VEGI (ocr) - Voice Enhanced Gaming Interface

Project Presentation

Need for Local AI

- Reduced Latency:
  - Instantaneous processing and response times.
  - Essential for real-time applications like gaming and video editing.
- Enhanced Privacy:
  - Data is processed locally, reducing the risk of data breaches.
  - Critical for sensitive information handling in healthcare and finance.
- Cost Efficiency:
  - Lower operational costs by minimizing cloud service usage.
  - Reduces the need for constant internet connectivity.
- Improved Reliability:
  - Functions independently of internet connection.
  - Vital for remote areas or in situations with unstable connectivity.
- Energy Efficiency:
  - Optimizes power consumption by leveraging local processing.
  - Extends battery life in mobile devices.

Applications

Intelligent gaming

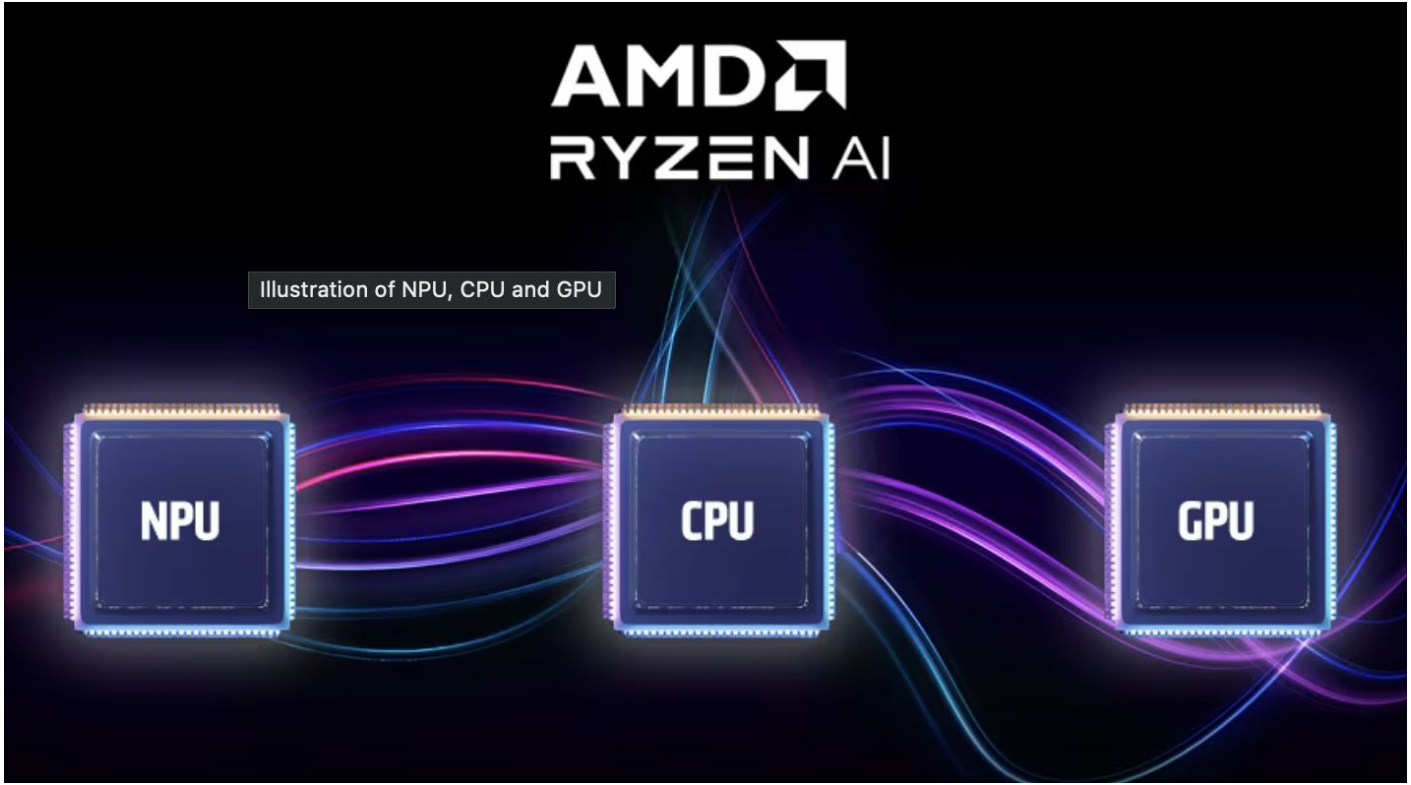
Higher quality audio and video

Virtual Assistants

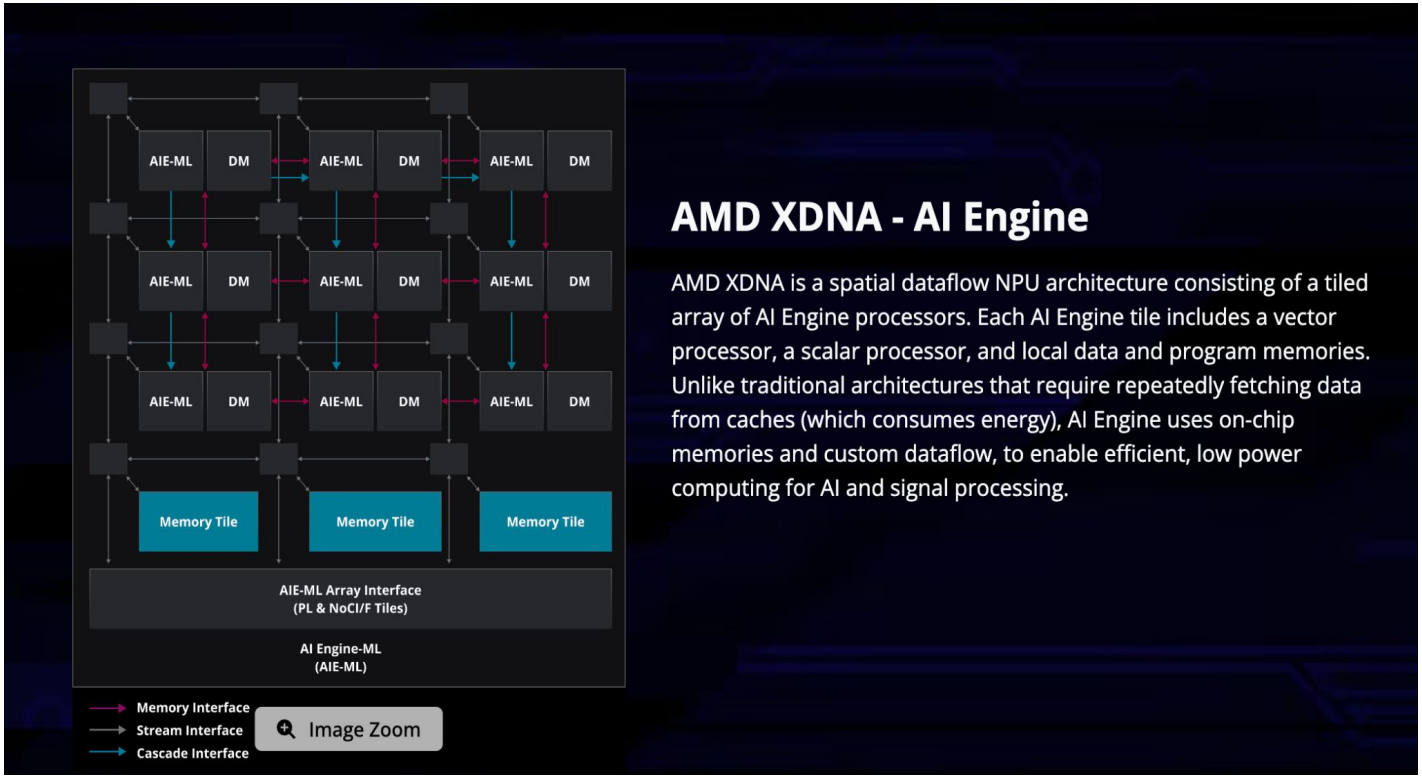
Enhanced video conferencing features

Smarter Security

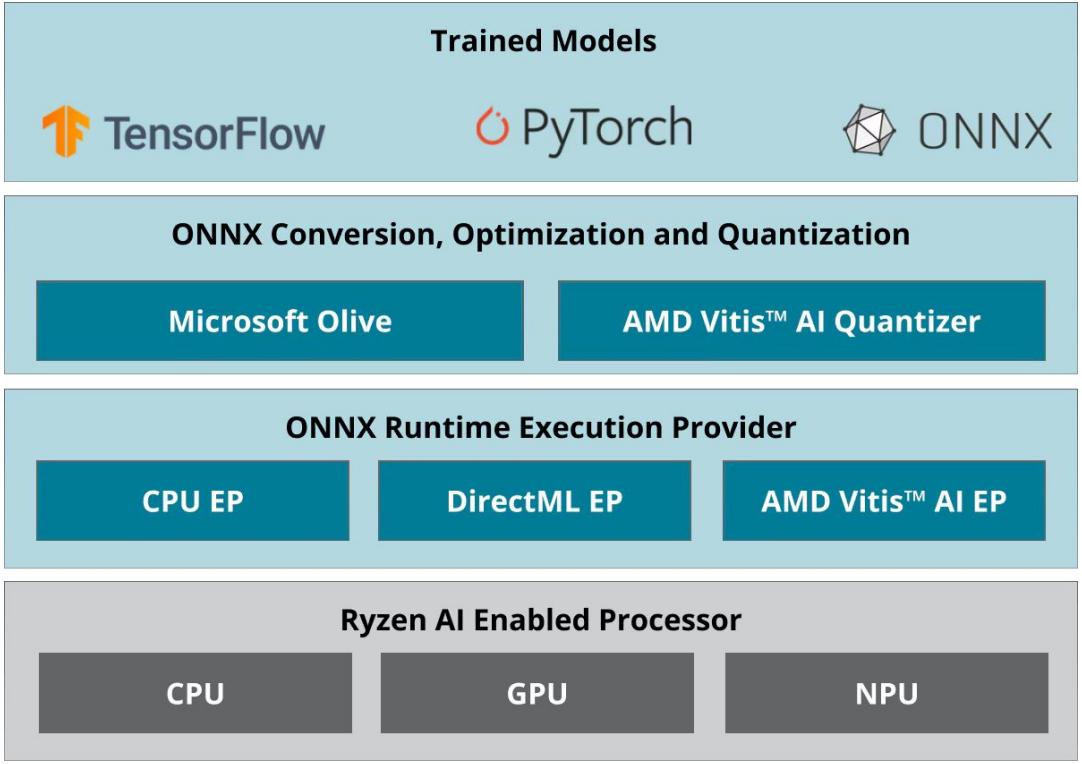
Ryzen AI



Ryzen AI architecture



Ryzen AI Software



ONNX


PyTorch
scikit learn
Chainer
K
Caffe2
mxnet
Cognitive Toolkit
ML
dmlc XGBoost
PaddlePaddle
ONNX
CPU
GPU
FPGA
ASIC

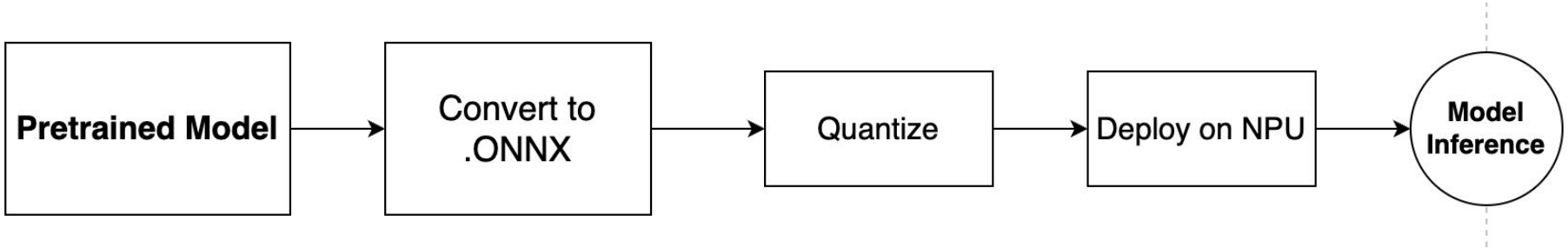
ONNX vs Compiler

Features	ONNX	Compiler Optimizations
Purpose	Standardizes ML models for interoperability and optimization	Translates and optimizes high-level code into machine code
Interoperability	Translates models across different ML frameworks	Translates code for different hardware platforms
Optimization Techniques	Graph optimizations like node fusion, constant folding	Code optimizations like loop unrolling, inlining, constant folding
Abstraction	Abstracts away framework-specific details	Abstracts away hardware-specific details
Execution	Produces an intermediate representation executed by a runtime	Produces executable machine code run directly on hardware

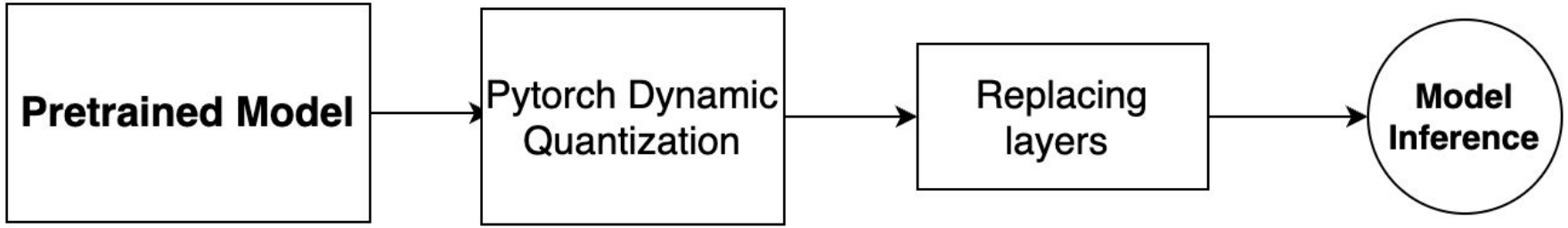
Challenges with ONNX

1. Complexity in Conversion. Converting models to ONNX format can be complex and time-consuming.
2. Performance Overheads. In some cases, there can be performance overheads in converting and running.
3. Version Compatibility. Ensuring compatibility with different versions of ONNX and ML frameworks can be challenging .
4. Might want to restructure the inference code to run Onnx model using onnx runtime

Development Flow using ONNX



Development flow using pytorch

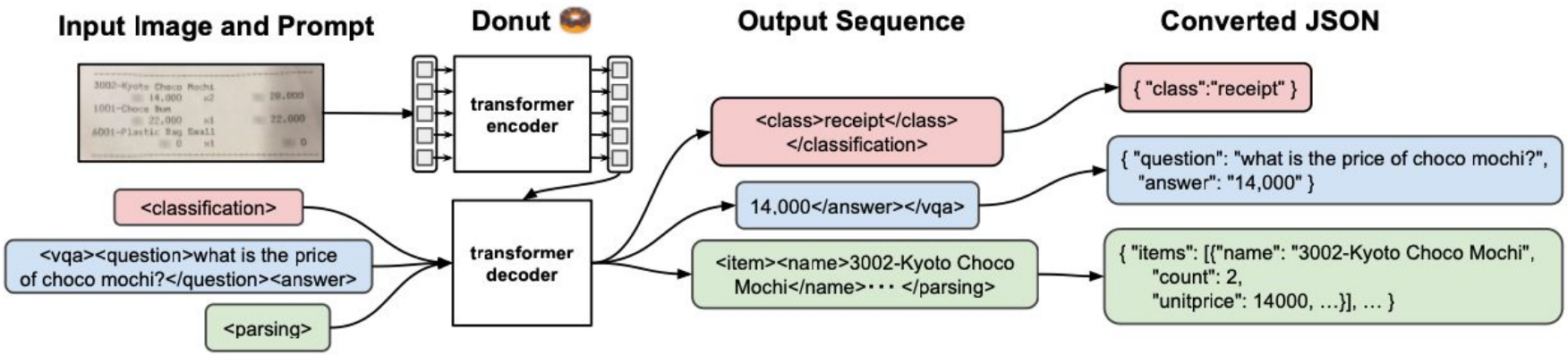


OCR (Optical Character Recognition)

Popular Models

1. PyTesseract
2. EasyOCR
3. TrOCR
4. DONUT
5. Paddle OCR

Donut(Document understanding transformer)



Results

```
{'text_sequence': 'Hey How are you doing This is Charan圖書館'}
CPU Total Time: 1.9429606000000001
{'text_sequence': ' I like to drink Brisk soda</s_changeprice></s_total>'}
CPU Total Time: 1.9721069000000009
{'text_sequence': 'Pikkal Pikkal Pikka! Pikka!</s_nm></s_total>'}
CPU Total Time: 1.9424080999999997
{'text_sequence': '"Right now I wish I was named Bob instead of Ash" Domini'}
CPU Total Time: 2.171251399999999
{'text_sequence': '"My dream is to become the greatest Pokemon masteri That way
CPU Total Time: 2.8705130999999984
```

```
{'text_sequence': 'Hey How are you doing This is Charan圖書館'}
NPU Total Time: 4.9293396000000005
{'text_sequence': ' I like to drink Brisk soda</s_changeprice></s_total>'}
NPU Total Time: 4.716383599999999
{'text_sequence': 'Pikkal Pikkal Pikka! Pikka!</s_nm></s_total>'}
NPU Total Time: 5.182793700000001
{'text_sequence': '"Right now I wish I was named Bob instead of Ash"'}
NPU Total Time: 5.857708900000002
{'text_sequence': '"My dream is to become the greatest Pokemon masteri That way
NPU Total Time: 12.6112675
```

Results

```
{'text_sequence': 'Hey How are you doing This is Charan</s_total>'}
GPU Total Time: 2.907313029000079
{'text_sequence': ' I like to drink Brisk soda</s_changeprice></s_total>'}
GPU Total Time: 0.5739927960000841
{'text_sequence': 'Pikkal Pikkal Pikka</s_total_price></s_total>'}
GPU Total Time: 0.4766519489999155
{'text_sequence': ' "Right now I wish I was named Bob instead of Ash飞机飞机飞机'}
GPU Total Time: 0.7711493949999522
{'text_sequence': '"My dream is to become the greatest Pokemon masteri That way the
GPU Total Time: 0.9299403290000328
```

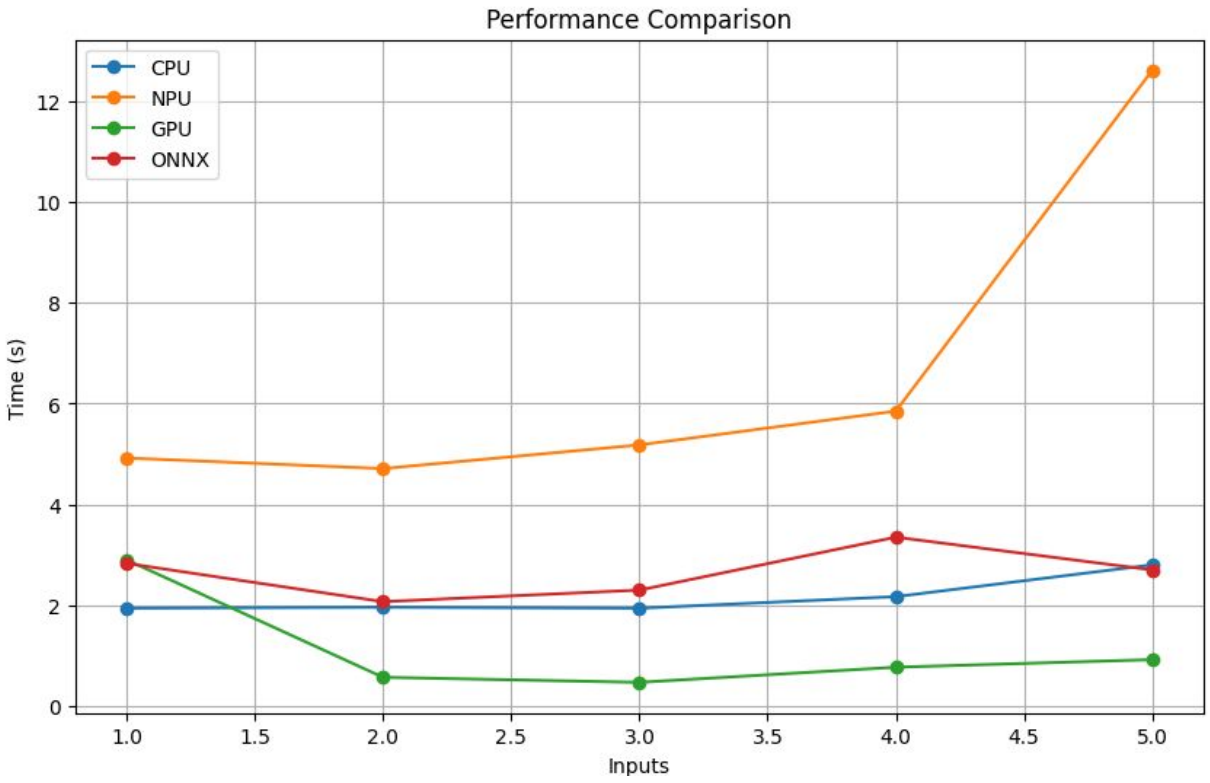
Results

```
[Vitis AI EP] No. of Operators :   CPU  6005    CPU   540    CPU   462
{'text_sequence': ' Hey How are you doing This is Charan editors emanded you doing momentalized
alism among actualism among a wealth are not wealth are not actual demoarch demoarch demoarch de
ng school demoarch domaine demoarch domaine demo'}
Onnx-CPU Total Time: 2.836526599999999
{'text_sequence': '设立 Brisk soda'}
Onnx-CPU Total Time: 2.0717595000000024
{'text_sequence': ' Pikkal Pikkal Pikka'}
Onnx-CPU Total Time: 2.3041193000000035
{'text_sequence': " 영화 was named Bob instead of Ash'I'll'I'll'll'll'll'll'll'll'll'll'll'll'll
Onnx-CPU Total Time: 3.358170600000001
{'text_sequence': ' wrote village "My dream is to become the greatest Pokemon masteri That way t
Onnx-CPU Total Time: 2.7186977
```

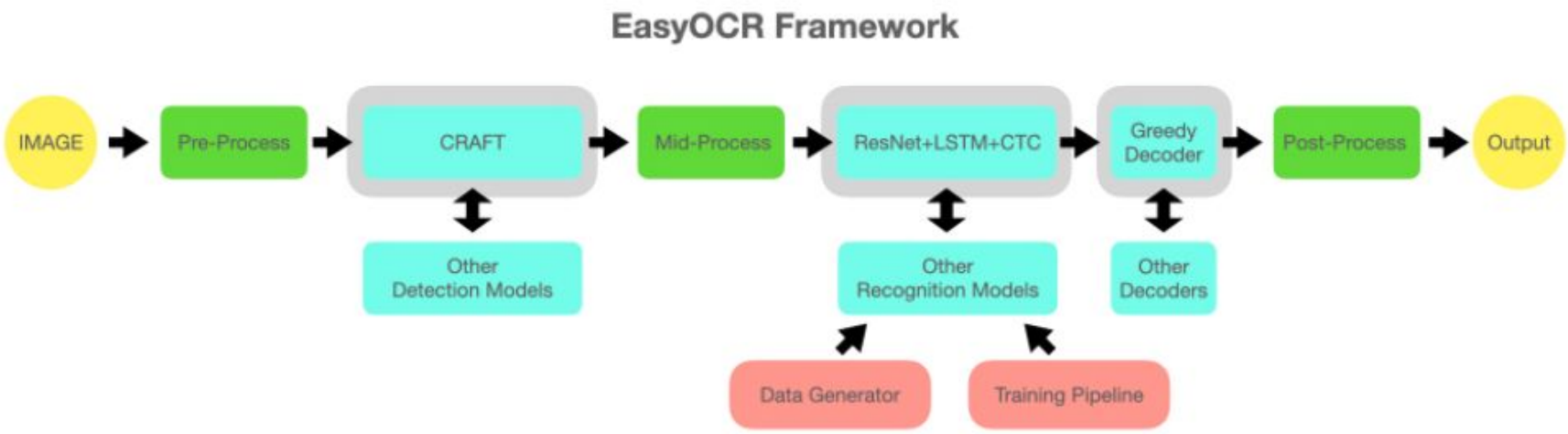
Comparison

CPU	NPU	GPU	ONNX
1.94s	4.92s	2.9s	2.83s
1.96s	4.71s	0.57s	2.07s
1.94s	5.18s	0.47s	2.30s
2.17s	5.85s	0.77s	3.35s
2.8s	12.61s	0.92s	2.7s

Graphical Analysis



EasyOCR


EasyOCR Framework
IMAGE
Pre-Process
CRAFT
Mid-Process
ResNet+LSTM+CTC
Greedy Decoder
Post-Process
Output
Other Detection Models
Other Recognition Models
Other Decoders
Data Generator
Training Pipeline

Results

```
['Hey Hou are you doing This Charan']
NPU Total Time: 0.20671510000000026
["Hello User! I'm Donut-OCR"]
NPU Total Time: 0.08585710000000013
['I like to drink Brisk soda']
NPU Total Time: 0.09162770000000009
['Pikka Pikka Pikka']
NPU Total Time: 0.07261190000000006
['Right now wish was named Bob" instead of Ash""']
NPU Total Time: 0.15565189999999962
['"My dream is to become the greatest Pokemon master! That way the whole vi
NPU Total Time: 0.2917114999999999
['"Pikachu, please stay Nith', 'forever']
NPU Total Time: 0.11504589999999926
```

Results

```
['Hey Hou are you doing This Charan']
CPU Total Time: 0.12622909999999976
["Hello User! I'm Donut-OCR"]
CPU Total Time: 0.07705270000000031
['I like to drink Brisk soda']
CPU Total Time: 0.07924259999999972
['Pikka Pikka Pikka']
CPU Total Time: 0.06389259999999997
['Right now wish was named Bob" instead of Ash""']
CPU Total Time: 0.13271760000000032
['"My dream is to become the greatest Pokemon master! That way
CPU Total Time: 0.2559649999999998
['"Pikachu, please stay Nith', 'forever']
CPU Total Time: 0.09894279999999966
```

Results

```
['Hey Hou are you doing This Charan']
GPU Total Time: 0.5963557500000434
["Hello User! I'm Donut-OCR"]
GPU Total Time: 0.07248877500001072
['I like to drink Brisk soda']
GPU Total Time: 0.07692647900000793
['Pikka Pikka Pikka']
GPU Total Time: 0.07177761200000532
['Right now wish was named Bob" instead of Ash""']
GPU Total Time: 0.10615609000001314
['"My dream is to become the greatest Pokemon master! That way
GPU Total Time: 0.13028006900003675
['"Pikachu, please stay Nith', 'forever']
GPU Total Time: 0.061609185999941474
```

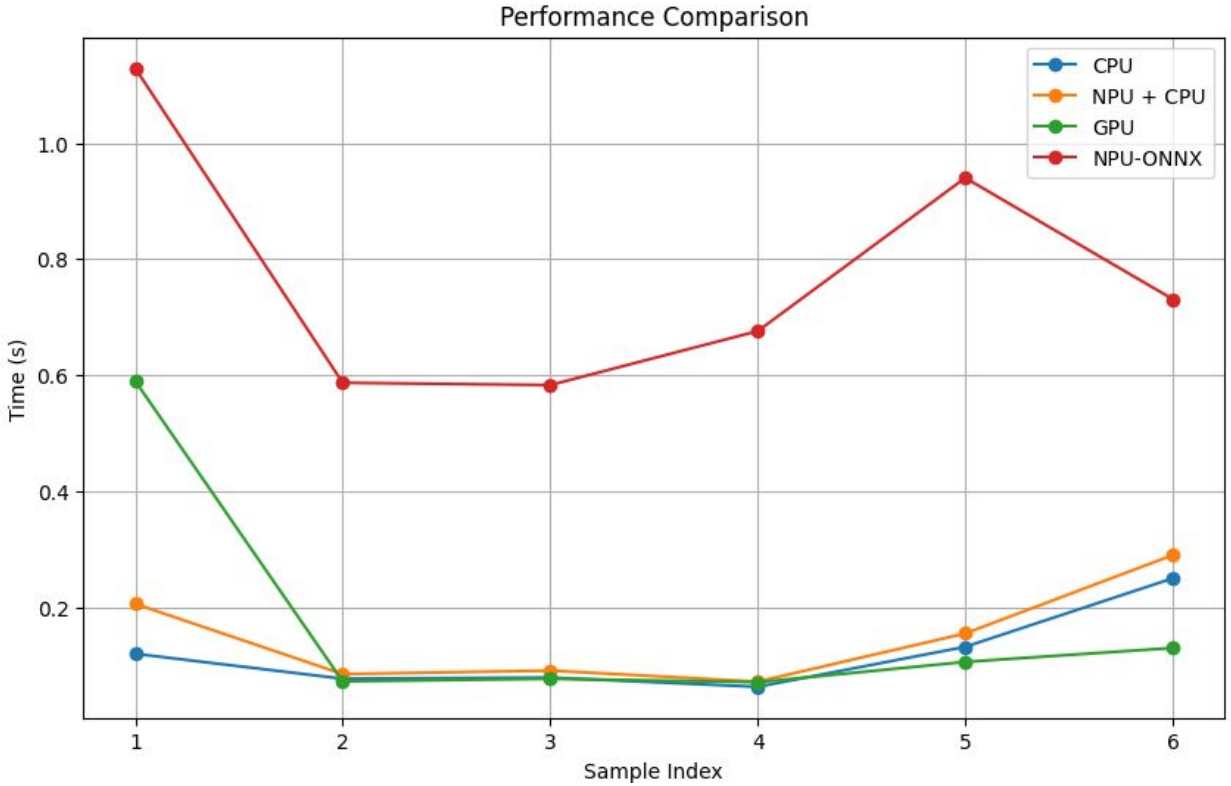
Results

```
Hey Hou are you doing This Charan
NPU-ONNX Processing Time: 1.1292216000000002
Hello User! I'm Donut-OCR
NPU-ONNX Processing Time: 0.5877799999999995
1 like to drink Brisk soda
NPU-ONNX Processing Time: 0.5839825999999997
Pikka Pikka Pikka
NPU-ONNX Processing Time: 0.6767198999999993
Right now wish was named Bob" instead of Ash""
NPU-ONNX Processing Time: 0.9409672000000002
"My dream is t0 become the greatest Pokemon master! That way
NPU-ONNX Processing Time: 0.7317740999999991
"Pikachu, please stay Nith
forever
NPU-ONNX Processing Time: 0.8428609999999992
```

Comparison

CPU	NPU + CPU	GPU	NPU-ONNX
0.12s	0.206s	0.59s	1.129s
0.077s	0.085s	0.0724s	0.587s
0.079s	0.091s	0.0769s	0.583s
0.063s	0.072s	0.071s	0.676s
0.132s	0.155s	0.106s	0.940s
0.25s	0.29s	0.130s	0.731s

Graphical Analysis




Demo

Future Work

- So far, I have tried multiple development flows, which at this point either don't run on NPU or have less utilization. Possible driver update from AMD could allow us to run the models on NPU effectively.
- Could work on Model optimization techniques.
- Improve model accuracy by fine tuning the model.
- Also trying different quantization techniques to increase model accuracy.