Build Sovereign AI

Sasank Chilamkurthy, Adithya Bhat

Introduction

Who we are

JOHNAIC

Sasank Chilamkurthy

- Founder of **JOHNAIC**
- Former CTO of Qure.ai
- Co-author of PyTorch

Adithya Bhat

- Co-founder of **JOHNAIC**
- Former data analyst at Caterpillar

AI Research Experience

Been at it from 2016 before it was cool

Sasank Chilamkurthy

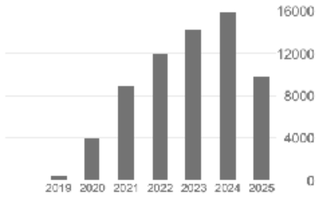
Qure.ai
Verified email at qure.ai - Homepage
Deep Learning Healthcare Machine Learning Radiology Open Source

FOLLOW

TITLE	CITED BY	YEAR
Pytorch: An imperative style, high-performance deep learning library A Paszke, S Gross, F Massa, A Lerer, J Bradbury, G Chanan, T Killeen, ... Advances in neural information processing systems 32	62866	2019
Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study S Chilamkurthy, R Ghosh, S Tanamala, M Biviji, NG Campeau, ... The Lancet 392 (10162), 2388-2396	991	2018
Pytorch: An imperative style, high-performance deep learning library B Steiner, Z DeVito, S Chintala, S Gross, A Paske, F Massa, A Lerer, ...	706	2019
Pytorch: An imperative style, high-performance deep learning library. arXiv 2019 A Paszke, S Gross, F Massa, A Lerer, J Bradbury, G Chanan, T Killeen, ... arXiv preprint arXiv:1912.01703 10	330	1912
2D-3D fully convolutional neural networks for cardiac MR segmentation J Patravali, S Jain, S Chilamkurthy International workshop on statistical atlases and computational models of ...	157	2017
Development and validation of deep learning algorithms for detection of critical findings in head CT scans S Chilamkurthy, R Ghosh, S Tanamala, M Biviji, NG Campeau, ... arXiv preprint arXiv:1803.05854	149	2018
PyTorch: An Imperative Style, High-Performance Deep Learning Library, Dec A Paszke, S Gross, F Massa, A Lerer, J Bradbury, G Chanan, T Killeen, ... arXiv preprint arXiv:1912.01703	83	2019

Cited by

	All	Since 2020
Citations	65625	64766
h-index	14	14
i10-index	16	15



Public access VIEW ALL

0 articles 1 article
not available available

Based on funding mandates

Co-authors EDIT

No co-authors

THE LANCET

This journal Journals Publish Clinical Global health Multimedia Events About

ARTICLES · Volume 392, Issue 10162, P2388-2396, December 01, 2018 Download Full Issue

Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study

Sasank Chilamkurthy, BTech [a] · Rohit Ghosh, BTech [a] · Swetha Tanamala, MTech [a] · Mustafa Biviji, DNB [b] · Norbert G Campeau, MD [c] · Vasantha Kumar Venugopal, MD [d] · et al. Show more

Affiliations & Notes Article Info Linked Articles (4)

Publication History: Published October 11, 2018
DOI: 10.1016/S0140-6736(18)31645-3 Also available on ScienceDirect
Copyright: © 2018 Elsevier Ltd. All rights reserved.
Check for updates

Get Access Cite Share Set Alert Get Rights Reprints

Show Outline

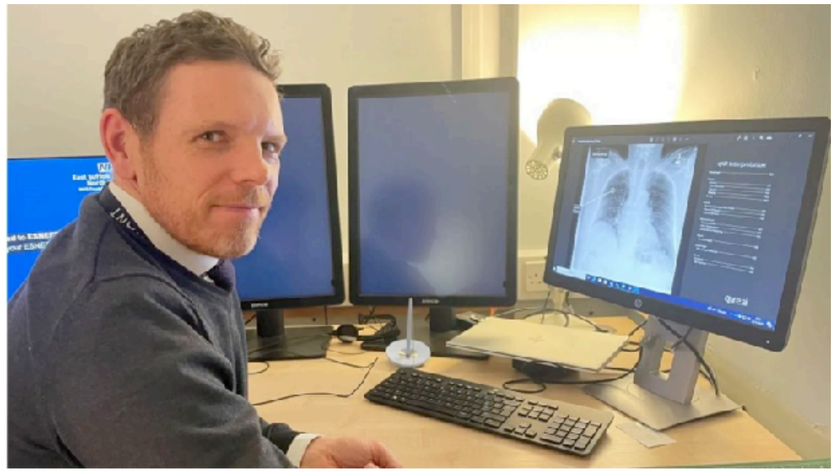
Summary

Background

Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms. We aimed to develop and validate a set of deep learning algorithms for automated detection of the following key findings from these scans: intracranial haemorrhage and its types (ie, intraparenchymal, intraventricular, subdural, extradural, and subarachnoid); calvarial fractures; midline shift; and mass effect.

Qure.ai: Top 3 healthcare AI company

Built here in India



- 50+ Countries
- 2500+ Centers
- 50,000+ Pts/day

We Design and Manufacture Computers

Computers are backbone of AI

Our Customers

Private and Cost Effective Cloud

Build vs Buy AI

Organizational Economics

Williamson: Transactional cost economics

be the efficient choice.

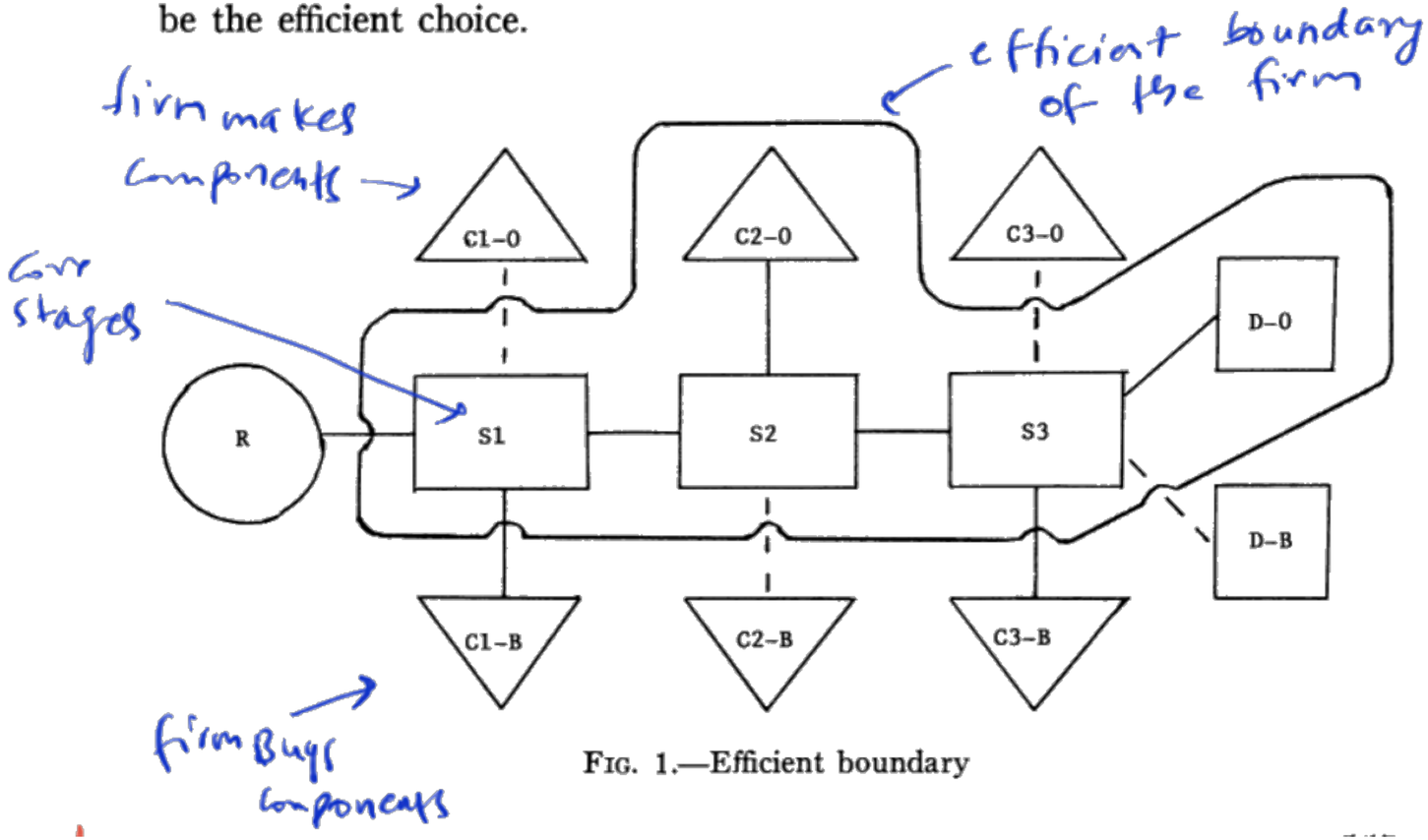


FIG. 1.—Efficient boundary

Good type	Example	Optimal Boundary
Non specific	Nuts and bolts, chairs, Slack, Jira	Buy
Specific	Car chassis for a car company, Employees, Company's app,	Build

Why do we build our own app?

It’s obvious but humour me

- Couldn’t use **shopify**/devshop because
  - Too specific requirements
  - Control over experience
  - Too critical to make mistakes
  - “Competitive Advantage”
- Why is AI different?

Marketing Lies about AI

What BigTech wants you to believe

- You need a huge datacenter and heck ton of GPUs
- You need million dollar talent
- Artificial “General” Intelligence
- Thus cheaper to buy

- Reality for Enterprises
  - ~7500 employees
  - $50/user/mo ChatGPT Enterprise
  - $4.5 million dollars / year bill
- Is it really cheap?

Is AI so hard?

Nope

- GPUs over cloud is expensive, true. 5-10x cheaper when you **own them**
- Great talent in India
- AI knowledge is not really a secret. Everything is published
- Opensource models offer a great crutch to start with

- Point of this talk is to prove we can do it

We trained Indic Voice AI

What we did

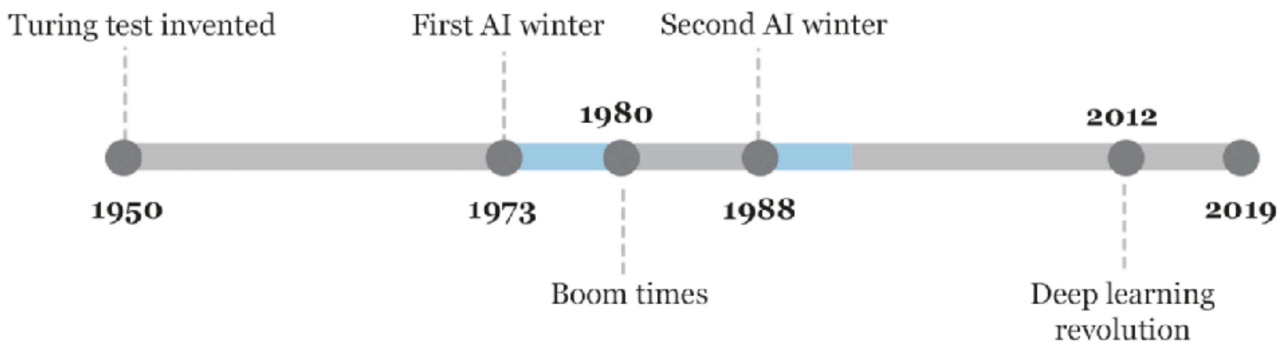
- We trained Kannada voice models
  - TTS (Text to speech): Make the computer speak
  - STT (speech to text): Make the computer listen
- We’ll talk more about TTS in this talk
- Used AI4Bharat models as the base model (aka ‘foundation’)
- Completely synthetic data based training
- Trained on JOHNAIC 16 - costs 3 lakhs

Why we are doing this

- Challenging Problem
  - Have to work in low data regime
  - Can do on a single GPU
- Explore the boundaries of synthetic data
- Larger models have also hit data bottle necks (GPT-4 vs GPT-5)
- More GPU sales with Voice AI :)

Overview of Deep Learning

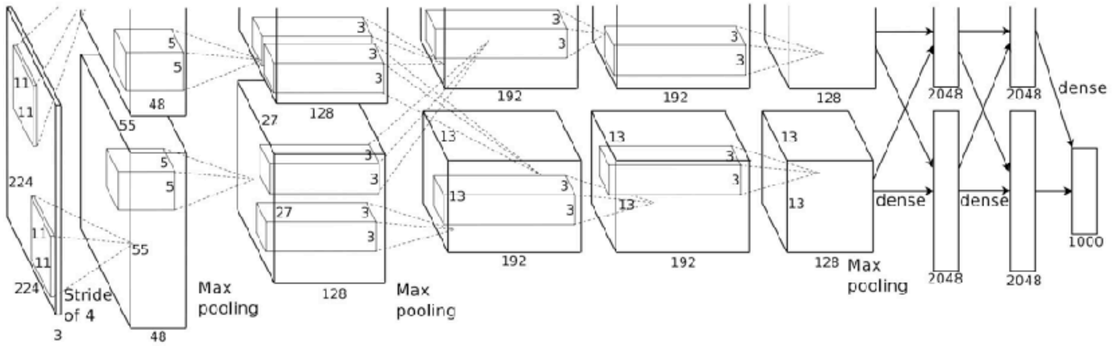
AI = Computers

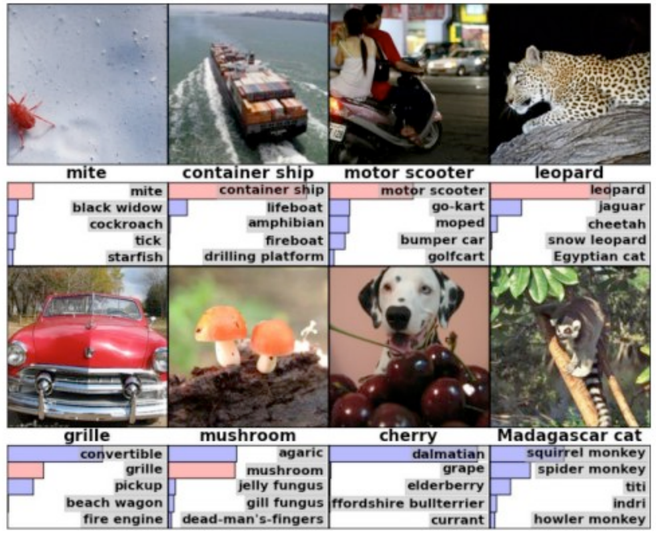


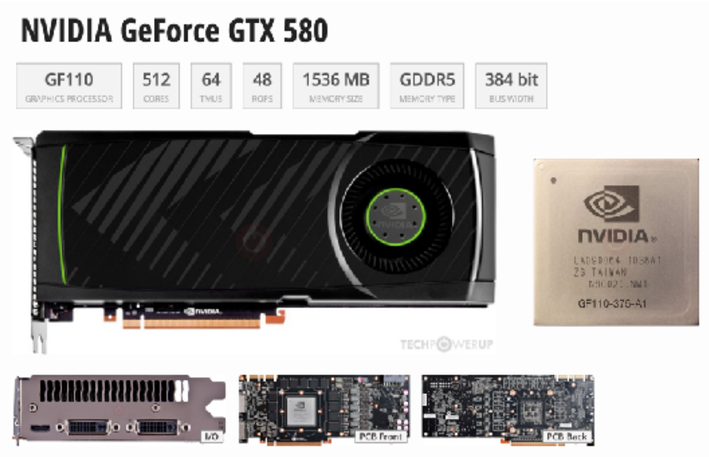
- Latest deep learning revolution of AI was caused by the availability of
  - Huge amount of data
  - Huge amount of compute to deal with that data
  - Highly open research and open source code

AlexNet: Convolutional Networks

Developed in 2012







Deep Q Learning: Reinforcement Learning

Developed in 2013

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih **Koray Kavukcuoglu** **David Silver** **Alex Graves** **Ioannis Antonoglou**

Daan Wierstra **Martin Riedmiller**

DeepMind Technologies

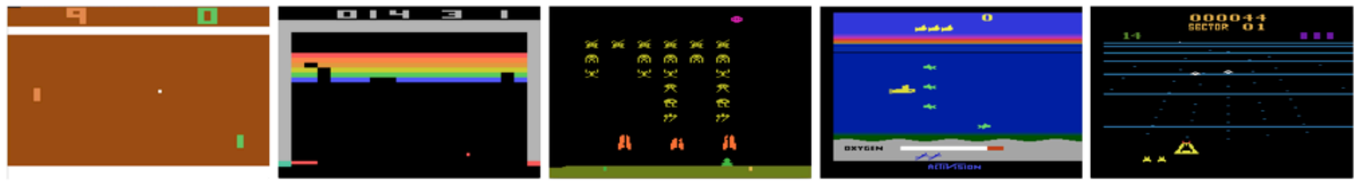


Figure 1: Screen shots from five Atari 2600 Games: (*Left-to-right*) Pong, Breakout, Space Invaders, Seaquest, Beam Rider

Seq2Seq: Recurrent Networks

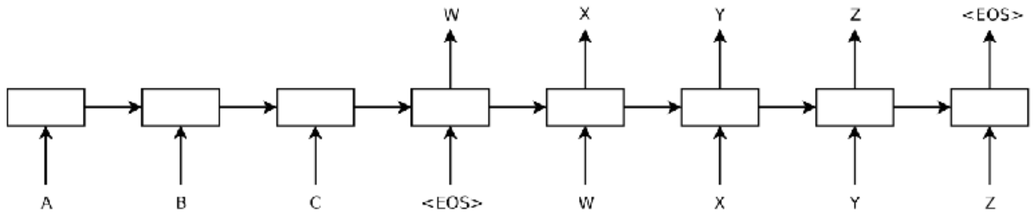
Developed in 2014

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com



Transformers: Upgrade over RNN

Developed in 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

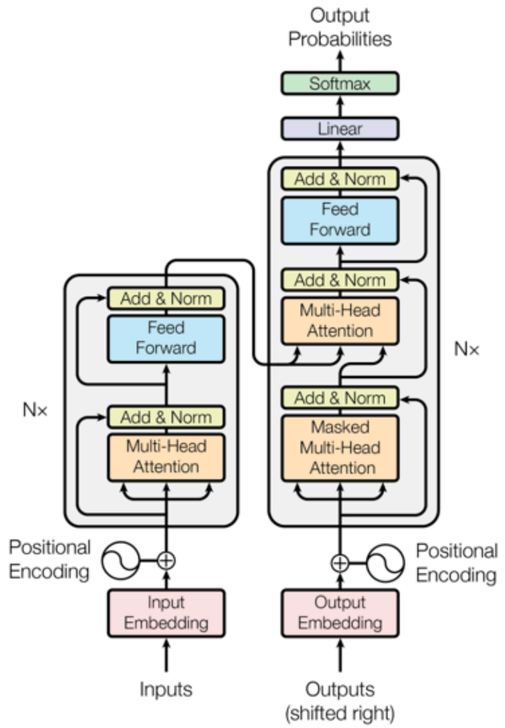
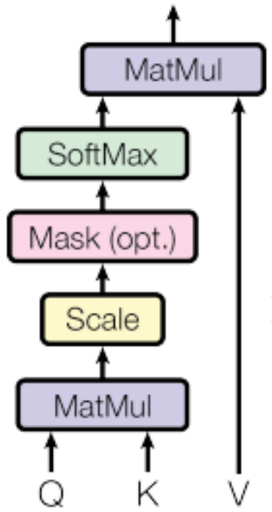


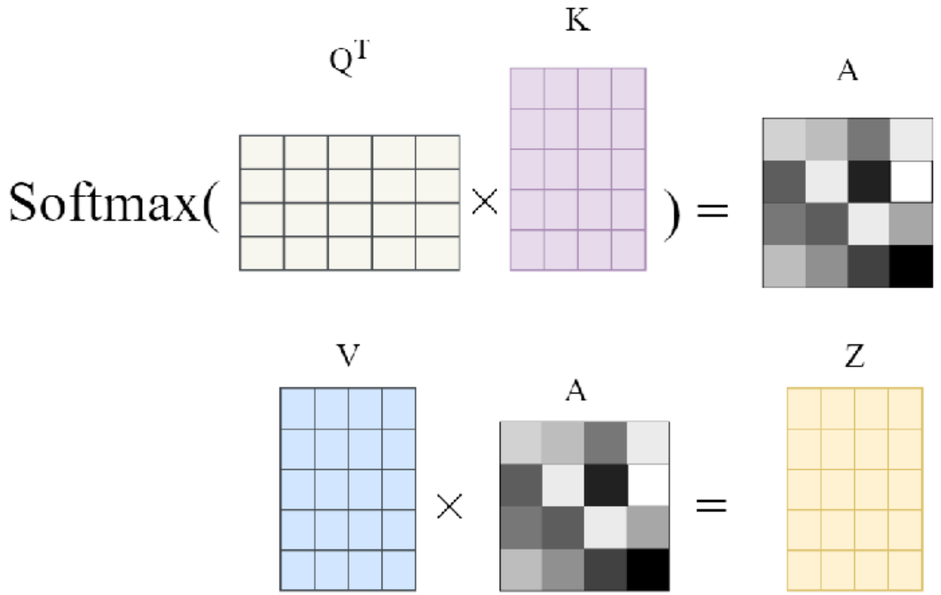
Figure 1: The Transformer - model architecture.

Attention is like database search

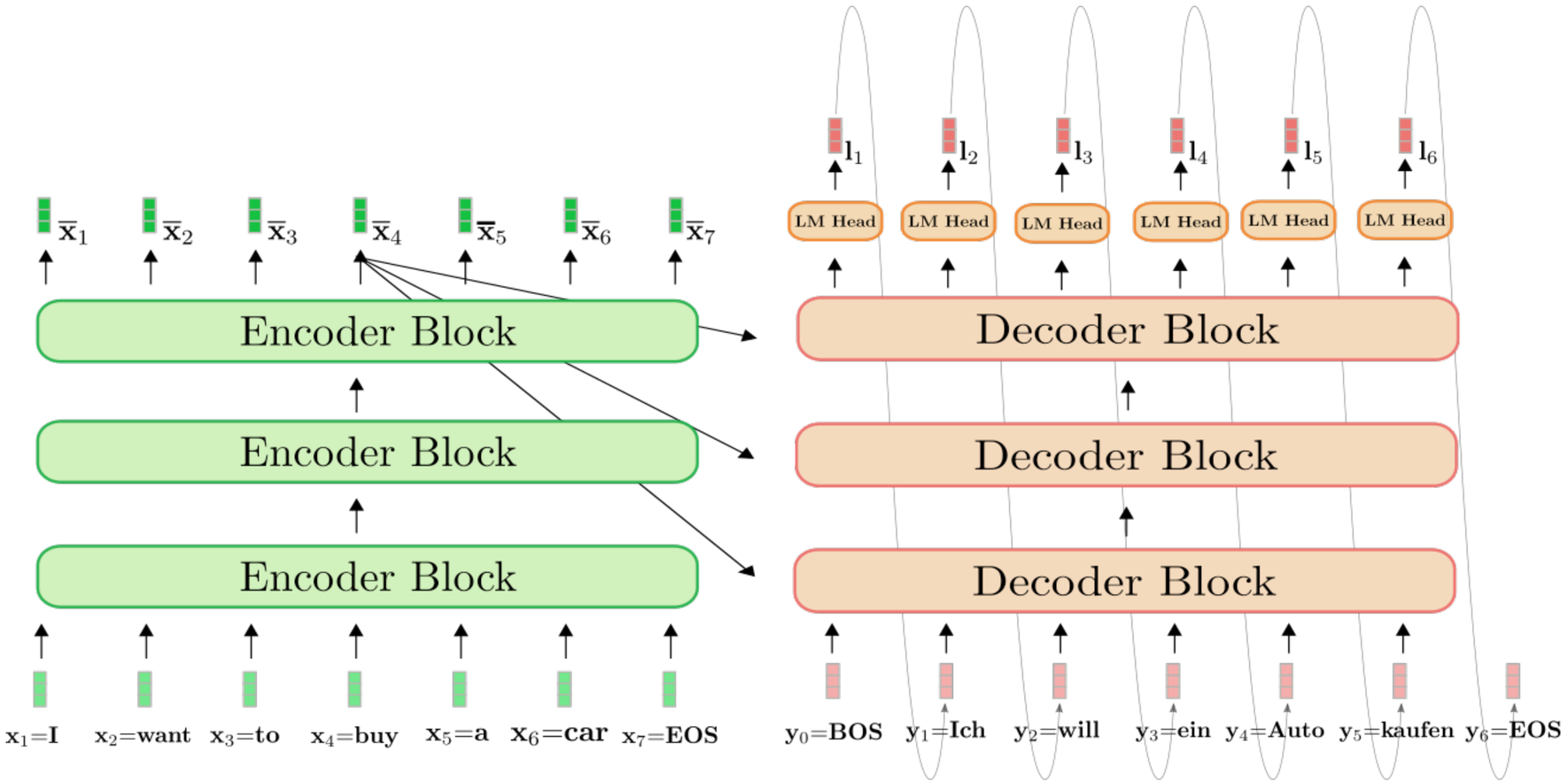


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Encoder - Decoder Architecture



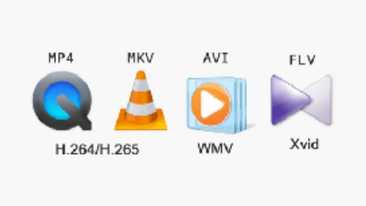
Three Types of Transformers

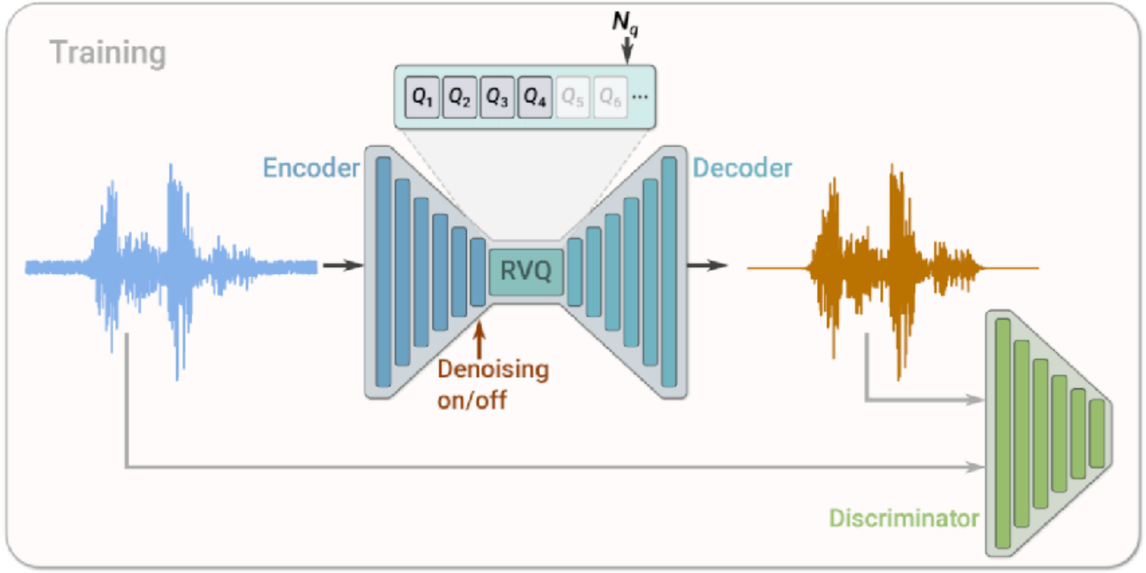
Type	Applications	Remarks
Encoder-only	Embeddings, Classification	You work with this when you don't have to generate new sequence
Decoder-only	Prompting	Generate sequences autoregressively like they are in training set
Encoder-Decoder	Translation, TTS/STT	Condition output sequence to generate based on an input sequence

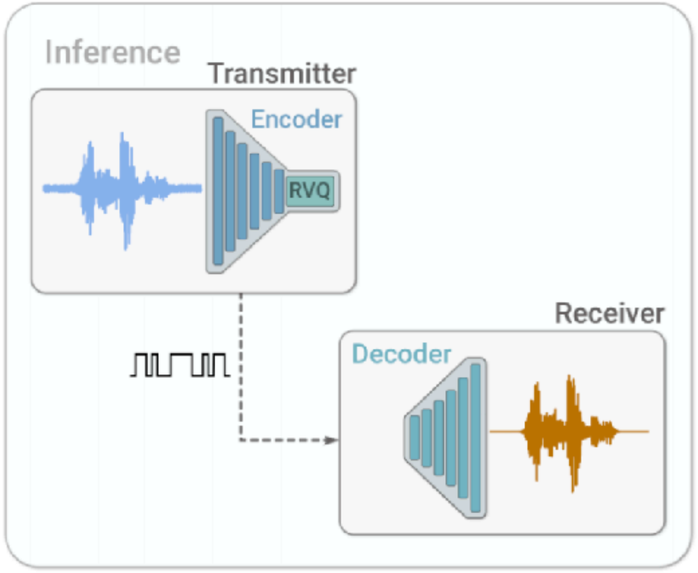
Voice AI

Codecs

Converts audio to vectors



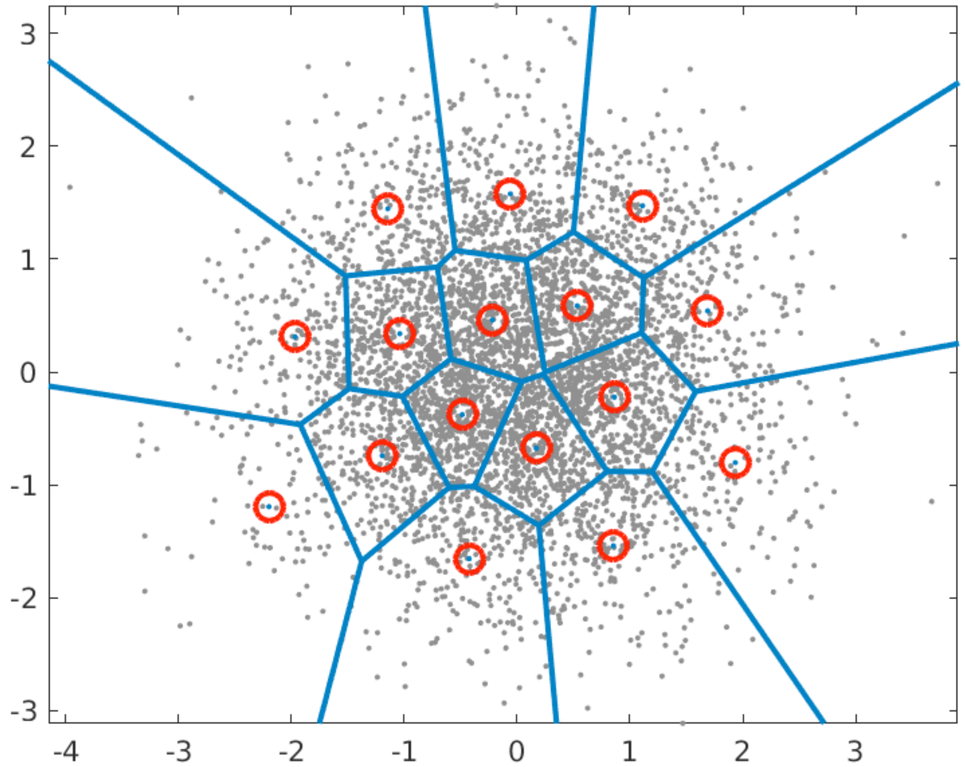




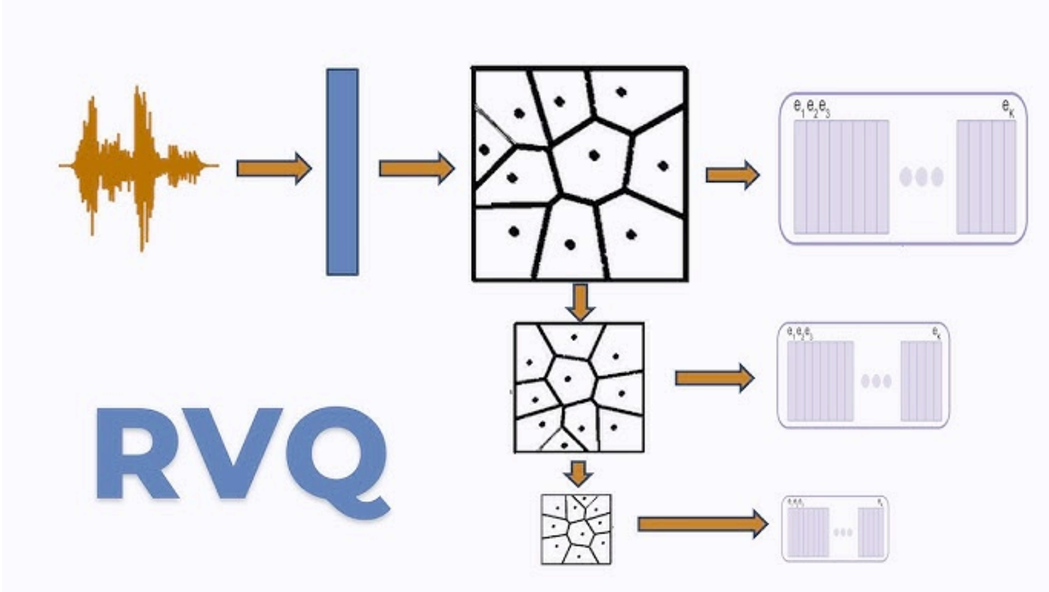
Residual Vector Quantization

Converts vectors to tokens

Vector Quantisation

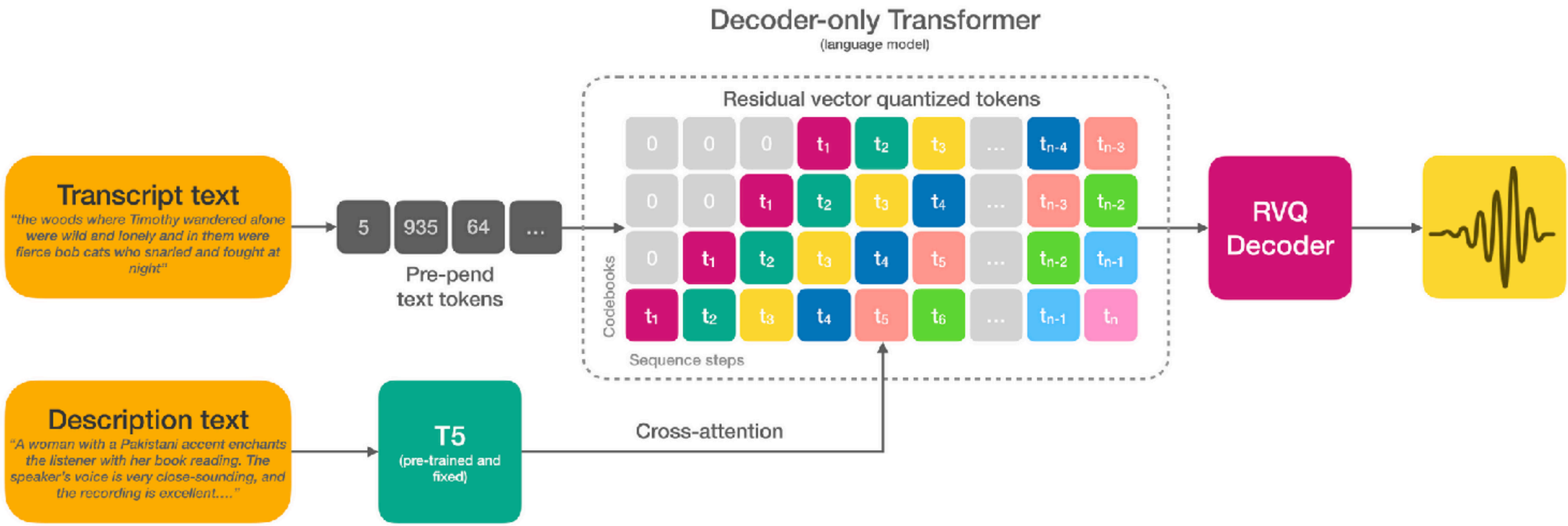


Residual Vector Quantisation



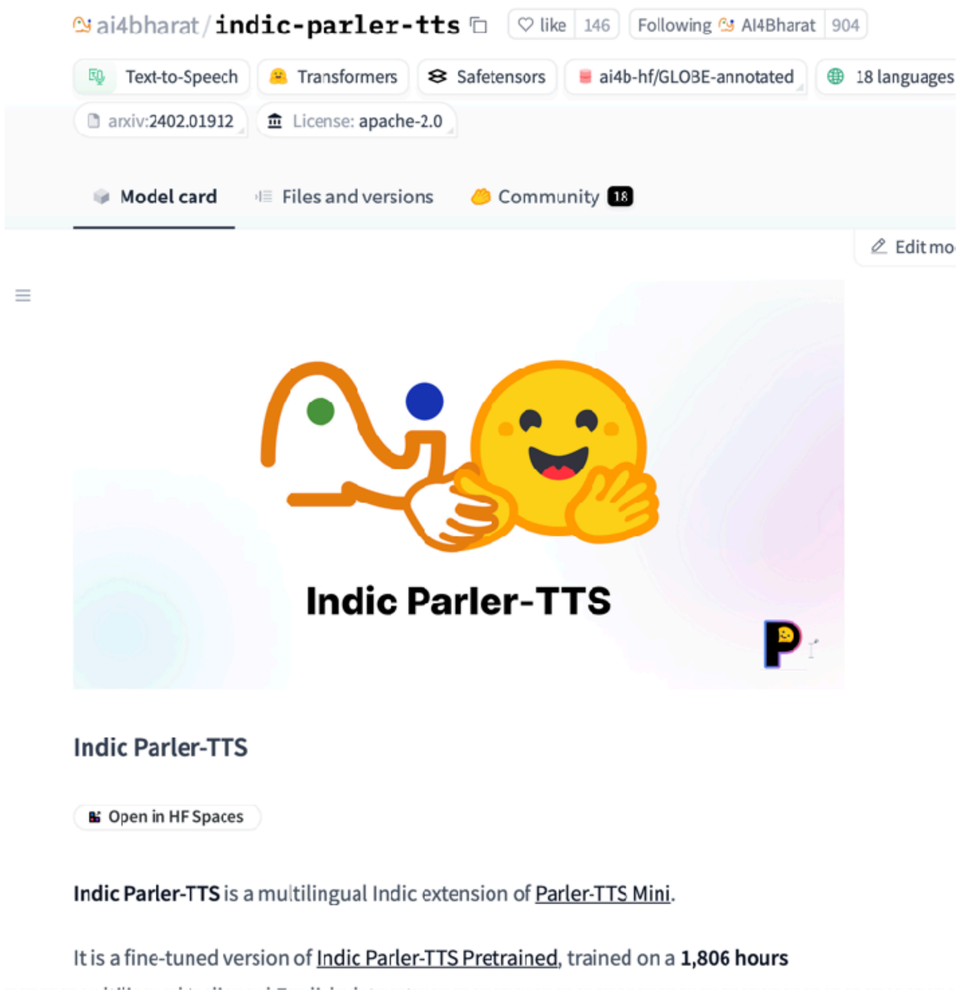
Parler TTS Architecture

It’s just a transformer with audio tokens



https://www.text-description-to-speech.com/

Base aka ‘Foundation’ Model



Problems we wanted to solve:

1. Too many languages, average quality in all
2. Too many speakers for 600m params

Our aims:

1. Solve one language really well
2. Replicate this for other languages

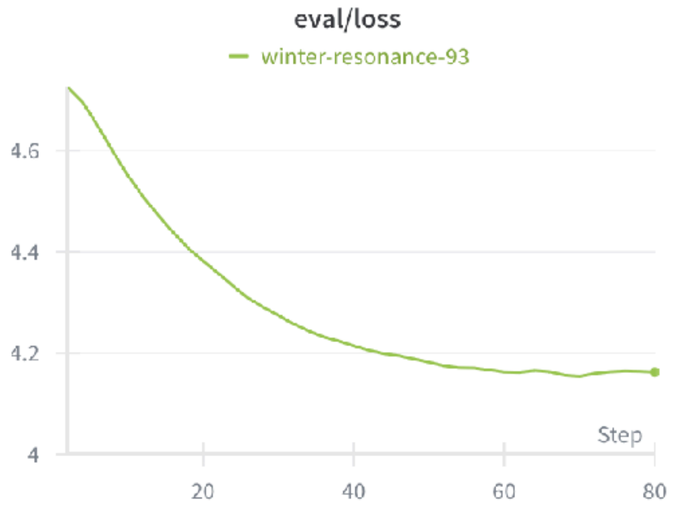
Data Generation

- We used synthetic data for training
- Scraped all of Kannada Wikipedia - you could download a data dump
- Split it into sentences and removed noise
- Converted numbers to textual numbers (100 -> hundred)
- Use Google cloud TTS to generate training data for ~ 25 hrs

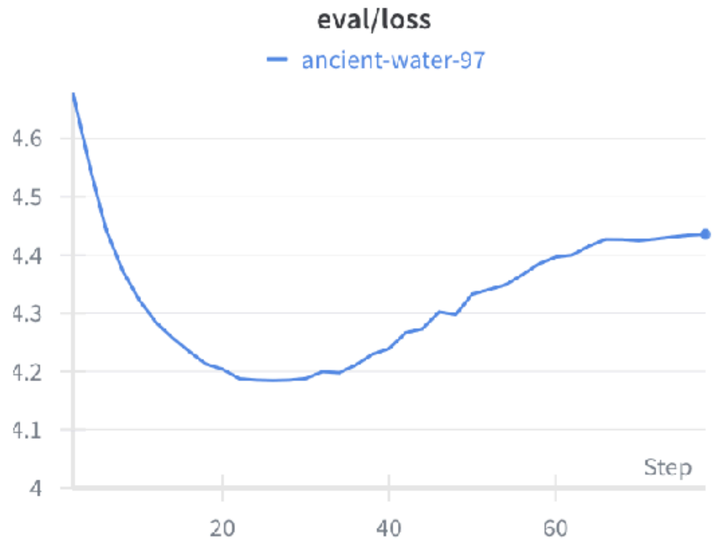
Ablation Experiments

Data size kept to <1 hr for finding right params

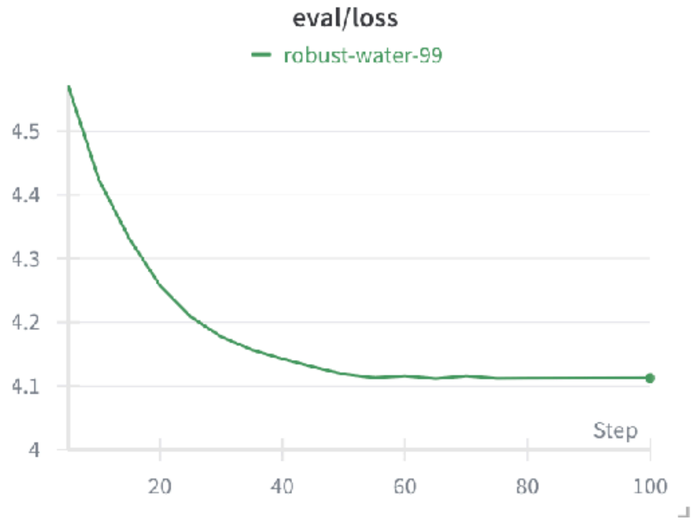
Lr = 1e-4, slow convergence



Lr = 1e-3, overfitting

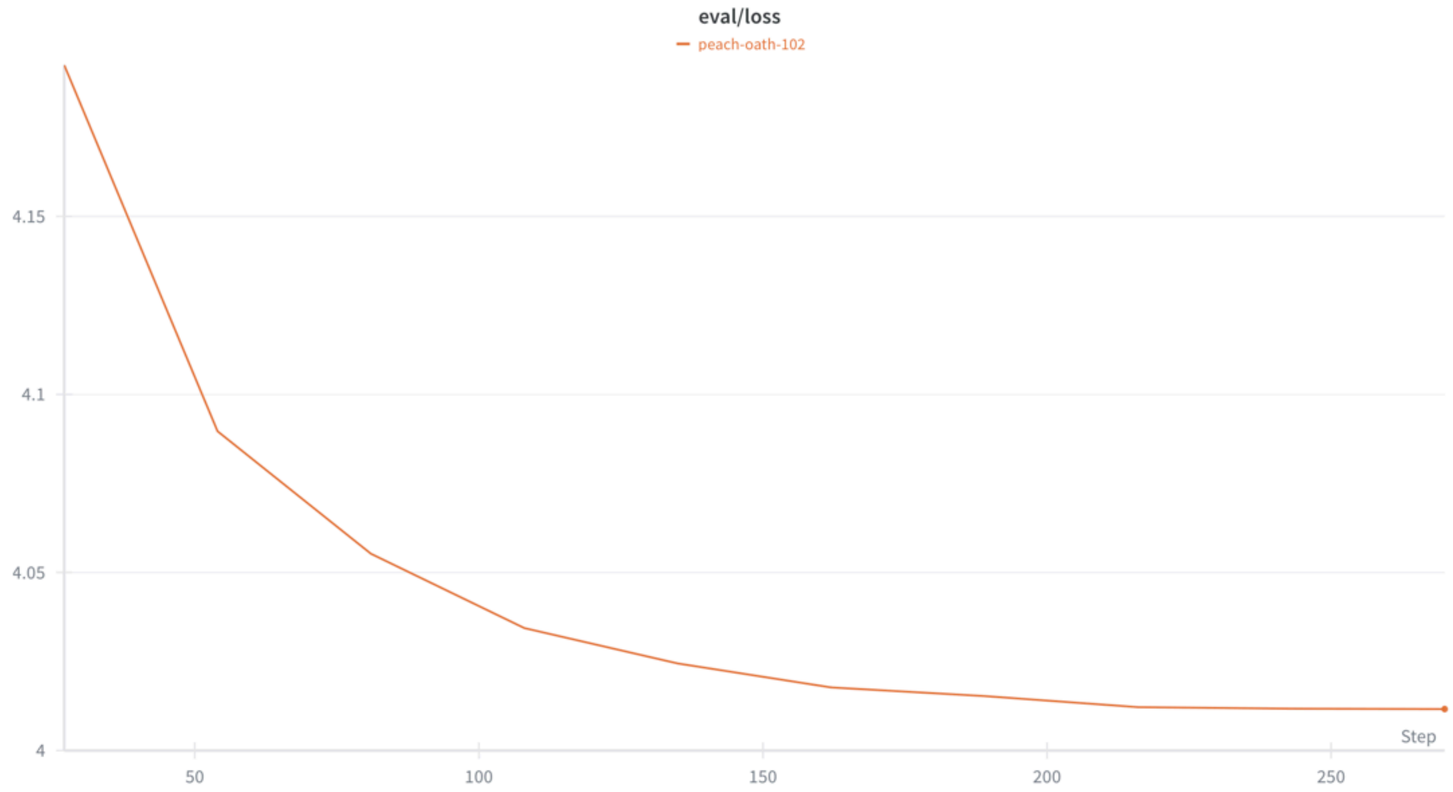


Lr = 2e-4, best results



Full Training

25 hrs of data



Future Work

- Hire a voice actor and create ‘emotional’ voice data/model
- Train STT/ASR model
- Use the TTS model to generate data for STT model
- Evaluate ASR model trained on synthetic data on unseen dataset
- Train STS model?

Conclusion

- We can do it!
- Synthetic data is surprisingly good
- Can build on top of open source
- Compute resources are not too expensive if done onprem
- It was surprisingly easy