Automatic Speech recognition on EC2-Habana Gaudi processors

habanalabs-base-ubuntu20.04-1.3.0-499-1644739541-9a75c51a-a4d1-4470-884f-6be27933fcc8

Mohamed A. Bencherif,
Center of Smart Robotics Research-CS2R
College of Computer & Information Sciences -CCIS
King Saud University-KSU
Saudi Arabia

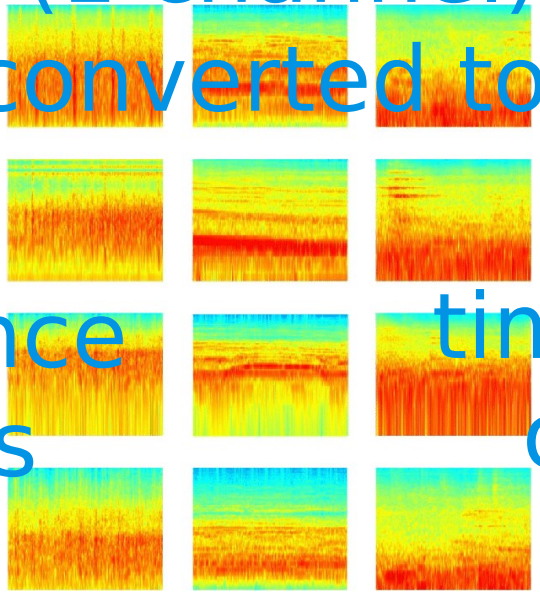
THANKS TO :
AWS, DEVOST, HABANA TEAM
ALL WHO HELPED....

Problem Definition

How to train an ASR in the Habana Gaudi processors on the AWS Cloud ?

Raw speech (1 channel) is in general converted to :

2D time Sequence of Matrices



3D time Sequence of Matrices

Candidate Networks

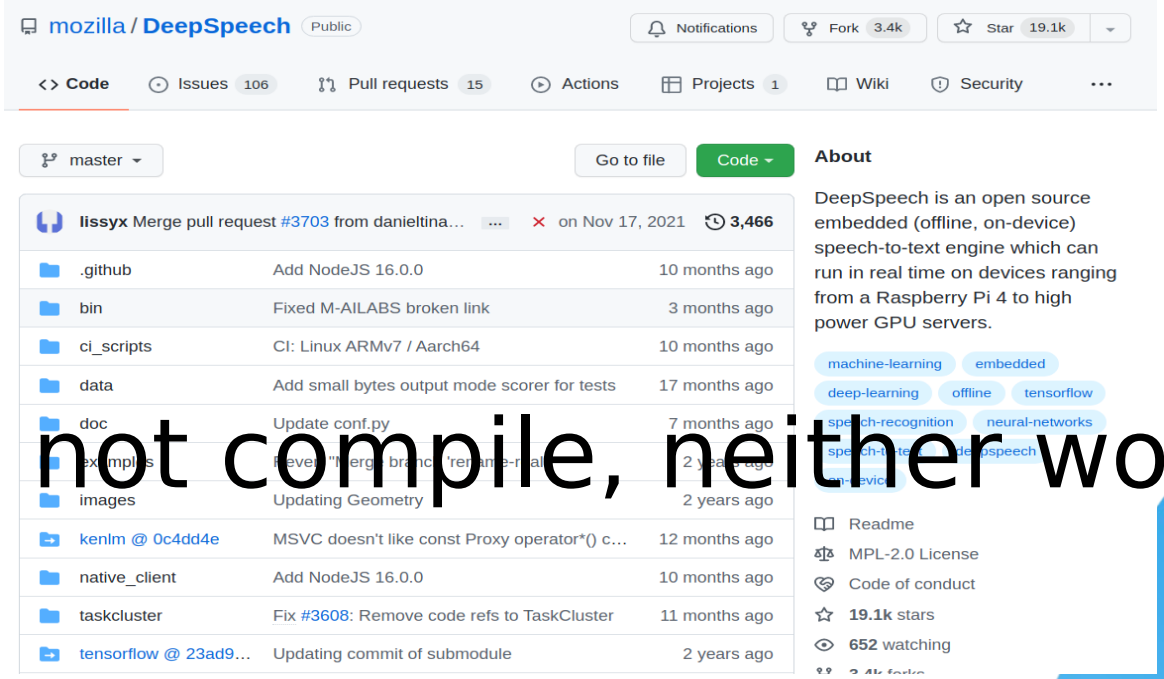
	Name	Type	Params	In sizes	Out sizes
0	cnn	Conv2d	320	[8, 1, 128, 1151]	[8, 32, 64, 576]
1	rescnn_layers	Sequential	56.3 K	[8, 32, 64, 576]	[8, 32, 64, 576]
2	fully_connected	Linear	1.0 M	[8, 576, 2048]	[8, 576, 512]
3	birnn_layers	Sequential	22.1 M	[8, 576, 512]	[8, 576, 1024]
4	classifier	Sequential	539 K	[8, 576, 1024]	[8, 576, 29]
5	criterion	CTCLoss	0	?	?

https://github.com/SeanNaren/deepspeech.pytorch

https://github.com/jiwidi/DeepSpeech-pytorch

https://github.com/mozilla/DeepSpeech
in tensorflow

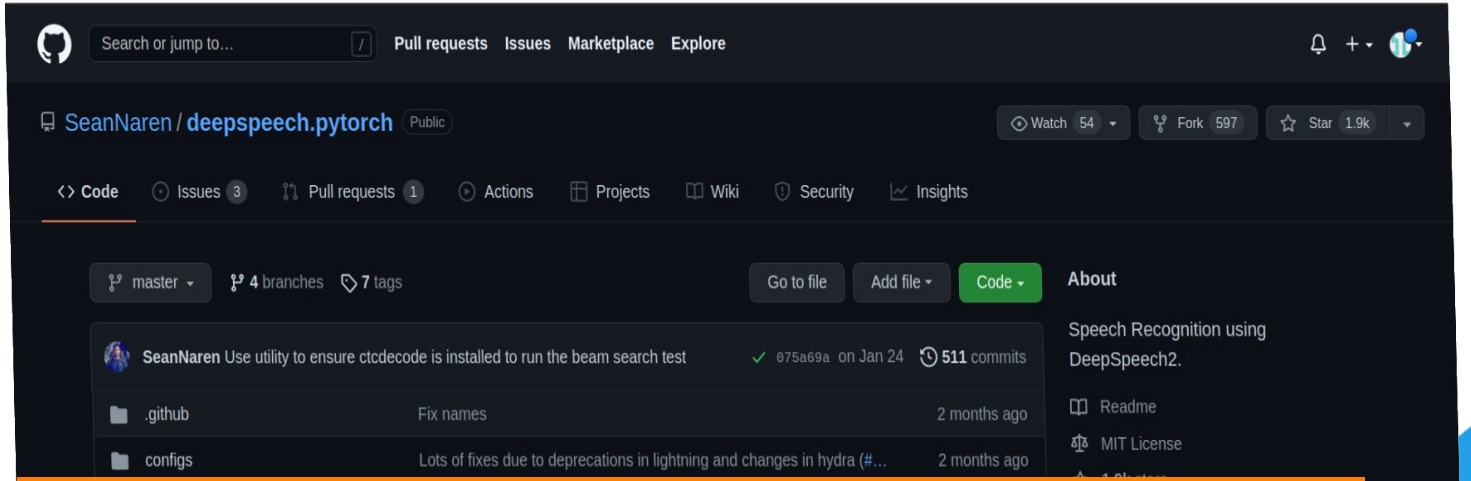
First Aws Credit 87$ spent on compiling and .....



Could not compile, neither work with

https://github.com/mozilla/DeepSpeech in tensorflow

Second Aws Credit ...

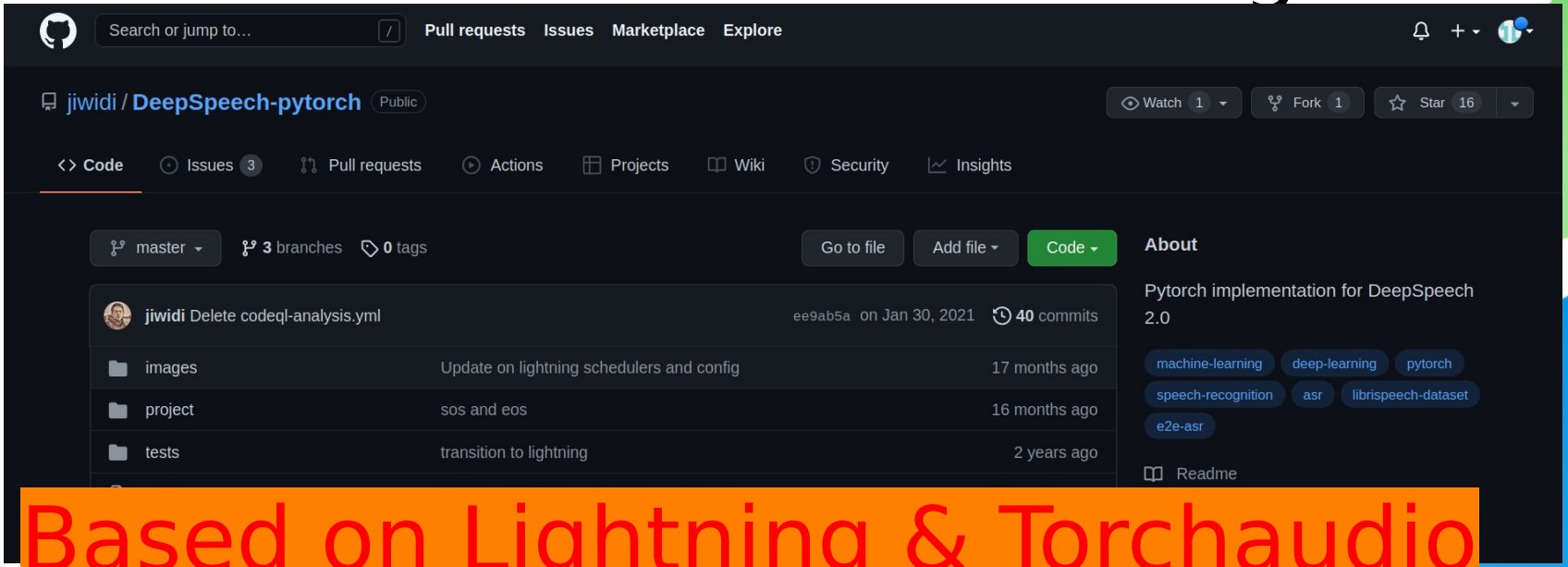


Based on Lightning & Hydra

Did not find a way to port hydra style to Habana

https://github.com/SeanNaren/deepspeech.pytorch

Second Aws Credit ... Still being Used



Based on Lightning & Torchaudio

https://github.com/jiwidi/DeepSpeech-pytorch

Worked fine on my local machine, ported to habana with torchaudio tweeking,
but some errors in negative dimension of the Tensors??
+ I did a toy example for the deepspeech in the same MNIST Style

Code Constraints

- **Do not break the Code of the habana framework**
- **Code in tensorflow or pytorch need to be adjusted to work on the habana machines.**
- **Tensorflow codes are restricted to tf>2.6**
- **Torchaudio does not work on Habana machine as it breaks the torch version of the machine**
- **Librosa can not be installed as it uses numba and this later requires numpy <1.21, but numpy used in habana is >1.22**
- **Working with PytorchLigtning**
  - **Which trainer ?? Model-References/Pytorch/central**
  - **Trainer on local machine and Ec2-HB**

Platform Constraints

- **Compiling from source requires complex procedure as CUDA is not supported and many available codes are CUDA dependent.**
- **Due to the hourly cost of the Ec2 instance, one needs to work on a local and a distant machine.**
- **But the Model-References Code needs to be run on the EC2.**

What went wrong?

- **Could not find the documentation on code adaptation easily.**

- **Workshops not working for every one as in the in the demo file,**
- **I could not benefit from it.**

- **Time Delay between question and answers, due to different time zones.**

Unet Example : my short nightmare...:)

- **One of the recommended code was the Unet Example.**
- **So much switches not find the documentation on code adaptation easily.**
- **Too much files to check and understand...**
- **Code should be improved and librairies should be better organized**
- **Workshop not working for every one in the demo file.**
-

Improvements

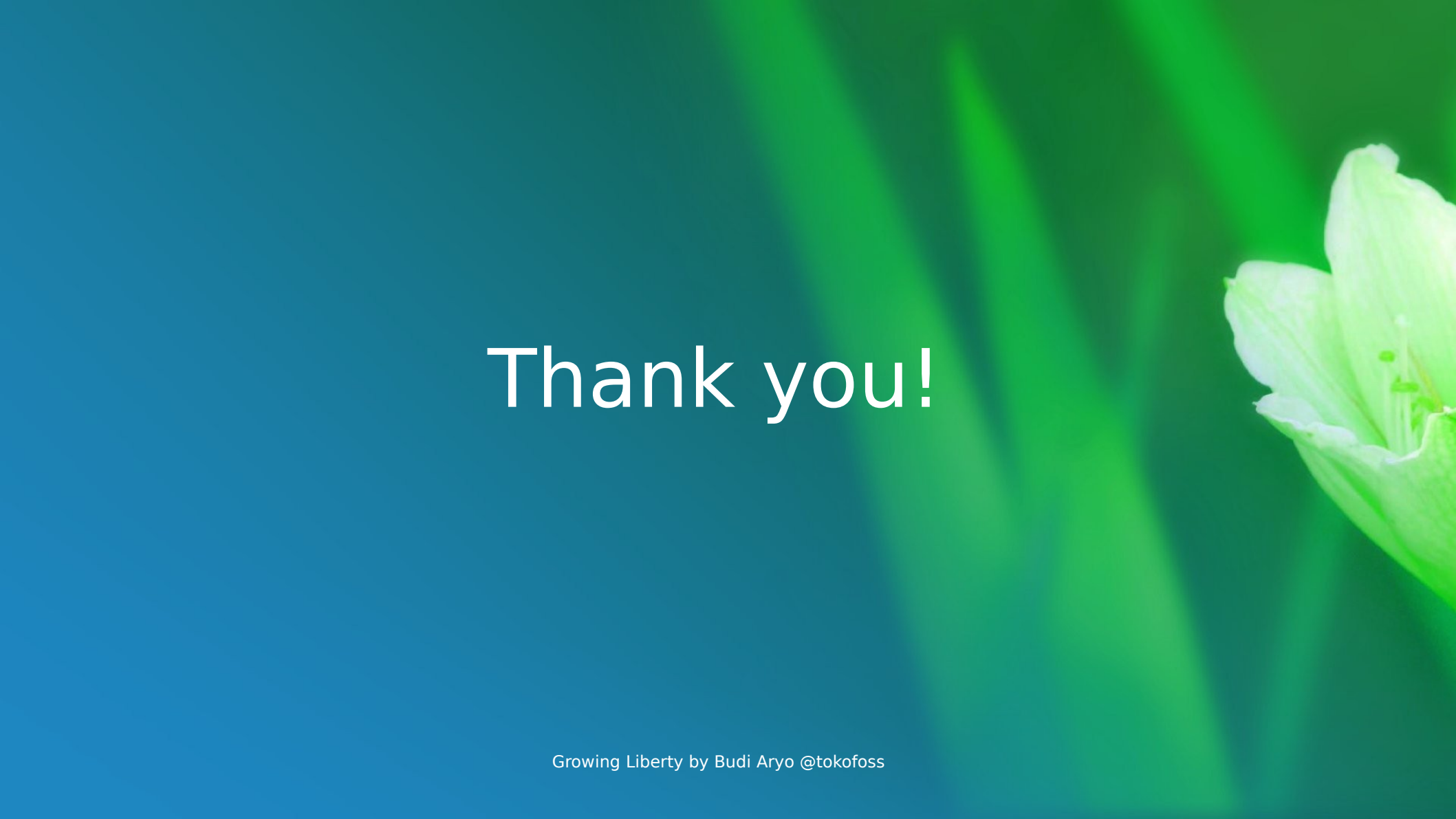
- - Used tweaked torchaudio on Gaudi Processors
- - Read flac files : replaced by soundfile
- Transformations
  - – Melspectrum:
  - - Time Masking
  - Frequency Masking
  -
  -

Gentle recommendations

- **Afford two types of machines, a low cost or free tiers machine containing 1 Gaudi processor for learning purposes and testing issues.**

- **A second machine for code deployment.**

- **An elastic IP for the Gaudi machine would have solved a lot copy-paste code, mainly when stopping and restarting the machine, each time a new allocated ip and new ec2-name, as well as file transfer via filezilla or aws cli commands, and ease the work with**


Thank you!