ANDREAS G. ANDREOU

http://andreoulab.net/
@andreoulab

Electrical and Computer Engineering,
Center for Language and Speech Processing
Johns Hopkins University
2022-2023: The Year of Generative-AI: Large Language Models and Stable Diffusion

A 1923 comic for New York World by cartoonist H. T. Webster (1885-1952)

ChatGPT 4, March 2024

TinyML Workshop on Generative AI and Foundation Models at the Edge

March 2024
Typical devices at the "Edge"

- Arduino Nano 33
- BLE Sense
- Kendryte K210
- Kendryte K510
- Rasp-Pico
- Arduino Pro
- Nicla Vision
- Google Edge TPU Dev Micro
- QuickFeather
- NVIDIA Jetson

March 2024
TinyML Workshop on Generative AI and Foundation Models at the Edge
but these as well ...


March 2024

TinyML Workshop on Generative AI and Foundation Models at the Edge

AI/ML in embodied/embedded systems at the EDGE

Sensing
World to Signals

Perception
Signals to Symbols

Knowledge
Contextual Model
Physical Reality

Learning Abstraction

Reasoning
Information to Action

Behaving
Actions to World

A.I. Machine

Model based systems with composability to drive and explain decisions as well as adaptation that is grounded in physical reality.

Neuromorphic Cognitive Computing
Action/gesture/gait recognition at the EDGE without a camera- like a bat

**Sensing**

Kinect

**Learning:** High dimensional data (Kinect cloud) + Low dimensional data (micro-Doppler signatures - time series)

**Inference:** Low dimensional data i.e. u-Doppler signatures - time series

**Models**

- High Dimensional Data (RGBD)
  - Skeletal Pose Sequences
  - Ultrasound Time-Series

- Low Dimensional Data (uDoppler signatures)
  - Rotation Representation
  - Spectrogram Representation

**Action Recognizer**

Train

Skeletal Pose Estimate

Action Label

Combined Ultrasound Percent Correct: 88.55%
Run time: LSTM Model

Default Training Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of epochs</td>
<td>200</td>
</tr>
<tr>
<td>Seed</td>
<td>1337</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.5</td>
</tr>
<tr>
<td>Max Sequence Length</td>
<td>404</td>
</tr>
<tr>
<td>Batch Size</td>
<td>100</td>
</tr>
</tbody>
</table>

March 2024 TinyML Workshop on Generative AI and Foundation Models at the Edge

Model | 5 Fold | LOUO |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>POE Baseline</td>
<td>87.8889</td>
<td>67.963</td>
</tr>
<tr>
<td>Stacked HMM</td>
<td>93.6296</td>
<td>89.0</td>
</tr>
<tr>
<td>LSTM 800</td>
<td>98.1852</td>
<td>95.6296</td>
</tr>
<tr>
<td>LSTM 1200</td>
<td>97.5926</td>
<td>95.7037</td>
</tr>
</tbody>
</table>

POE: Product Of Experts
LOUO: Leave One User Out
Learning new actions and new end users at the EDGE

Why?
Do learning at the EDGE to:
1. Enroll new end users and actions
2. Alleviate data-center energy expenses. i.e. end user pays for electricity (the problem with Alexa business model; Amazon pays!)

Technical Challenge:
When deployed at the EDGE, high dimensional data (RGBD) are not available to facilitate learning of novel actions/gait/gestures or new users.
1. We need a model base approach that is grounded to physical reality.
2. Need a generative model to create novel poses and introduce stochasticity.
Contextual Modeling: Physics and Generative AI

Knowledge
Contextual Model
Physical Reality

A.I. Machine

Learning
Abstraction

Conditional Deep Belief Network (CDBN)

CRBM
Gaussian CRBM

Single Time Slice
(Without Auto-Regressive Connections)

Time

Generative AI model

Physics Model
Hallucinating skeletal poses and micro-Doppler signatures

Use these generated micro-Doppler signatures with local zero/few shot learning to adapt the models.
Hallucinating can be troublesome!

Real micro-Doppler data

Hallucinating micro-Doppler data
(sometimes it is missing or has extra limbs!)


March 2024

TinyML Workshop on Generative AI and Foundation Models at the Edge
LLMs for creating new generation of AI chips
Writing chips using natural language and LLMs: Design a Spiking Neural Network chip

Neuromorphic chip design using chatGPT

Now that we have a working neuron, can you create a new module that instantiates a network of neurons with 2 layers in a fully connected fashion?

Certainly! Let's define a simple 2-layered neural network with neurons fully connected between the layers. For simplicity, let's assume:

1. Layer 1 has \( N \) neurons.
2. Layer 2 has \( M \) neurons.

I don't think Verilog allows for 2D arrays in the port list, can you fix this?

You're right. Verilog does not support 2D arrays for port lists. I apologize for the oversight. We can fix this by using 1D arrays (vectors) to represent the input currents and spike outputs.

Let's modify the design to use concatenated vectors for input currents and spikes:
Conclusions

• Fully LLM written Spiking Neuron Array with SPI interface.
• Generated Verilog verified with functional simulation
• Design was submitted on Tiny Tapeout 5 using an open source RTL-> GDS flow
• 33% Density in a 320um x 200um tile
• Chips expected back on July 30 for testing
• All code, scripts, and ChatGPT transcripts are available on Github.

• One of the first fully AI described ASICs
• Not an easy process, but a fully Natural Language -> Verilog flow is possible
• For now, it requires extensive verification
• Future work – Given a top module can LLMs automate functional verification & stimulus generation?
Variations of our team picture generated by Dall-E
Telluride Neuromorphic Engineering Workshop

Topic Areas 2024

The core of Telluride is broken into separate Topic Areas. Each Topic Area is guided by a group of experts who will provide tutorials, lectures, and hands-on project ideas. Participants should expect to be exposed to all of the Topic Areas, but will generally focus on one or two to work on during the event; see past workshops for examples.

There are also a series of (mostly) neuroscience talks during 2nd week; see below.

NIC24
Neuromorphic integrated circuits

AUD24
Understanding the auditory brain with neural networks

L&T24
Language and thought

SPA24
Neuromorphic systems for space applications

CNS24
Computational Neuroscience Talks

30th year anniversary:
Join for 3 weeks of fun @ 9000 feet

March 2024
TinyML Workshop on Generative AI and Foundation Models at the Edge