIPTION
JESSICA: WHAT WILL YOU HAVE READY FOR WEDNESDAY?
DOUGLAS: I EXPECT TO HAVE EARLY DESIGNS OF THE PACKAGING.
JESSICA: GREAT.

Visual Diarization

Multiple speaker transcription based on video and audio streams

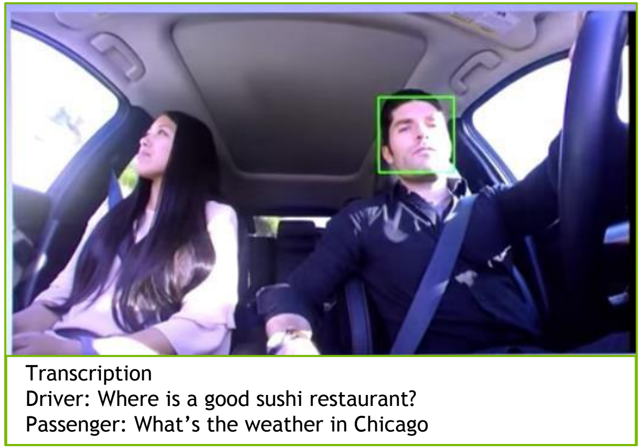
Interaction: Jupyter notebook with live video stream overlaying gaze detection and lip activity detection and producing a text transcript per person from the audio stream

Technology of sensor fusion:

- Video stream
    - Gaze detection to engage the system
    - Lip activity to determine who is speaking
- Audio stream:
    - Transcribe the audio
    - Label transcriptions per individual speaker

Implementation:

- Fusion graph via JSON to combine the multiple inference models
- gRPC end points for direct interaction with the inference models
- Jupyter notebook demonstrates Python APIs for interaction

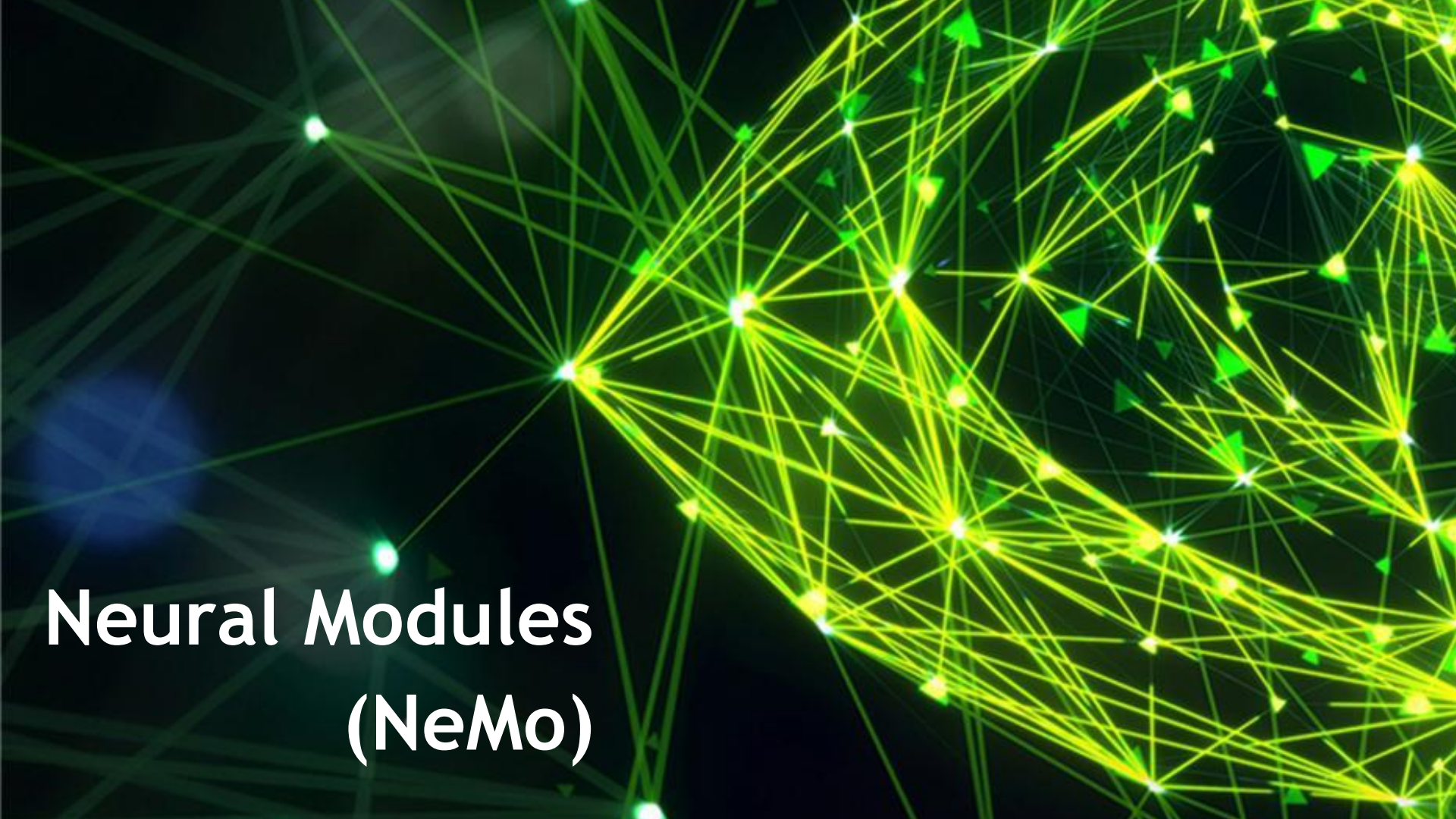


Transcription
Driver: Where is a good sushi restaurant?
Passenger: What’s the weather in Chicago

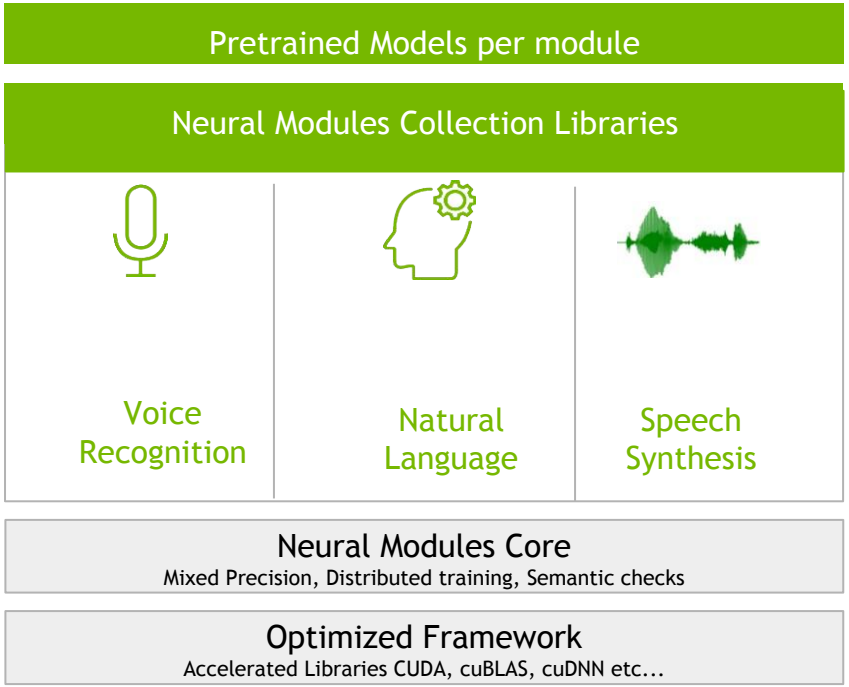
Gaze & Speech
https://www.youtube.com/watch?v=r264lBi1nMU

Model Developer: Improve the conversational model accuracy via fine-tuning with NeMo

Developer Operations: Deploy via docker containers from NGC into Kubernetes (EGX)


Neural Modules
(NeMo)

NEMO: TRAINING CONVERSATIONAL AI MODELS



- Open source deep learning Python toolkit for training speech and language models
- High performance training on NVIDIA GPUs
  - Uses TensorCores
  - Multi-GPU
  - Multi-Node
- Based on concept of Neural Module – reusable high level building block for defining deep learning models
- PyTorch backend (TensorFlow on Roadmap)

NEMO COLLECTIONS

pip install nemo_asr

nemo_asr
(Speech Recognition)

- Jasper acoustic model
- QuartzNet acoustic model
- RNN with attention
- Transformer-based
- English and Mandarin tokenizers and dataset importers

pip install nemo_nlp

nemo_nlp
(Natural Lang Processing)

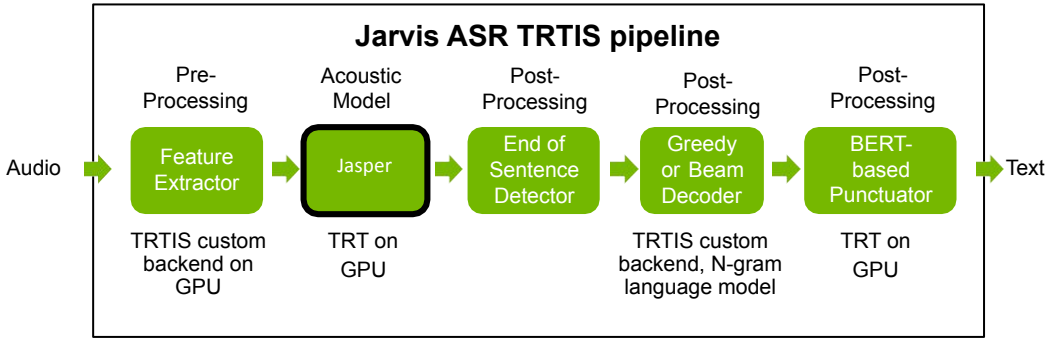
- BERT pre-training & finetuning
- GLUE tasks
- Language modeling
- Neural Machine Translation
- Intent classification & slot filling
- ASR spell correction
- Punctuation
- English and Mandarin dataset importers

pip install nemo_tts

nemo_tts
(Speech Synthesis)

- Tacotron 2
- WaveGlow
- English and Mandarin output and datasets importers

Jarvis ASR Service



Jarvis ASR API

Method Name	Description
Recognize	Given audio file as input, return transcript
StreamingRecognize	Process audio from a file or a microphone as it's being captured, returning partial transcripts

A FEW EXTRAS

Jasper:

A CNN based acoustic model, developed by NVIDIA.

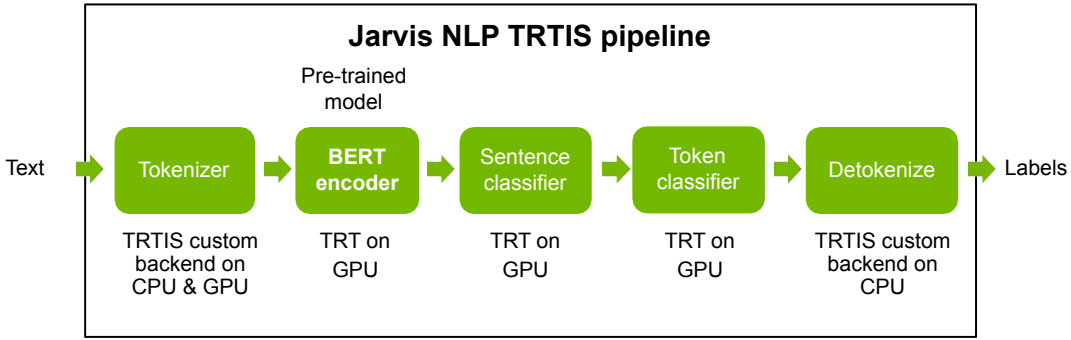
It used to represent the statistical relationship between an audio signal and the distinct sounds that makes words in a language.

Decoder (Beam/Greedy):

Once the acoustic model provides the statistical options for words, we'd like to construct a sentence out of them - using decoders.

- Greedy: The greedy decoder treats every word independently, as in each step it selects the word with the highest probability.
- Beam: Unlike the greedy one, the Beam Search decoder taking it's previous steps into statistical consideration. In fact, the decoder maximizes the *log-likelihood* of the output sequence.

Jarvis NLP Service



Jarvis NLP API

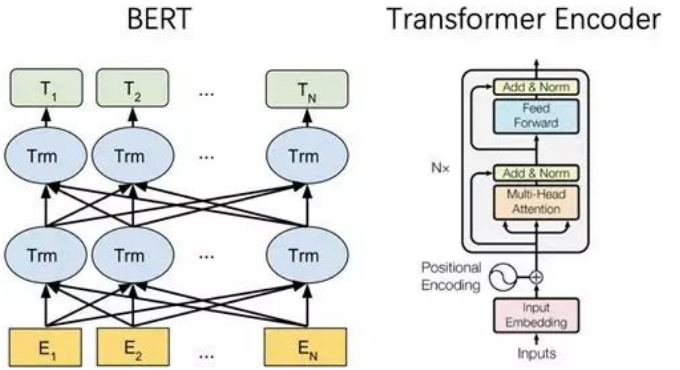
Method Name	Description
ClassifyText	Given text input, return a class label and score
ClassifyTokens	Given text input or array of tokens, return a class label and score per token
TransformText	Given input text, return output text

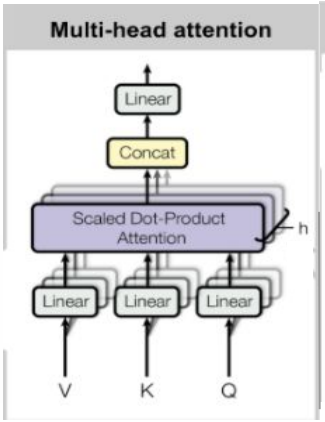
Jarvis NLP Provided Models

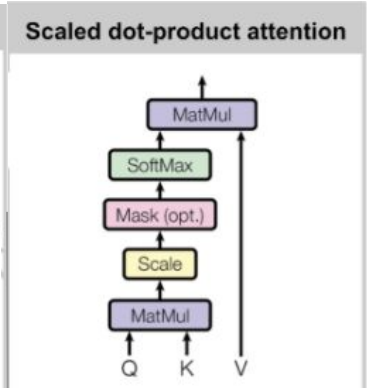
Method Name	Description
AnalyzeSentiment	Run sentiment detection on input and return label/score
AnalyzeEntities	Given text input, return named entities (NER)
Punctuate	Take text without punctuation (e.g. ASR output) and add periods, commas, question marks

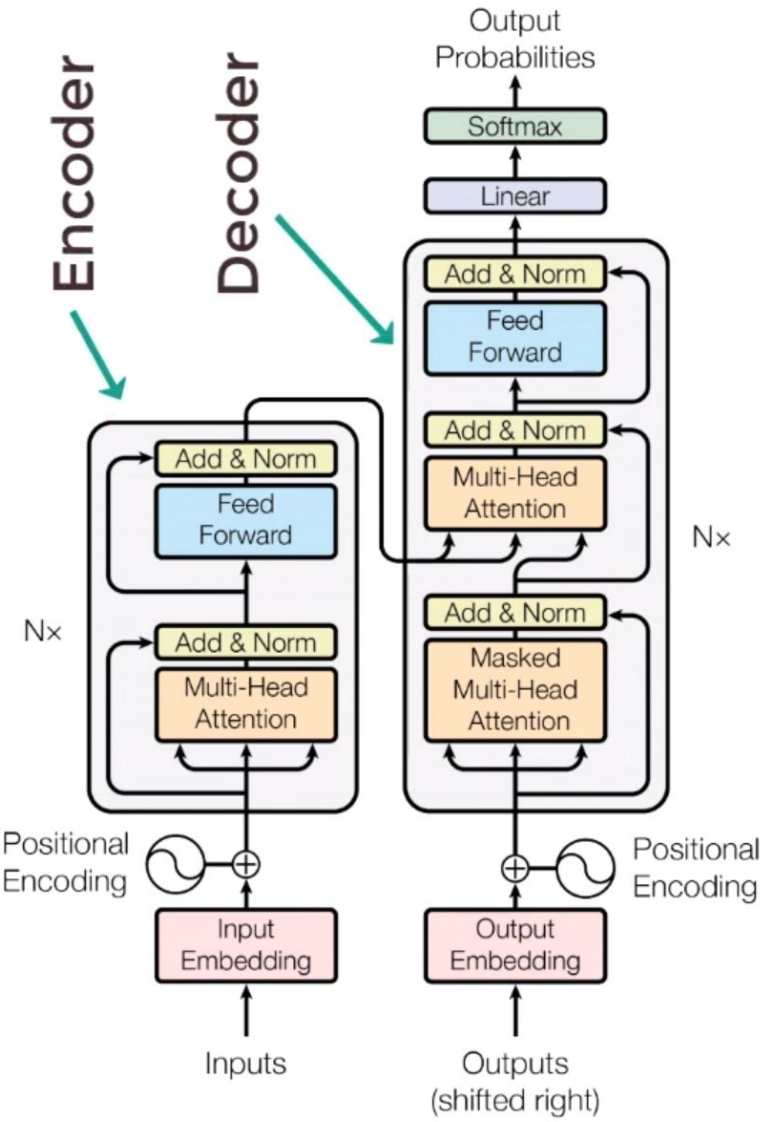
BERT

- BERT is a state-of-the-art, Transformer based model, developed by Google.
- The model is pre-trained from unlabeled data from English Wikipedia and BookCorpus (a large-scale text corpus, which is no longer distributed).
  - BERT has been proven as capable of high quality performances in many NLP tasks, such as sentence similarity, sentiment analysis and many more.

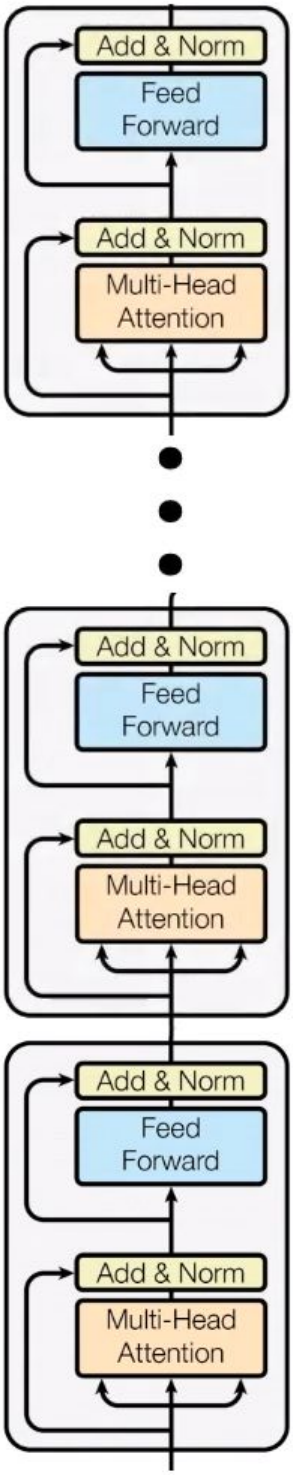




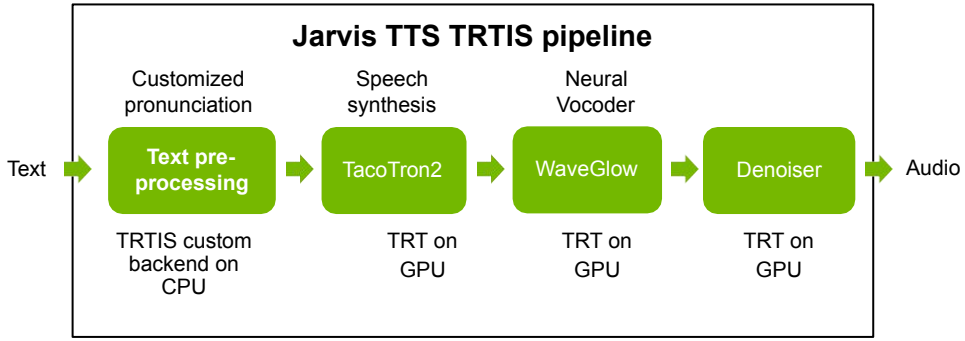



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Encoder
Decoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

BERT


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention

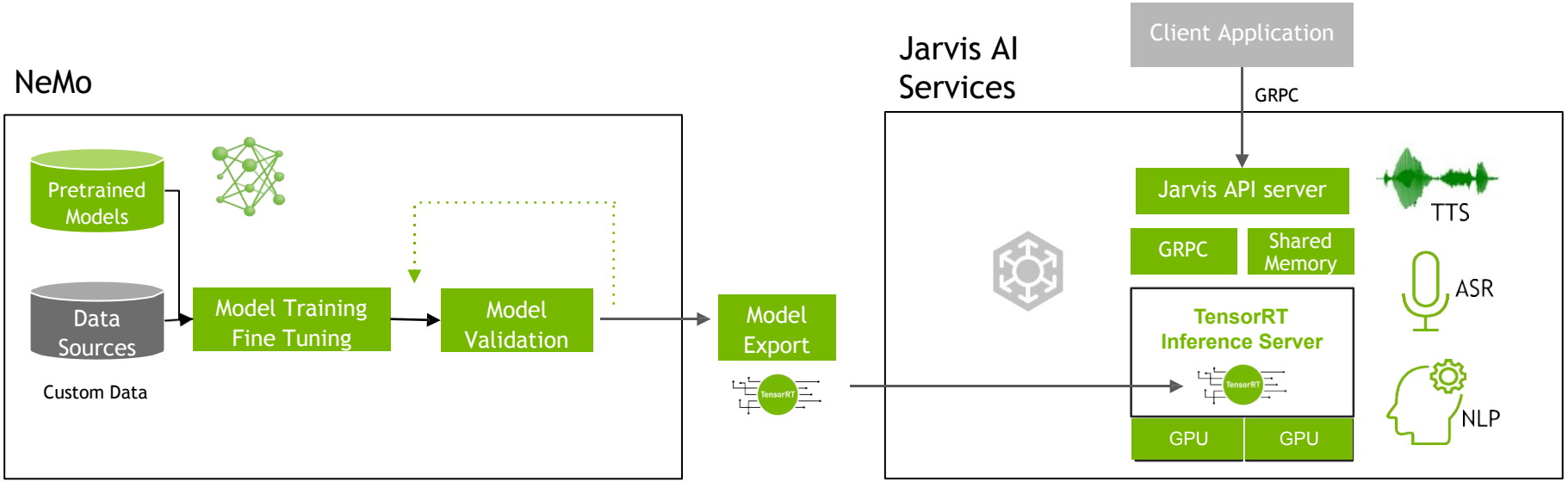
Jarvis TTS Service



Jarvis TTS API

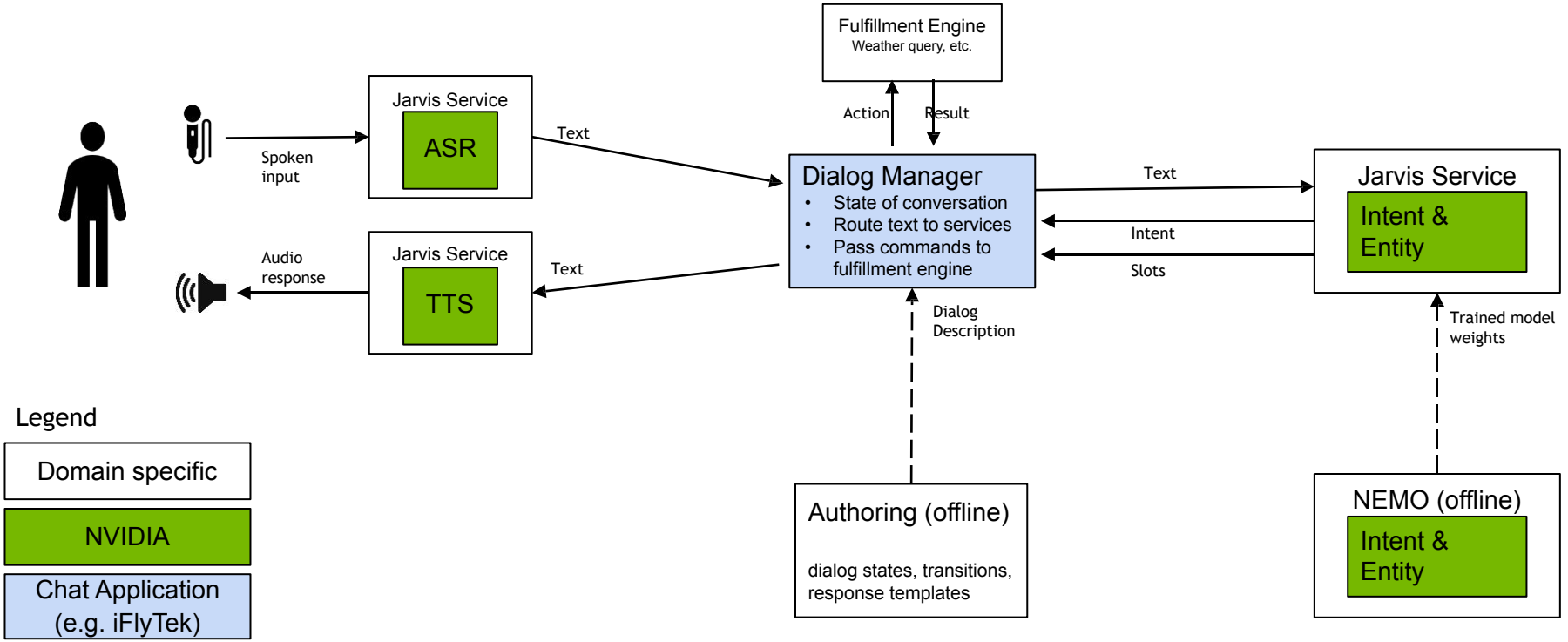
Method Name	Description
Synthesize	Given text input, return audio of spoken version as a single audio clip
SynthesizeOnline	Given text input, return audio of spoken version as an audio stream

JARVIS AND NEMO TOGETHER



Jarvis - Weather Bot Architecture

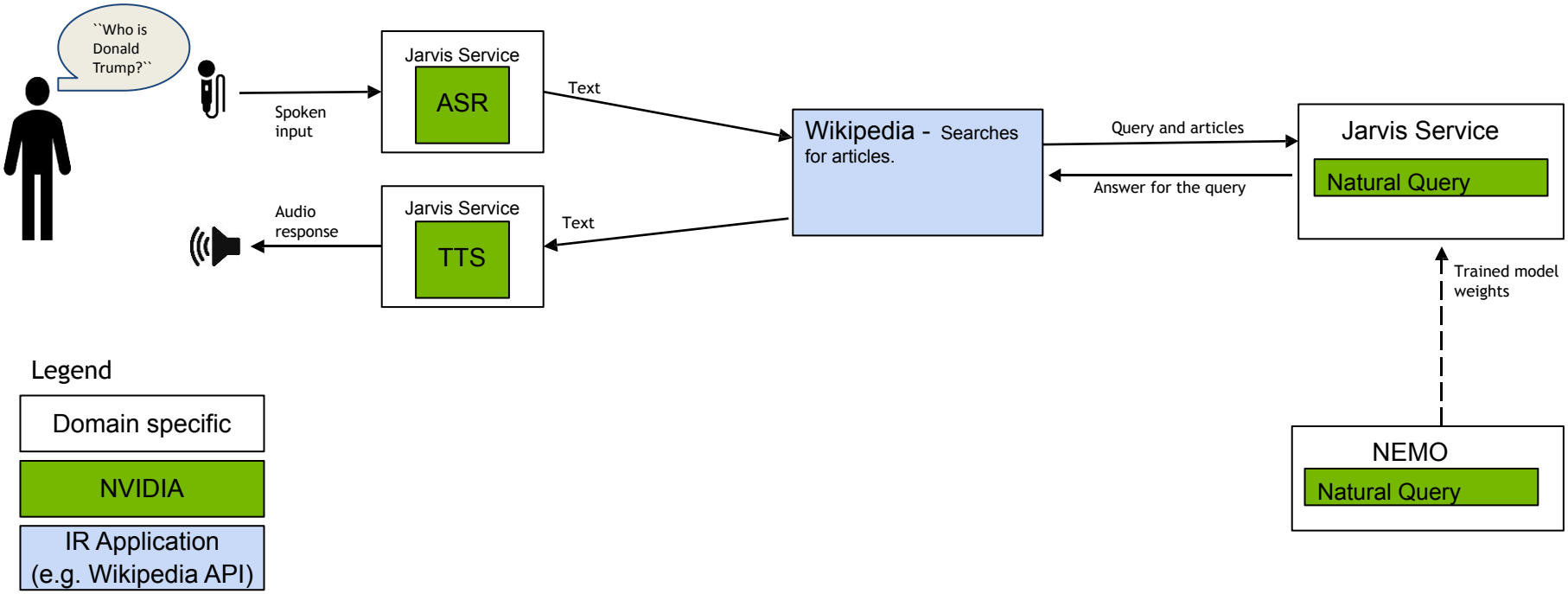
Deployment of Jarvis components with simple dialog manager




Jarvis
Chatbot

Jarvis - Chatbot

Application's Workflow



Jarvis - Chatbot

Application's Architecture