Healthcare Services Transformation
Deep Learning Use Cases

- Dima Rekesh, Sr. Distinguished Engineer
- Julie Zhu, Distinguished Engineer/Chief Data Scientist

- GTC 2019, Mar 20, San Jose

Agenda

- Deep Learning in Healthcare
- Poly Chronic Predictions and Automations
- Speech and NLP Use Cases
- Using the OpenSeq2Seq framework

OPTUM™

Deep Learning Transformational Capabilities in Healthcare

ASR and NLP

- Live and faster than real-time speech to text
- Call center optimization
- Disease identification and prediction in speech

Poly Chronic Care

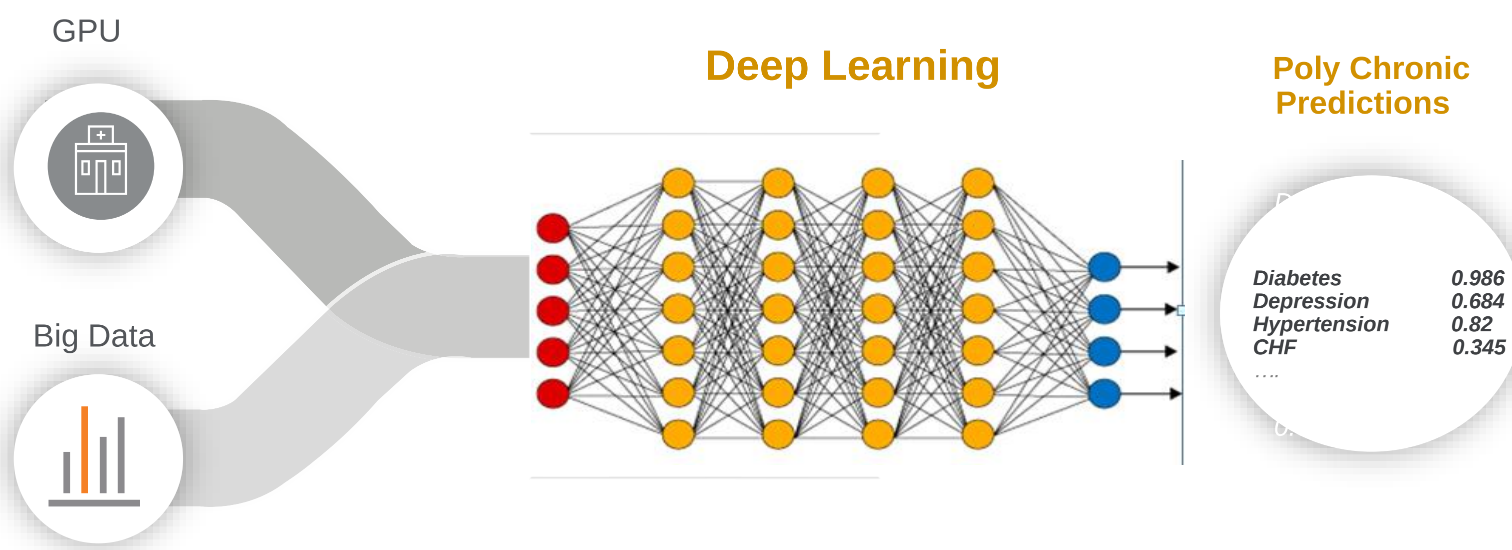
- Predict poly chronic conditions to identify high risk patients
- Evidence Care
- Personalize treatment recommendations.

Operation Automation

- Prior authorization, fraud identification and auto-adjudicated claims processes
- Resources and staffing optimization to adapt and manage fluctuations in need

OPTUM™

Poly Chronic Disease Predictions



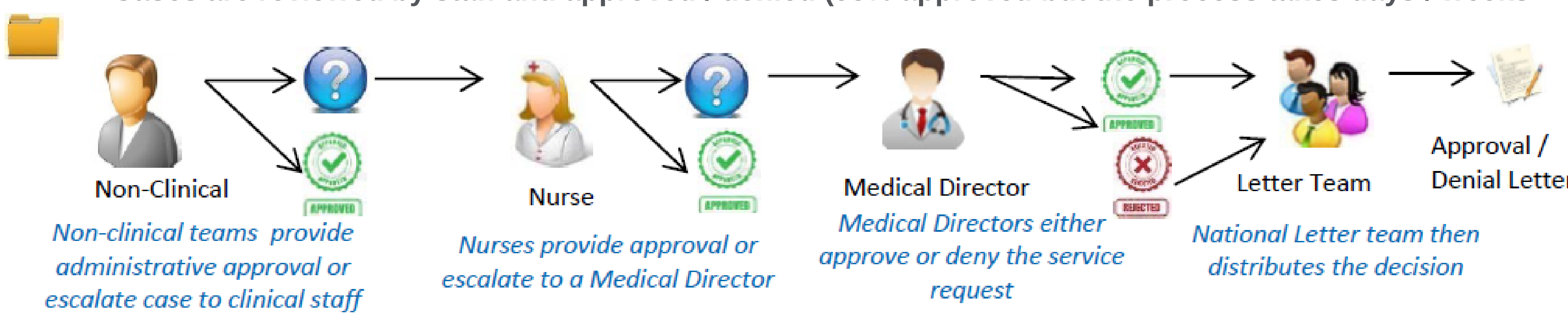
- **Building a sustainable network to predict the poly chronic conditions and high risk patients.**
- **Provider Evidence-base care to optimize the treatment.**
- **Multidisciplinary intervention for preventions and reduce the cost.**

Prior Authorization Workflow

Challenge

- Prior authorization is the process by which health care providers obtain advanced approval for a procedure, service or medication to confirm coverage by the health plan.
- Necessary to ensure appropriate care is provided and to mitigate overutilization
- Prior authorization processes are often disruptive for both providers and patients resulting in inefficiencies and reduction in time spent providing care

Cases are reviewed by staff and approved / denied (95% approved but the process takes days / weeks

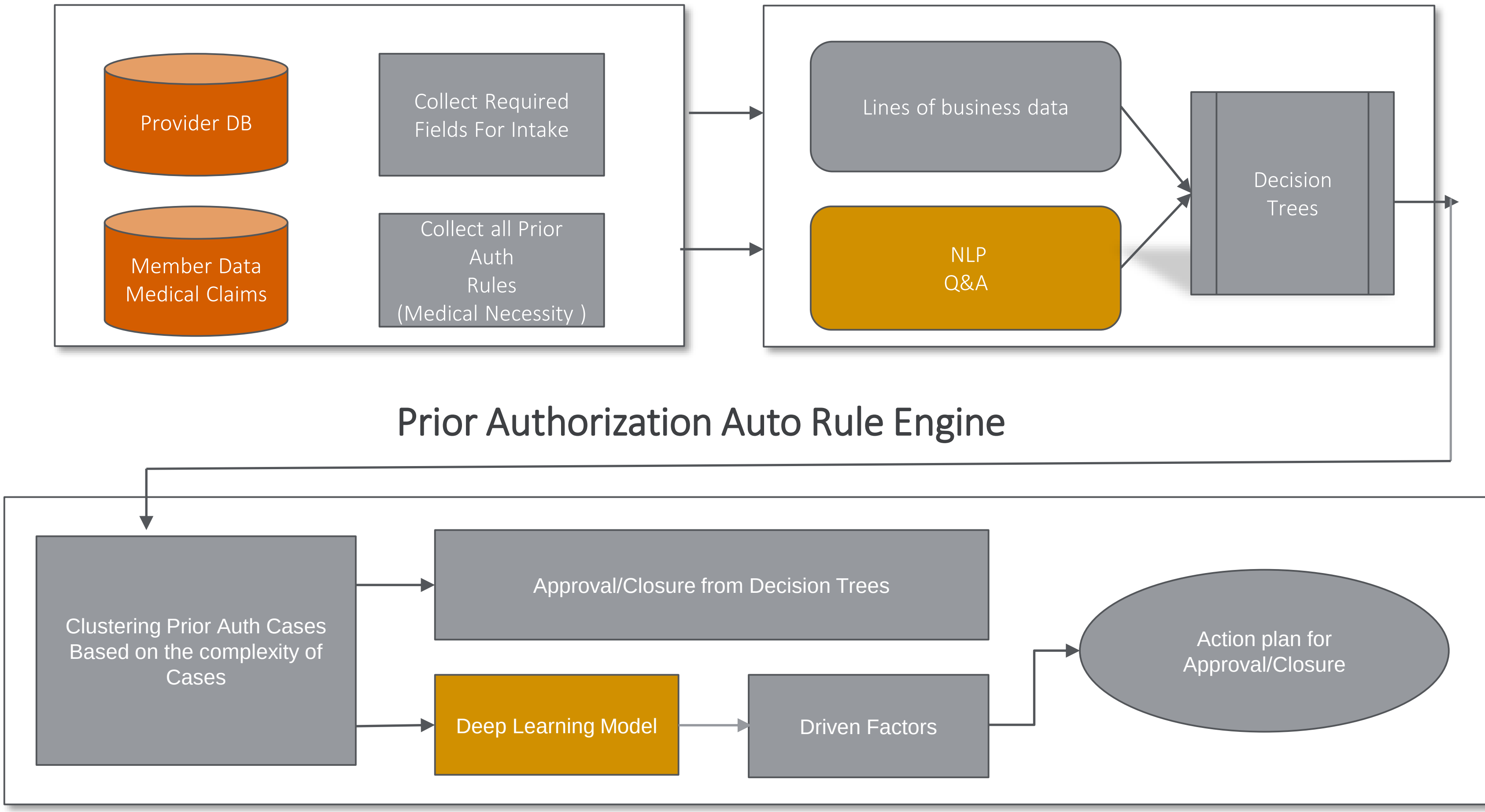


Business Benefits:

- Authorizations now take seconds, instead of days or weeks, and 95% of cases are approved.
- Payers can improve member satisfaction for Prior Authorization process using AI (specifically, Deep Learning) to accelerate Prior Authorization approvals.

Prior Authorization Automation

Prior authorization processes contain both routine components and components of high variability and complexity.



Prior Authorization Clinical Outpatient Approval Prediction

Deep Learning Model Results

Model	Data	AUC	Accuracy(%)	Sensitivity /Recall(%)	Precision(%)	F1-Score(%)
Neural Networks	Train	95	92	97	94	95
	Test	94	91	97	93	95

XGBoost Model Results

Model	Data	AUC	Accuracy(%)	Sensitivity /Recall(%)	Precision(%)	F1-Score(%)
XGBoost	Train	87	89	96	91	94
	Test	87	89	97	91	94

OPTUM™

7

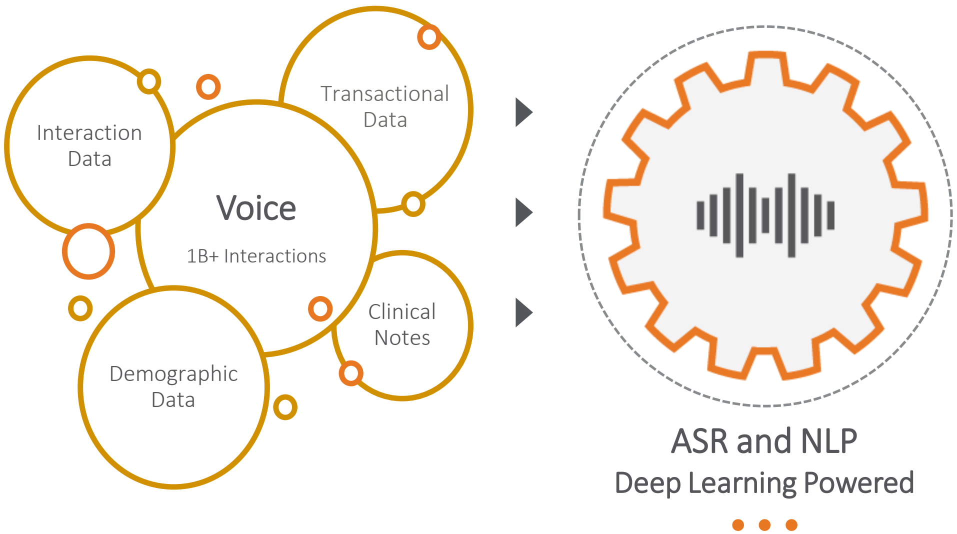
Prior Authorization Process Automation by Deep Learning Prediction

Procedure Description	Model Prediction
Continuous positive airway	0.99
Physical therapy	0.99
positive airway pressure device	0.99
glucose monitor	0.99

The key successes of production is to monitor the process and automate the modeling updates.

OPTUM™

ASR/NLP & Business Capabilities



Recognition and Response

Understands conversational interaction and speaker's intent. Provide automated response or informed guidance to advocate.

Informed Action

Accurately identifies or predicts issues or opportunities such as customer satisfaction or likelihood of repeat calling. Recommends appropriate response in the real time of an interaction.

Clinical Insights

Linkage between human speech production and neurodegenerative disease types, cognitive impairment, and behavioral and mental health can assist in their detection and ongoing treatment.

Speech Recognition

Feature Parameterization

Transcription Normalization

Phonetic Indexing

Natural Language Processing

Named Entity Extraction

Sentiment and Topic Classification

Text Summarization

OPTUM™

Contact Center KPIs

Average Handle Time Reduction

- *"I don't want to be on the phone longer than I have to be!"*
- AHT reduction achieved through accurate caller identification and authentication, recognizing caller intent, efficient call routing, reaching the right agent, avoiding transfers, eliminating unnecessary holds, etc.

First Call Resolution Improvement

- *"I shouldn't have to call 3 times to get the answer I need"*
- FCR increase achieved by identifying and fixing gaps in processes, reaching the right agent the first time, identifying (and training to) gaps in knowledge management, identifying repeat callers and agent education, etc.

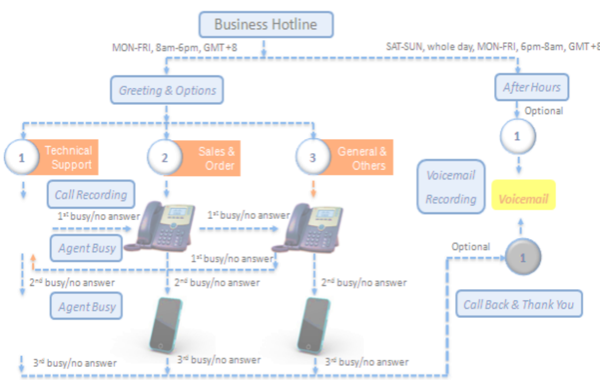
Call Volume Reduction

- *"How did I end up here? I thought I was waiting for a licensed agent"*
- Call volume reduction achieved by eliminating unnecessary and misdirected calls, self-service (bots/digital assistants) to answer questions that don't need an agent, identifying how and why customers are getting to the wrong places

Understanding Caller Intent

Challenge

- Poor speech recognition in IVR remains one of the largest contributors to long handle times, multiple transfers and caller frustration
- A hierarchy of IVR prompts feels slow, complex and outdated compared with today's open speech experiences such as Siri, Cortana and Alexa
- Current IVR speech recognition models require manual training for any improvement or change – a lengthy and costly activity



Proposal: *Streamline the call experience through Deep Learning*

1. Identify and authenticate the caller
2. Understand the reason for the call
3. Optimize the call routing
4. Personalize the interaction
5. **Self-Learn**

Applying OpenSeq2Seq

ASR and NLP Use Cases for Deep Learning

Disease Predictors

Use the member's voice to help screen for disease. The voice analysis technology would supplement the nurse's "human" decision making, with the expected result of increasing early intervention to improve clinical outcomes.

VoiceMail Assistants

Voicemail Assistant transcribes and extracts essential information for clinician's review and processing and facilitates submission of appropriate content into backend system. Clinician productivity significantly increases by reducing manual transcription and content entry thus allowing clinician to focus on patient care.

Wellness Visit Predictors

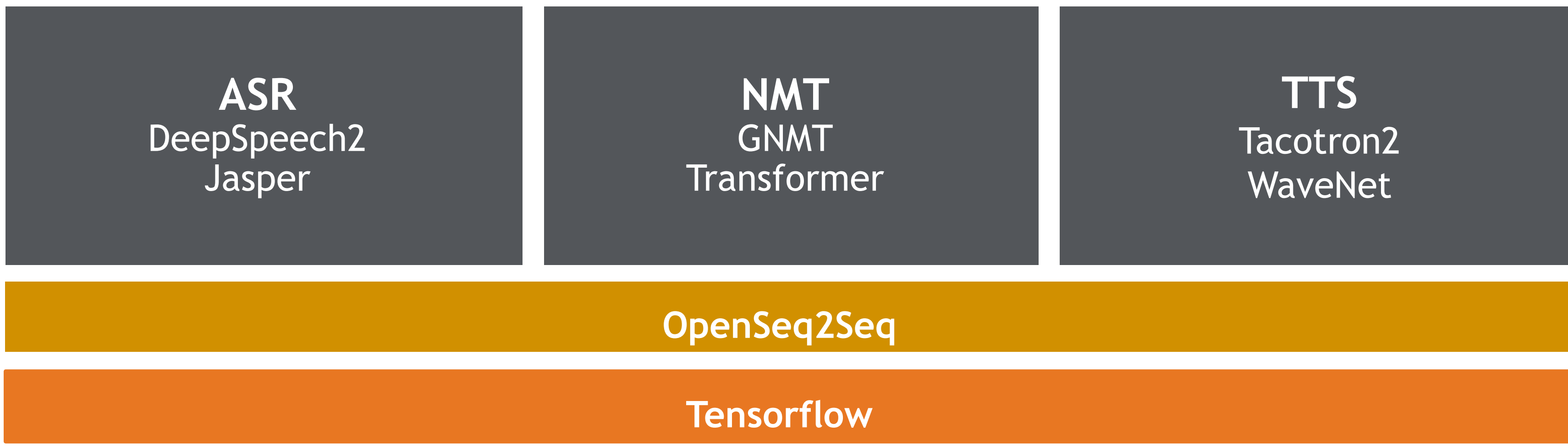
There is a limited understanding of the key factors that motivate a member to schedule a wellness visit, most notably, there is a lack of visibility of those features that occur during the course of the advocate-member voice interaction. Immediate focus is to increase member acceptance rate by identifying the most impactful and persuasive conversational features and assist advocates in making beneficial adjustments to that dialogue.

OPTUM™

OpenSeq2Seq

NVIDIA Research sequence-to-sequence framework for speech*

- Toolkit for for distributed and mixed-precision training of Seq2Seq models
- Pre-defined (and growing) support multiple encoder-decoder model



*Reference: https://arxiv.org/abs/1805.10387, https://nvidia.github.io/OpenSeq2Seq

OpenSeq2Seq

Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, Paulius Micikevicius

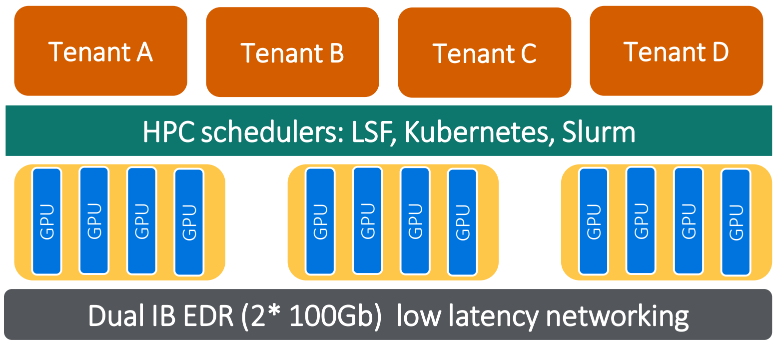
- NVIDIA Research team in Deep Learning Software Development
- Working on DL algorithms for Fast, Scalable training
- Research areas:
- NLP, speech recognition, text-to-speech
  - Recommender systems
  - Auto-ML
  - Large batch training
  - Low precision training

OPTUM™

The Infrastructure

Deep Learning Model Design and Training

HPC style cluster

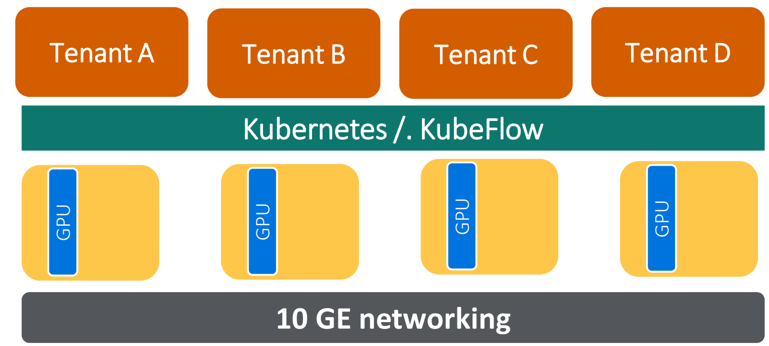


- 10-100 times more computationally demanding
- Tenants can experience the power of the entire cluster (queue based)
- Distributed DL jobs can span the entire cluster
- Nodes act as one [super-] computer

Challenges: full support for docker in HPC schedulers, low latency overlay networking

Deep Learning Model Deployment

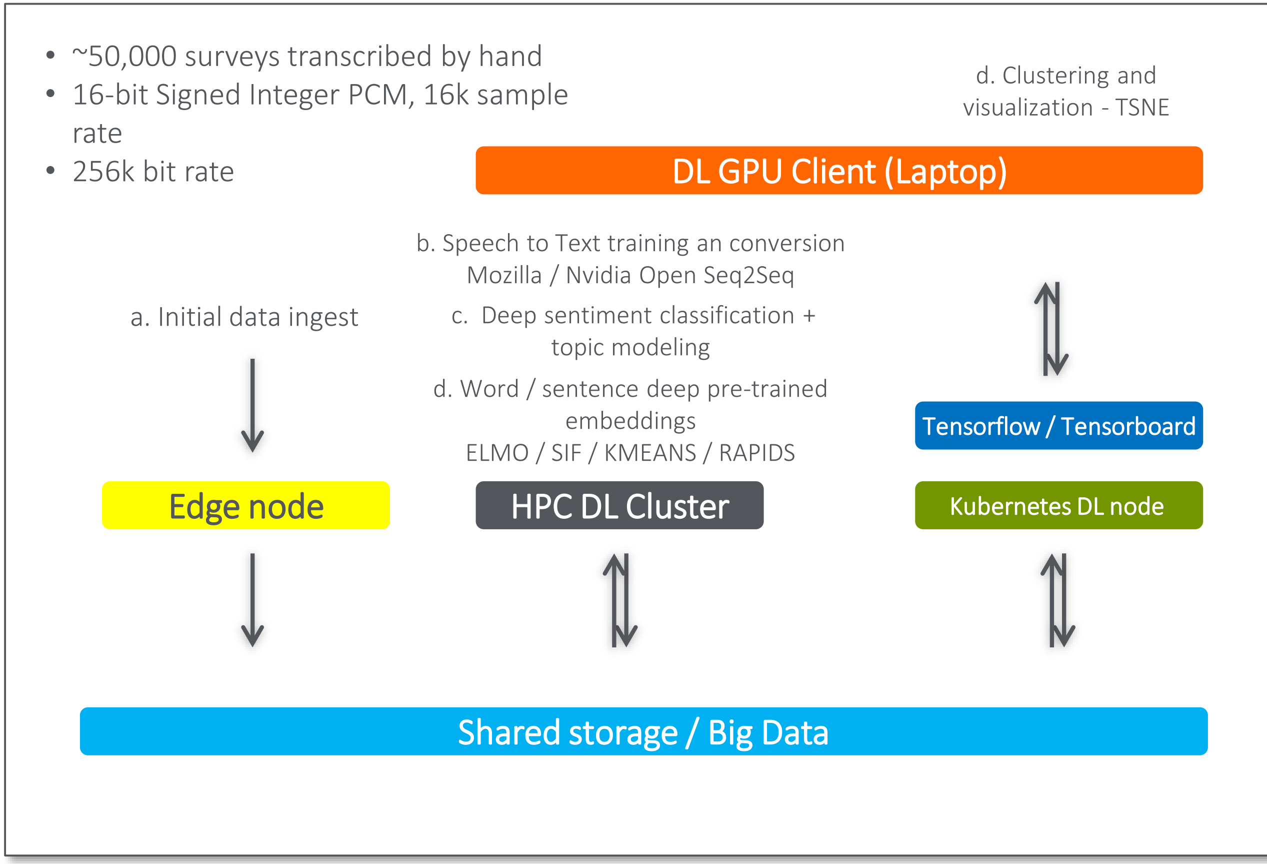
Enabling GPUs as first class citizens to existing micro-services / big data clusters



- Share resources of the cluster among tenants.
- DL Jobs are typically contained in one node or require a small portion of the cluster
- Nodes are relatively weakly connected

Challenges: user access controls and enforcing limits / quotas on groups of users in Kubernetes / queues

GPU Clusters working together: sample end-to-end flow



OPTUM™

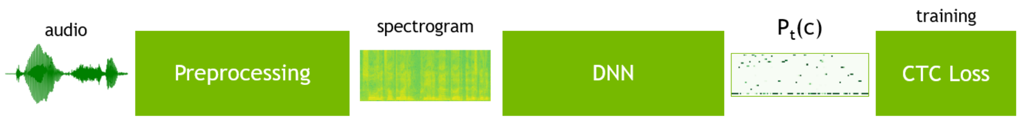
OpenSeq2Seq: Training

- In LSF, and in Kubernetes clusters
- In containers, but externalize the source and config files for ease of tweaking
- Looking at the same storage, allowing for TensorBoard
- Building containers: requires a powerful a server, currently..
  - KenLM requires building TF, which takes a long time
  - GPU Direct, IB: out of the box container compatibility issues
  - LSF vs Slurm: OpenMPI needs to be recompiled

OPTUM™

Networks for Speech Recognition

Audio pre-processing and Deep Speech 2



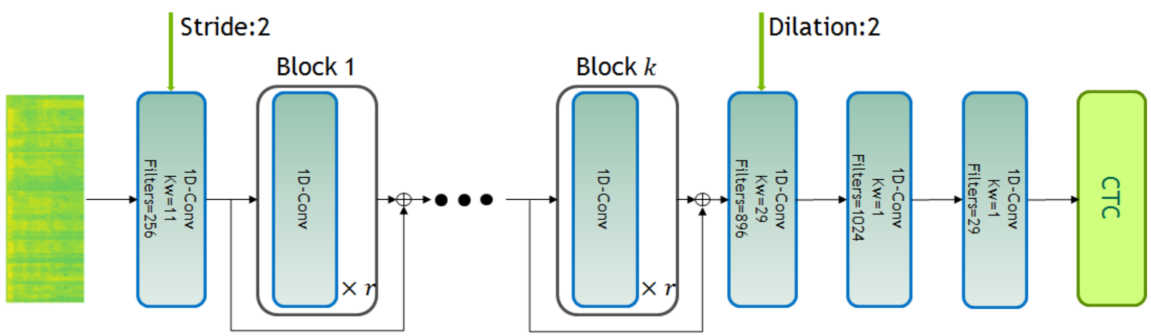
- Data augmentation
  - additive Gaussian noise
  - time stretch (resampling)
- Windowing 20-25 ms, stride 10 ms
- FFT, log
- Normalization

- 2-3 convolutional layers
  - ch=32, ks=[11, 41], s=[2, 2]
  - ch=32, ks=[11, 21], s=[1, 2]
  - ch=64, ks=[11, 21], s=[1, 2]
- 3-7 bi-/uni- directional GRUs/LSTMs
- 1 row conv layer (for unidirectional)
- 1-2 fully connected layers

- BN/dropout for regularization

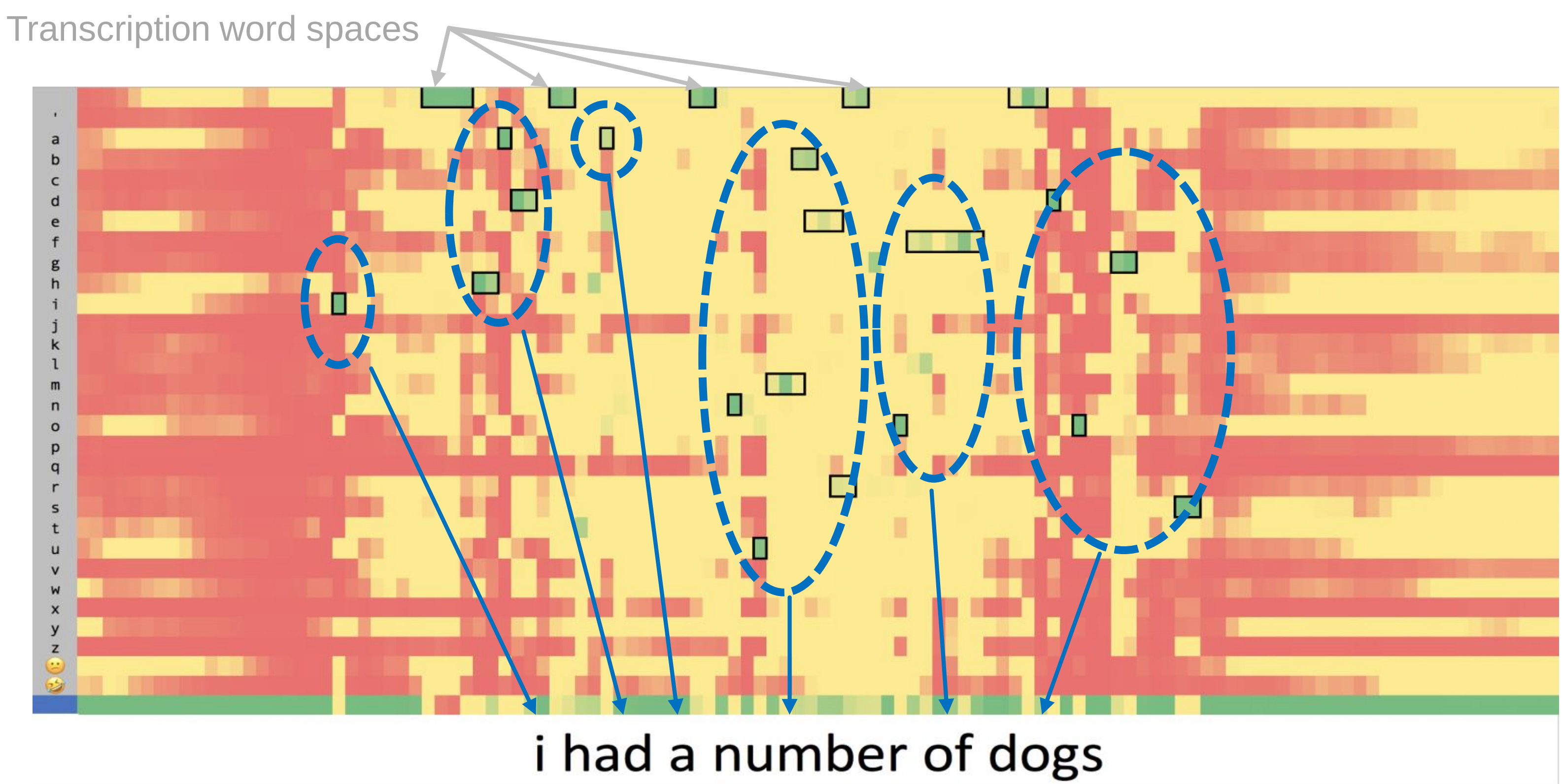
- greedy (argmax)
- beam search
- beam search with language model (width=2000-8000)

Jasper



https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition/deepspeech2.html
https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition/jasper.html

Decoding the output and the CTC loss



- “Greedy” decoding simply takes the most likely symbol in each position – it’s fast
- Non-neural language models run on CPUs today and are slow
  - The default KenLM implementation today runs on just 1 CPU core.

Scaling: Deep Speech 2

Scales well over GPU Direct / RDMA

Intra-node check: Horovod does not affect the speed of training

# nodes	0.75	1	1	1	1	2	2	3	3	4	4	5
job #	9475	9470	9444	9449		9485	9445	9446	9451	9447	9452	9448
hvd?	yes	yes	yes	yes		yes	yes	yes	yes	yes	yes	yes
optimizer	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum	Momentum
lr	0.001	0.001	0.001	0.001	0.004	0.001	0.001	0.001	0.001	0.001	0.001	0.001
lr power	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Avg time /step	1.618	1.759	1.624	1.752		1.746	1.687	1.729	1.766	1.767	1.794	1.828
Avg objects / s	36584.25	44871.25	48598.88	45048.24		90366.497	93557.64	136954.4	134049.2	178616.3	175939.1	215843.6
total run time, s	24874	26880	24511	26826	27187	26586	25453	26097	26737	26675	27141	27652
total run time, hrs	6.91	7.47	6.81	7.45		7.39	7.07	7.25	7.43	7.41	7.54	7.68
avg objects /s		45972.3233				91962.069		135501.801		177277.723		215843.579
scaling factor						1.0		1.0		1.0		0.9

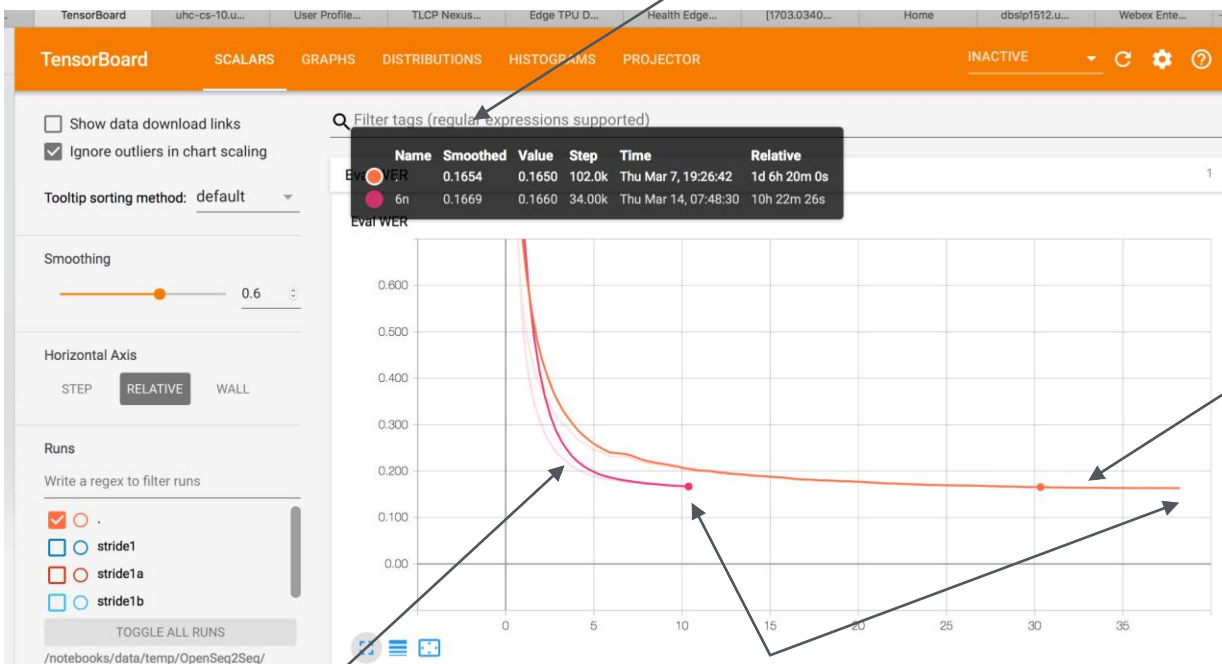
Consistent runtimes

Consistent scaling

Scaling: Jasper (10x5)

Scales well over GPU Direct / RDMA

WER is almost identical



2 nodes, 8 GPUs

About 3x faster, just 10 hours

6 nodes, 24 GPUs

https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition/jasper.html

Data Preparation – Audio

The basics are reasonably simple

- The data is already collected, compressed, stored in a variety formats
- Size matters: just make sure it's collected all in one place!
- OpenSeq2Seq currently requires signed PCM 16 bit integer sampled at 8 kHz or 16kHz
- Single channel
- One speaker
- Non-English languages may be present! How do you find them?
- Conversion: sox in docker, etc.

Data preparation – transcripts

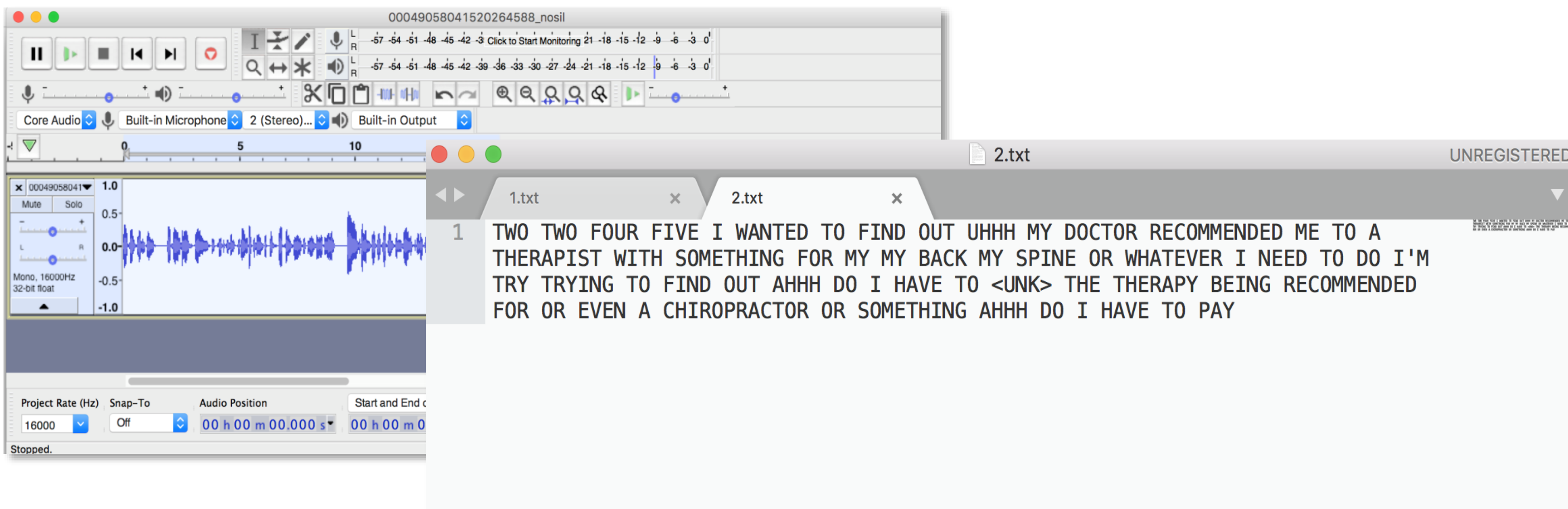
A little less simple

- The data is already transcribed according to a variety of [best] practices
- OpenSeq2Seq allows for a vocabulary file
- If using a checkpoint, helpful to be consistent to what the checkpoint is trained on (librispeech)
  - To lowercase
  - Keep letters only, remove punctuation, spell out numbers (5 => “five”)
- Replace voice tags with spaces (<unclear> <unk> => ”space”)
- List of suppressed words
- Collapse multiple spaces

Handling Longer Files

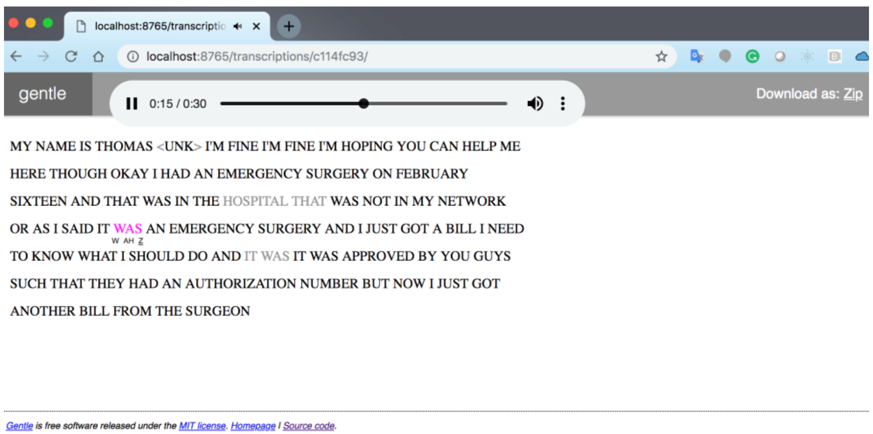
DeepSpeech 2 and especially Wave2Letter do not train well or at all on longer fragments (e.g. 60s), so

1. Just filter the longer fragments out?
2. Split the files [by hand?]



Splitting and aligning audio [semi-] automatically

Gentle



- Alignment is a task worthy of a small cluster – e.g. ~1,000 core-hours for a 1k hour dataset
- Split between detected words with an offset, so that the fragments overlap
- With Jasper, 8s appears to work well


```
{
  "alignedWord": "i",
  "case": "success",
  "end": 2.8600000000000003,
  "endOffset": 19,
  "phones": [
    {
      "duration": 0.12,
      "phone": "ay_S"
    }
  ],
  "start": 2.74,
  "startOffset": 18,
  "word": "I"
},
{
  "alignedWord": "am",
  "case": "success",
  "end": 2.94,
  "endOffset": 22,
  "phones": [
    {
      "duration": 0.01,
      "phone": "ae_B"
    },
    {
      "duration": 0.06,
      "phone": "m_E"
    }
  ],
  "start": 2.87,
```

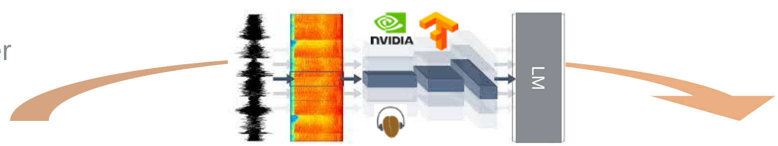

- Aligns by word
- Based on Kaldi
- Likely needs to be combined with human curation

https://github.com/lowerquality/gentle

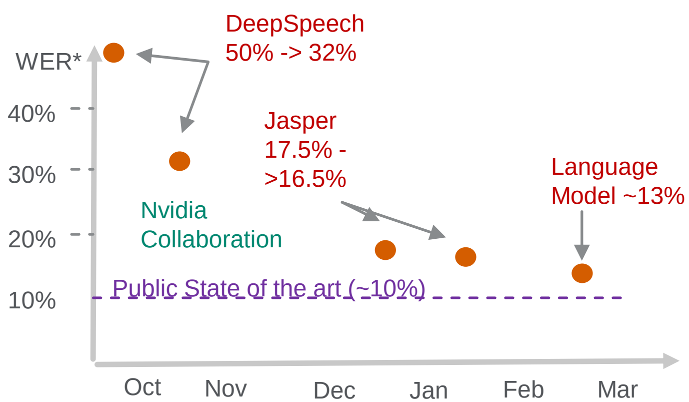
Transfer Learning form Public Data

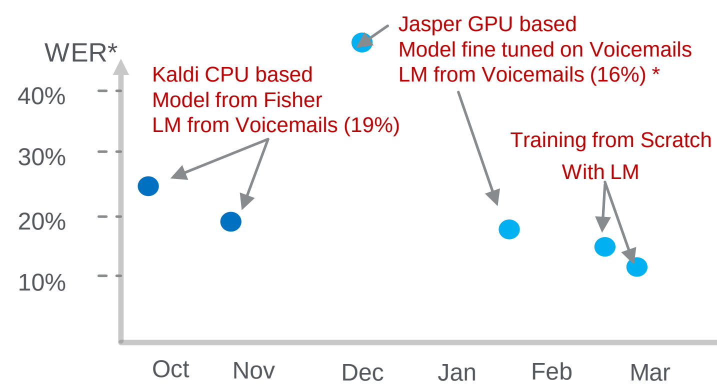
Fisher Data Set	Voice Mail Dataset
Public available data, acquired thru LDC	Optum UHG Data
2k hours, phone calls, noisy data	1k hours, phone calls, noisy data
Short segments (10s to 20s)	Long Segments (~3min) (…difficult for learning)
High quality transcriptions, Few medical terms	Poor quality transcriptions, Medical terms

Model learned on Fisher



Continue learning or directly
apply on UHG data
Looking for best approach





- *Currently benchmarking exactly Kaldi vs Jasper
- Improving Fisher & NursesVM in parallel; GPU capacity bottleneck

Transfer Learning

Often, best results, but can’t use directly

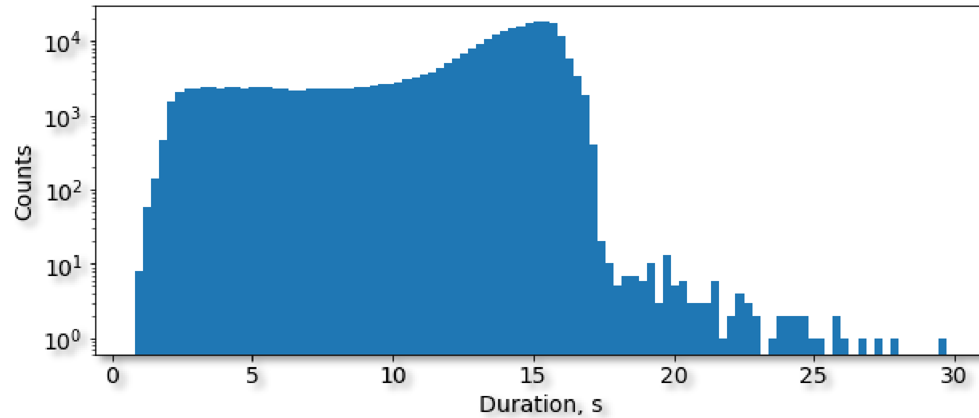
- Specify the checkpoint: train yourself or grab one from Nvidia
- Can’t use directly, resulting WER is not good.
- Add your own data set (can on the command line)
- Watch out: Learning rate won’t change unless you force it to!
- Watch out: number of epochs will be recalculated depending on your data set
- Adjust Dropout
- Adjust regularization

OPTUM™

Starting Point: LibriSpeech Data Set

One of the largest open voice data sets

- 1000 hours of audio
- Aligned
- Clean
- 99.9% < 16.7 s
- DeepSpeech2 : 6% WER
- Jasper: 4% WER



http://www.openslr.org/12/

OPTUM™

Starting point: Fisher data set

Noisy, telephone, accented speech

- Commercial
- 2742 hours
- 16,454 calls
- Female / male: 53/47

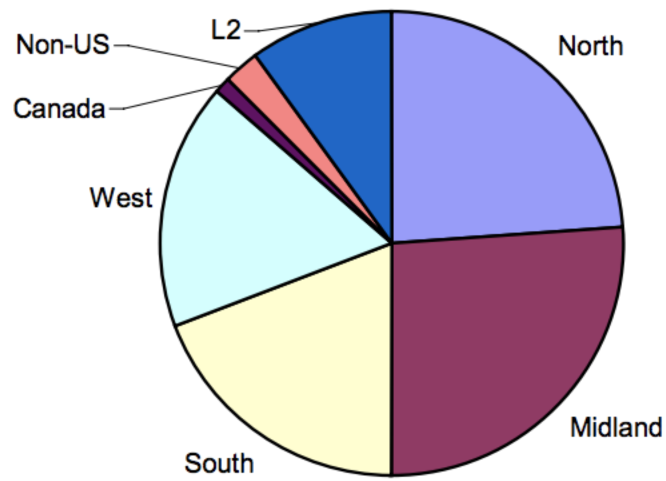


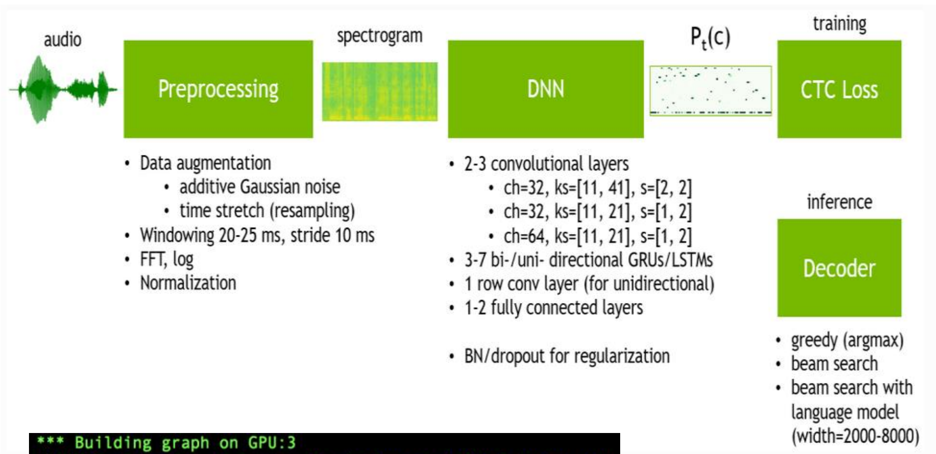
Figure 1: Regional/Dialect Distribution of Fisher Speakers

https://catalog.ldc.upenn.edu/LDC2004T19

Freezing layers

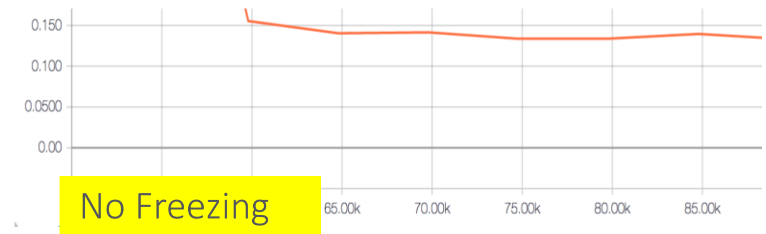
Now part of OpenSeq2Seq

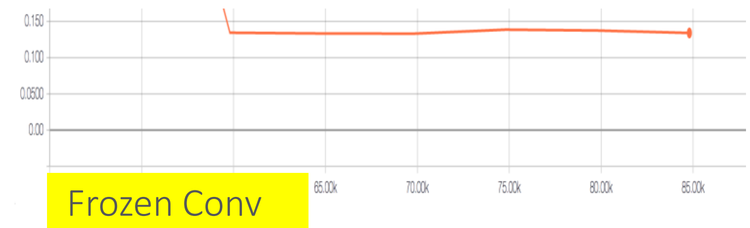
Deep Speech 2 Architecture




```
*** Building graph on GPU:3
WARNING: freezing of layers not implemented for Decoders yet.
Freeze variables...
 ForwardPass/ds2_encoder/conv1/kernel:0
ForwardPass/ds2_encoder/conv1/bn/gamma:0
ForwardPass/ds2_encoder/conv1/bn/beta:0
ForwardPass/ds2_encoder/conv2/kernel:0
ForwardPass/ds2_encoder/conv2/bn/gamma:0
ForwardPass/ds2_encoder/conv2/bn/beta:0
*** Trainable variables:
```

Eval WER





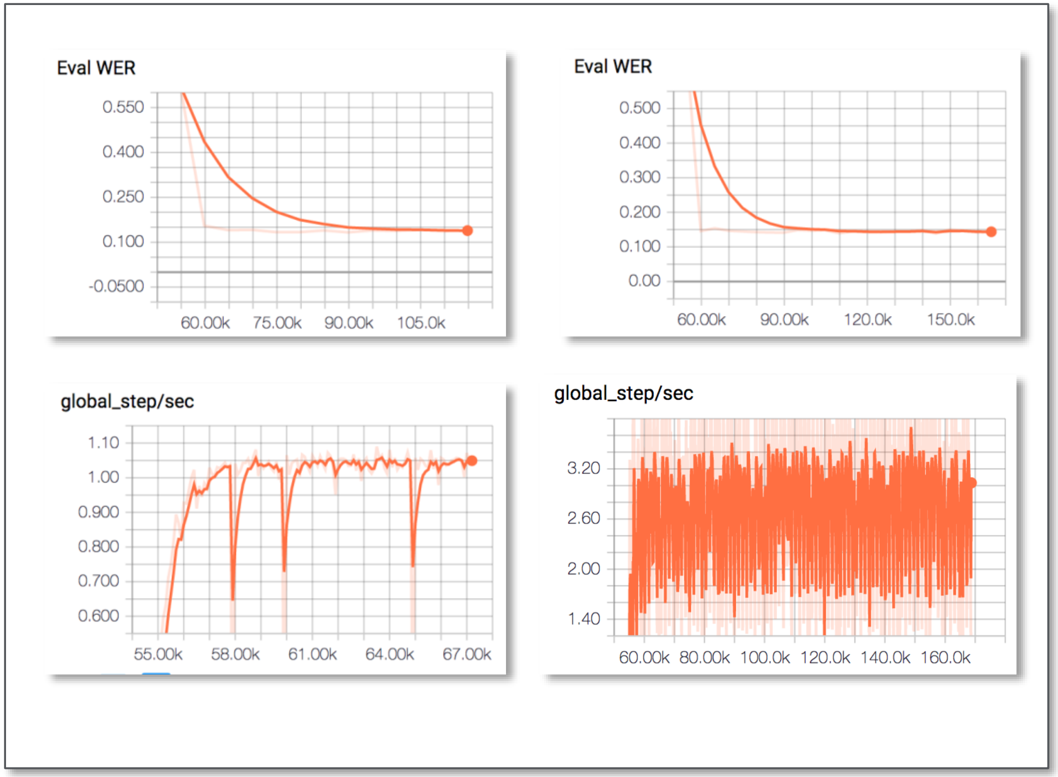
Sorted Batches

Make sure to turn shuffle off

Sorting batches on audio length lead to approx. 2X faster training, without minor impact on WER.

This minimizes padding in each batch, bay making batches have similar sized audios.

Good way to accelerate distributed training

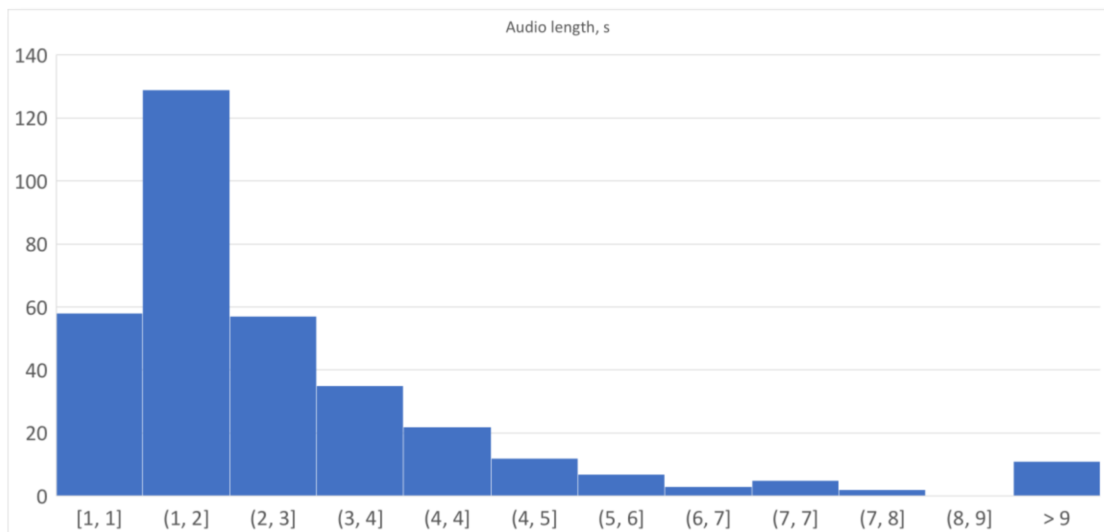


Example: an interactive voice response (IVR) data set

Short utterances direct the call path

- Single channel
- Single speaker
- “Greedy” WER: 17%
- [Ken]LM WER: 13%
- Massive overfit
- (add data on shorter vs WER)

Avg,s	Samples	Total, hrs
2.68	28776	21.4



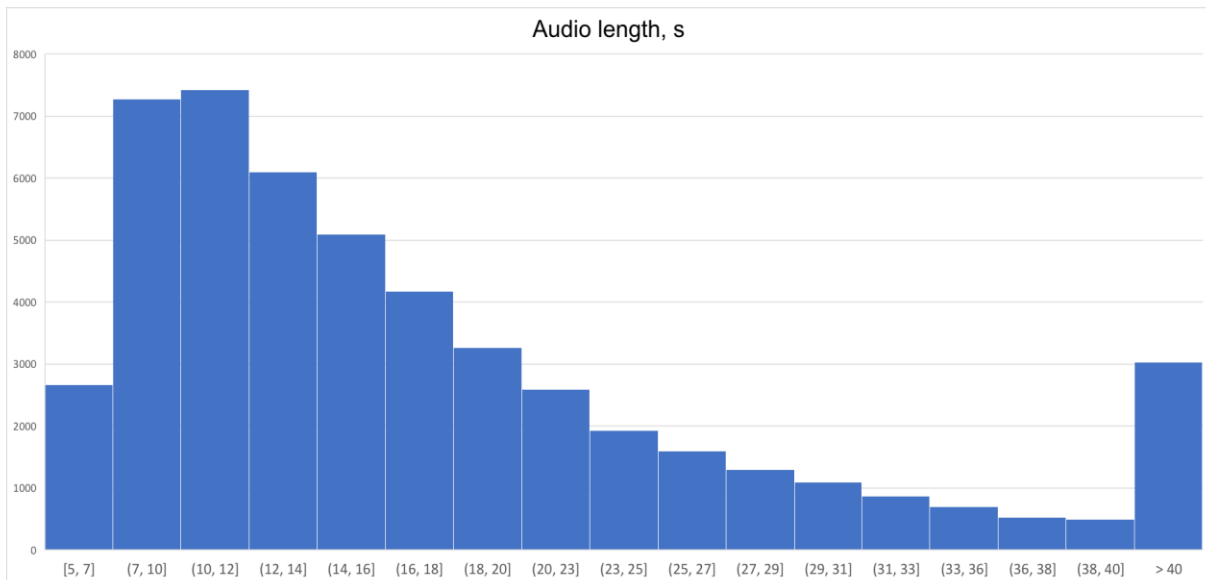
OPTUM™

Example: Surveys Data Set

Members evaluate quality of service received

- Single channel
- Single speaker
- Best result: fragments longer than 16.7s filtered out (92 hrs)
- Current WER: 17%
- (validate that number)?

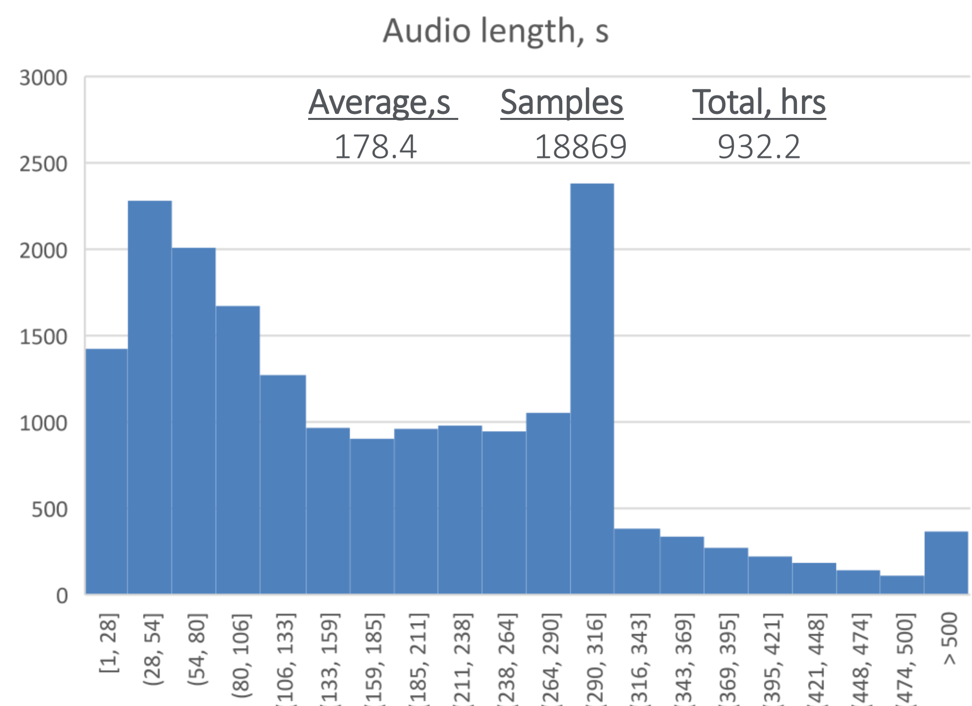
Average,s	Samples	Total, hrs
18.5	50114	257.1



Example: Voicemail data set

Voicemails regarding member condition

- Single channel
- Single speaker
- Must be split for training
- Current WER: 15.3% (no LM), < ~12% (KenLM)
- Explain more here



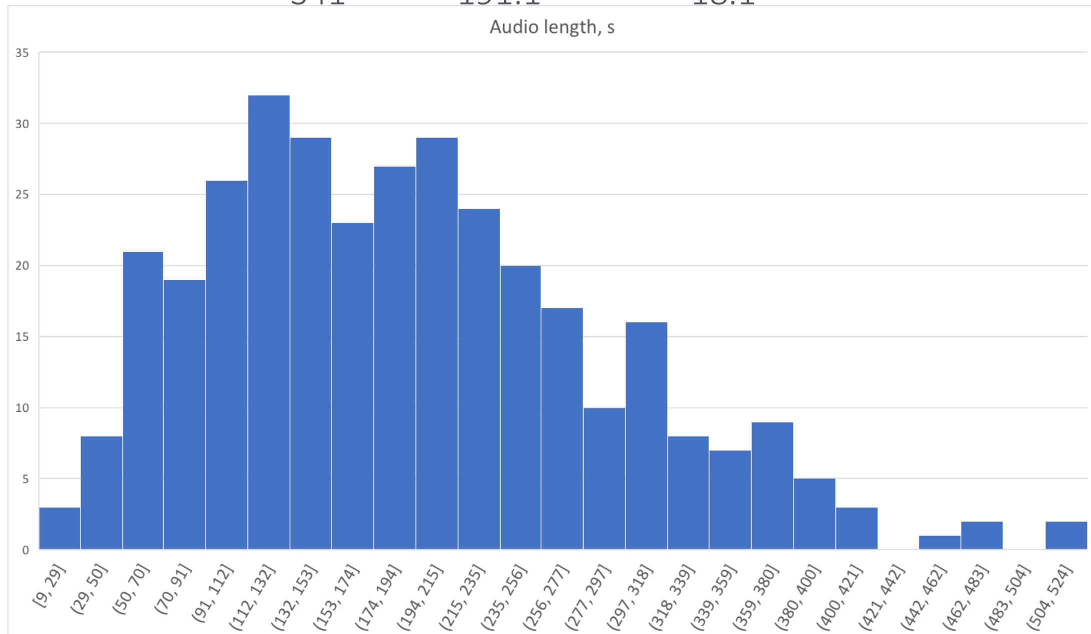
OPTUM™

Example: a conversational data set

Callers call with specific questions

- Two channels, one speaker each
- Member quality noisier
- 8kHz
- Best results: 10% hand-split files with frozen layers and removal of short files
- Current WER: ~40%
- Massive overfit

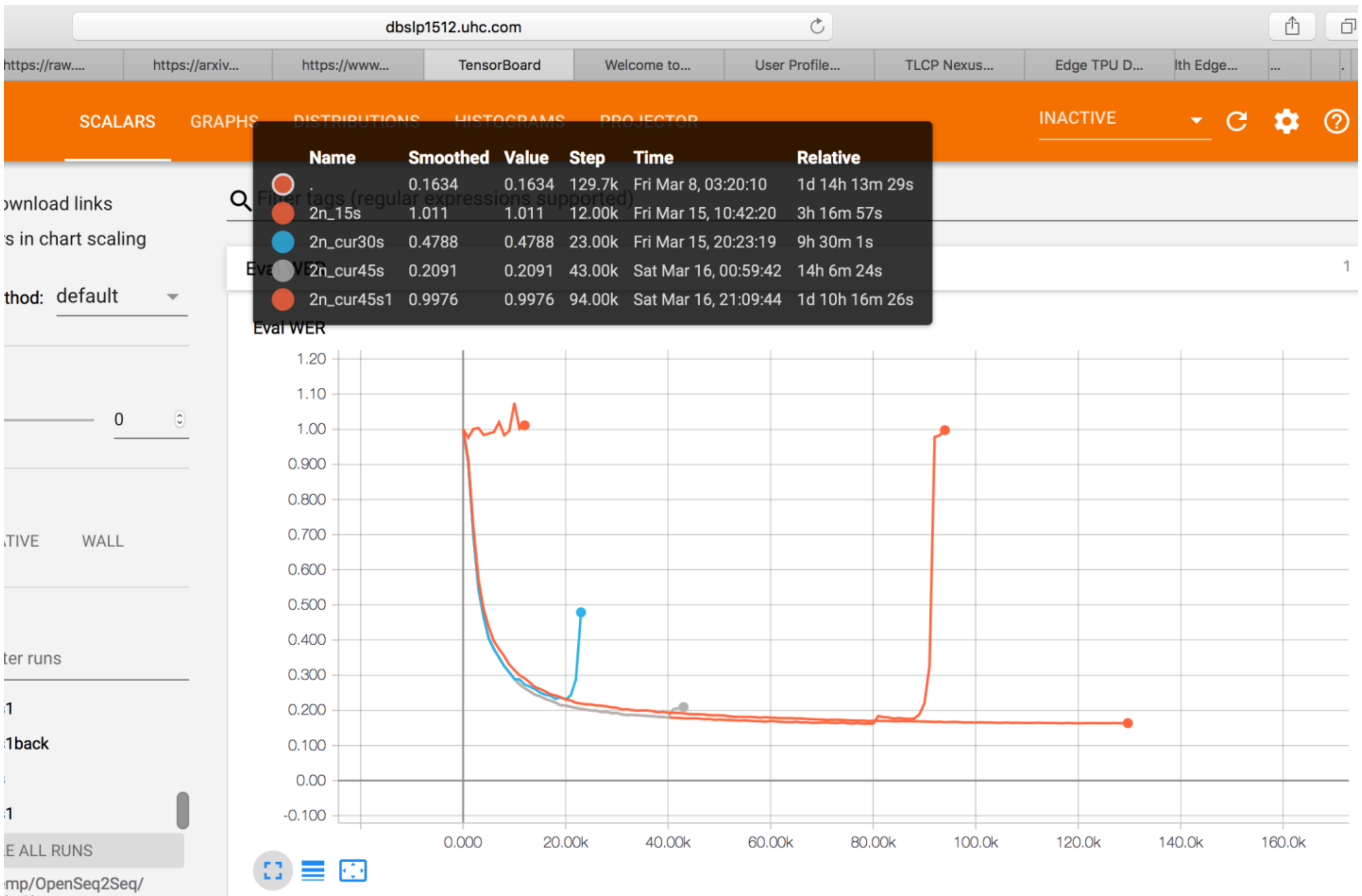
Samples	Avg,s	Total, hr
341	191.1	18.1



Experiments

- Curriculum learning – gradually introducing new data.
- Lower learning rate (0.001 ➤ 0.0001)
- Higher Decay
- Regularization not possible on Cudnn layers with given pretrained weights.
- Omitting very short audios (under 0. seconds) – slight improvement
- Freezing Conv and all RNNs
- Stretching incoming audios
- Silence removal
- Volume adjustment

Attempts to train on longer fragments



Cloud Service Comparison

sample1

Ground Truth

hi nancy it is amanda that sunshine hospital i got your message excuse me noise on caitlin laurel date of birth nine twenty two ninety six admitted to me on seven two seventeen no i appreciate the reminder nancy absolutely she was last covered day on this to review on six thirteen but we are a platinum protocol facility on any commercial um just want you to confirm that with you again we are platinum facility and will be will review on the thirteenth today which lead us to the twenty first please feel free to reach out at three four three six seven five nine eight zero two thank you

Jasper with greedy

- hi nancy as inmanda san shine hospital i got your message excuse me noise on kaitlyn laurel date of birth nine twenty two ninety six admitted to me on seven two seventeen now i appreciate the reminder nancy absolutely she was last covered day on thee on six thirteen but we are a platinum protocol facility on any commercial just wanted to confirm that with you again we are platnin facility and we will we will review on the thirteenth day which wead us to the twenty first please feel free to reach out at three four three six seven five nine eight zero two thank you

Cloud Service

Hi Nancy is Amanda at sunshine hospital? I got your message. Excuse me, no lies and it Caitlin Laurel date of birth 92295 admitted to me and 7 to 17 know. I appreciate the reminder Nancy. Absolutely. She was last covered day on 31613, but we are a platinum protocol facility on any commercial. Just wanted to confirm that with you again. We are platypuses ility, and we hope you will review on the 13th day which lead us to the 21st. Please feel free to reach out at 343-675-9802. Thank you.

Cloud Service – normalized, WER: 0.31

hi nancy is amanda at sunshine hospital i got your message excuse me no lies and it caitlin laurel date of birth nine two two nine five admitted to me and seven to one seven know i appreciate the reminder nancy absolutely she was last covered day on three one six one three but we are a platinum protocol facility on any commercial just wanted to confirm that with you again we are platypuses ility and we hope you will review on the one three th day which lead us to the two one st please feel free to reach out at three four three six seven five nine eight zero two thank you

Jasper with LM, WER: 0.135

hi nancy i as amanda sunshine hospital i got your message excuse me noise on caitlyn laurel date of birth nine twenty two ninety six admitted to me on seven two seventeen now i appreciate the reminder nancy absolutely she was last covered day on the on six thirteen but we are a platinum protocol facility on any commercial just wanted to confirm that with you again we are planning facility and we will we will review on the thirteenth day which lead us to the twenty first please feel free to reach out at three four three six seven five nine eight zero two thank you

Cloud Service Comparison

Sample 2

Ground Truth

hi kendra this is a potto calling you on diamond recovery center patient sherry woods date of birth is four thirty one ninety five i d is zero nine eight seven six five four three two but just forgot to leave you my phone number i ran over time at the end of the message so could you return my call or palm return my call with the authorization number from thirty for ongoing that would be great five four three six seven eight four three two one extension is three forty five thanks have a good day bye

Jasper with greedy

hi keno this is renia paul calling you on dimond recovery center patient is sherry woods date of birth is four thirty one ninety five i d zero nine eight seven six five four three two this is forgot to leaveing my phone number irando o time at the end of the message so could you return my call or call return my call with authorization number from thirty four ongoing that would be great five four three six seven eight four three two one it function is three forty five thank this is have a good daye

Cloud Service

Hi Kendall, this is Renee Apollo calling you on diamond Recovery Center patient i s Sherry Woods date of birth of store. 3195 ID 0 9 8 7 6 5 4 3 2. I just forgot to lea ve you my phone number. I ran out of time at the end of the message. So, could y ou return my call or come return my call with the authorization number from 30 sal on on going that would be great. 543-678-4321 extension 345. Thanks. Have a good day.

Jasper with LM, WER: 0.17

hi ken this is renea paul calling you on diamond recovery center patient is sherry woods date of birth is four thirty one ninety five i d zero nine eight seven six five four three two this is forgot to leaving my phone number aranda of time at the end of the message so could you return my call or call return my call with authorization number from thirty four ongoing that would be great five four three six seven eight four three two one it function is three forty five thank this is have a good day

Cloud Service Normalized, WER: 0.23

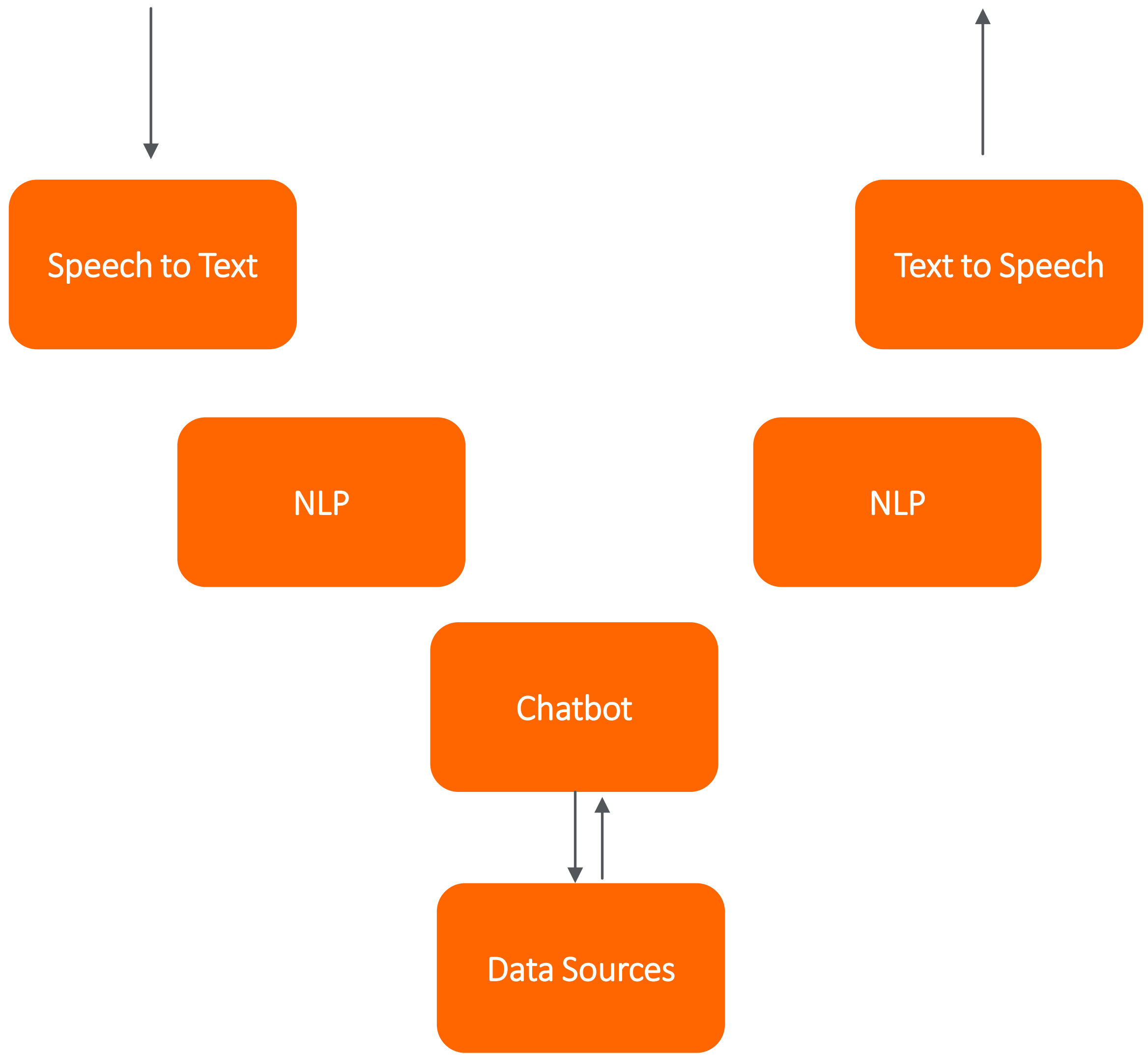
hi kendall this is renee apollo calling you on diamond recovery center patient is sherry woods date of birth of store three one nine five id zero nine eight seven six five four three two i just forgot to leave you my phone number i ran out of time at the end of the message so could you return my call or come return my call with the authorization number from three zero salon on going that would be great five four three six seven eight four three two one extension three four five thanks have a good day

OPTUM™

ASR + NLP Evaluation

- WER not good enough for most applications.
  - Most words in the speech are glue words
  - Informational words are only used once or twice
- Normalized NER:
  - Remove speech disfluencies and plurals
  - Normalize number sequences and contractions
- Extrinsic evaluation: A metric that only includes “Informational tokens”.
  - Named Entities: Important words/phrases
  - Relations and Topics
  - Average F-score on all named-entities, relations and topics.

NLU (Natural Language Understanding) : End to End



Conclusions

Very promising, but much to do

- Transfer learning appears to be the way to go
- However, it’s not as easy as with imaging
- It’s not clear that “Speech is Speech is Speech”, limiting the accumulation of data sets
- Stability (during training) needs work; networks often do not converge
- Cluster acceleration: much better, near perfect on our scales (24 GPUs)
- Data prep tooling is obviously lacking
- OpenSeq2Seq is not super user friendly but improving (containers, code readability)

OPTUM™

Next steps

- Ability to explore files & transcriptions
- Ability to edit the sound files and align with text
- Gradually, add intelligence (DL models) to this task
- Iterative improvement of the task
- Maximize the files / minute / human
- How long does it take to transcribe a 200 hour data set?
- Proper transfer learning
- Volume adjustment, silence / pause removal, leveraging noise..
- Neural Language models
- Speaker stratification
- The Spell Checker approach
- Use a trained model for splitting

OPTUM™

Next Steps – Leveraging Unlabeled Data

Autoencoders

- There's much much more untagged data – hundreds of millions of calls
- Obviously, we should find some use for it
- There's lots of noise, crosstalk, pauses
- Distortions are somewhat similar, an outcome of specific audio formats?

TTS / Speech Generation: Tacotron

Training&validation set – LJSpeech1.1

Training variations:

- OpenSeq2Seq versions – 18.09, 18.10
- Configurations – float, mixed
- Number of GPUs:
- 1 P5200 laptop
- 2 GTX1080TI desktop
- 1 and 4 V100
- Output "mel", "magnitude", "both "

Observations:

- On what GPU available, mixed precision is ~20% faster than float
- Training with 100000 steps on single GPU takes ~100hours
- Training with 40000 steps on 4 GPUs takes ~50hours
- Memory of GPU is important. On V100 and P5200 with 16GB memory batch size is 48, while on GTX1080TI with 11GB memory batch size is 32.
- Because of batch size training on 2 GTX1080TI and not very efficient multi gpu training, training on 2 GTX 1080TI is just little faster than on single P5200.

OPTUM™

Neural Machine Translation

Data Preparation

- Training Data
  - Pulling data from public databases
  - Adding domain specific data
    - Extracting text from documents
    - Aligning sentences
  - Clean and shuffle data
- Create common vocabulary and language model from training data
- Tokenize training and evaluation sets using common model and vocabulary

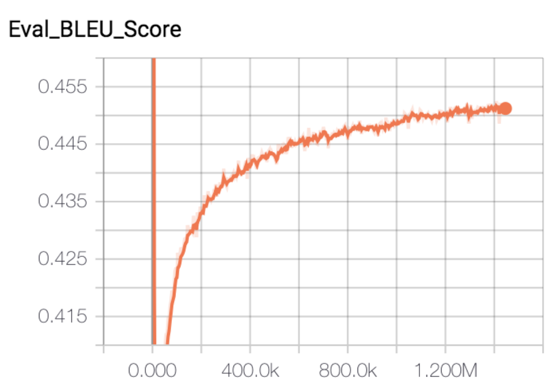
Training

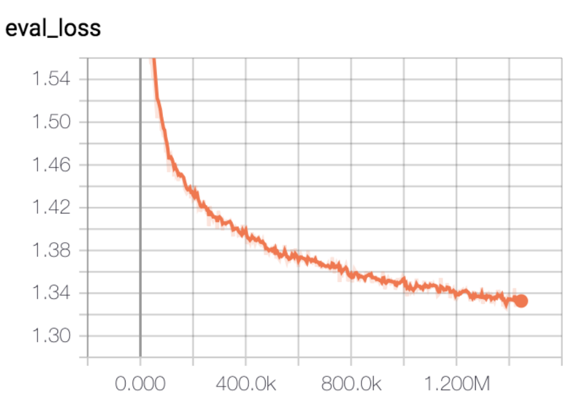
- 4 GPUS
- Transformer model

Neural Machine Translation – English to Spanish

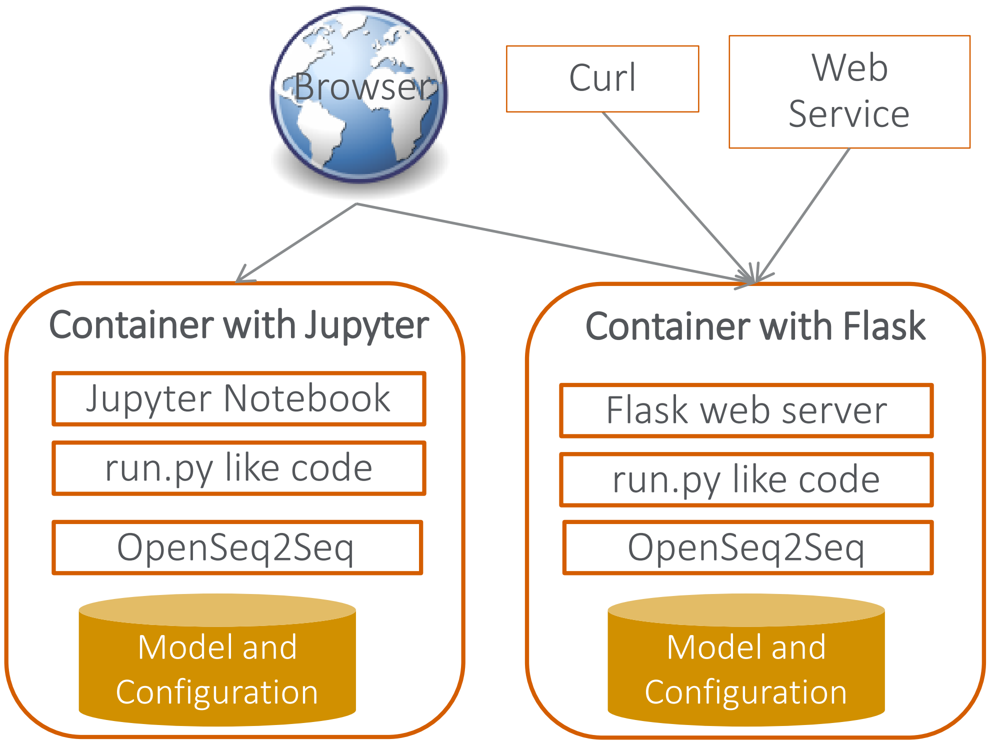
Use Case: Correspondence

- Data used for training:
    - Public data: 19550410 sentences
    - Translation memory and extracts from bilingual documents 1602561
    - Evaluation set – public data: 3000 sentences
    - Best model BLUE score: 45%
- Test on 118649 sentences extracted from documents: 59.75%





Inference with OpenSeq2Seq



text2speech inference is

- Very minimally using one GPU
- Takes 1.5-2 times less time then the length of produced file
- Cannot produce longer than 16-17 sec of recognizable speech
- Ends with last sec babble for any phrase producing file longer than 5-6 sec.

text2text inference

- **Requires GPU**
- GPU two order of magnitude faster than CPU

OPTUM™

HUGE THANKS TO:

Kiran Meetakoti
Galina Grunin
Jacob Glozman
Darragh Hanley
Kathrin Bujna
Tom Sullivan
Danita Kiser
The openseq2seq team

OPTUM™

Thank you

Questions?