FPGA Deployment of LFADS for Real-time Neuroscience Experiments

Elham E Khoda

University of Washington, Seattle

Xiaohan Liu, ChiJui Chen, YanLun Huang, LingChi Yang, Yihui Chen,
Scott Hauck, Shih-Chieh Hsu, Elham E Khoda, and Bo-Cheng Lai.

Accelerated AI Algorithms for Data-Driven Discovery

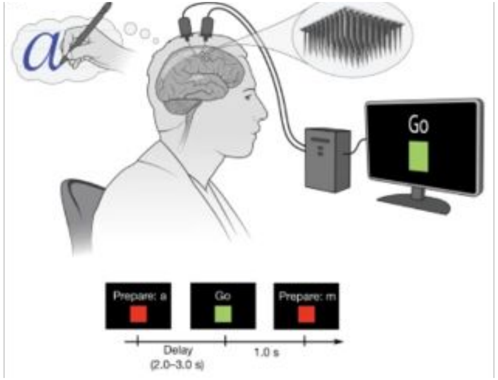
2nd November, 2023

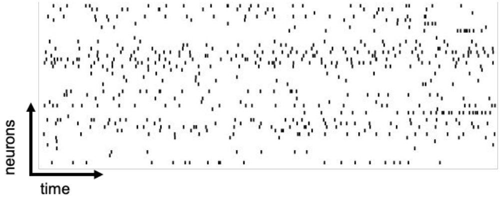
FastML@ICCAD 2023

UNIVERSITY of WASHINGTON

Introduction

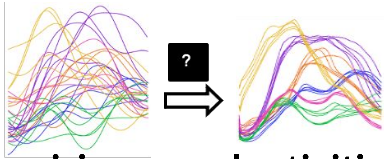
- Brain encodes behaviors through neural activities
- **Information about cognitive and motor processes is distributed within many neurons in the brain**
- Understanding this neural code can help us relate the neural activity to behavior
  - Neural activities are noisy
  - → *Encoding behaviors is very difficult*
- Need algorithms that model the neuron activity to uncover the underlying dynamics
  - Clean version





Latent Factor Analysis via Dynamical Systems (LFADS)

- SOTA for inferring single-trial neural dynamics

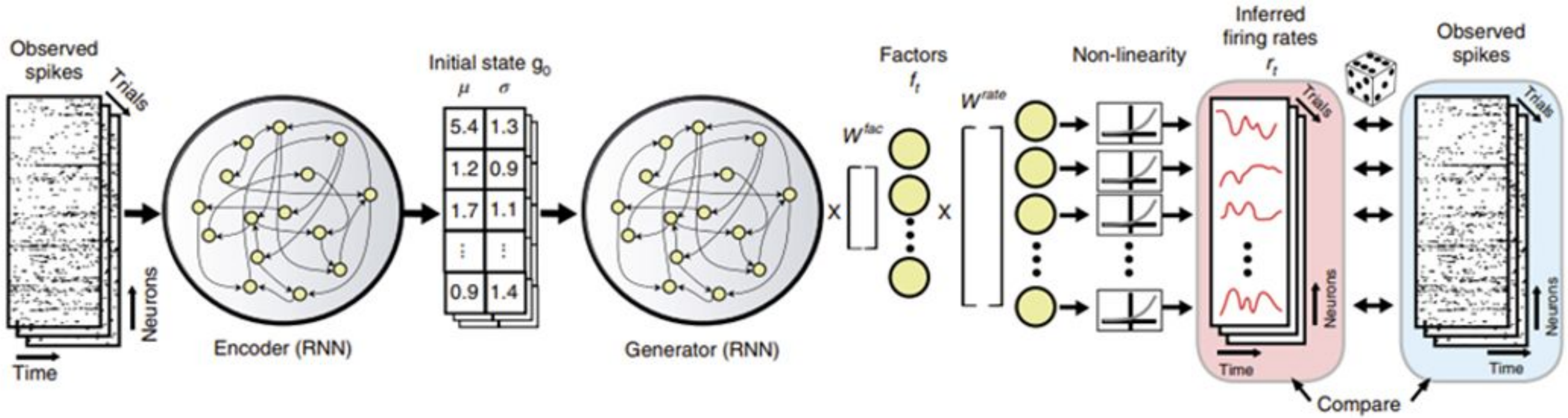


Denoising neural activities

Latent Factor Analysis via Dynamic Systems (LFADS)

LFADS is a sequential model based on Variational Autoencoder

- LFADS assumes the observed spikes are samples from a Poisson process with firing rates
- Decoder learns the firing rates a function of time
- **Training objective**: Decoder is trained to infer a reduced set of latent dynamic factors

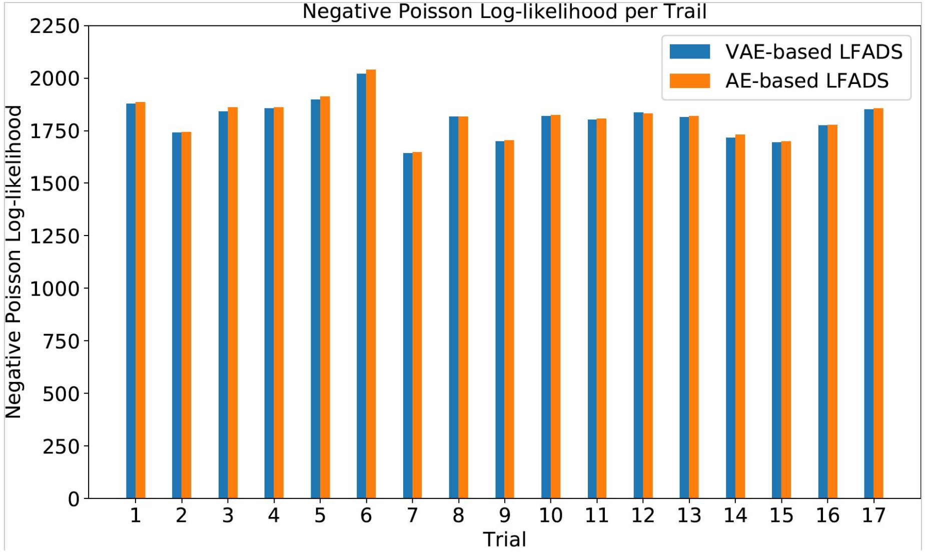


Autoencoder-based LFADS

Started with a simplified model:

Variation Autoencode → Autoencode

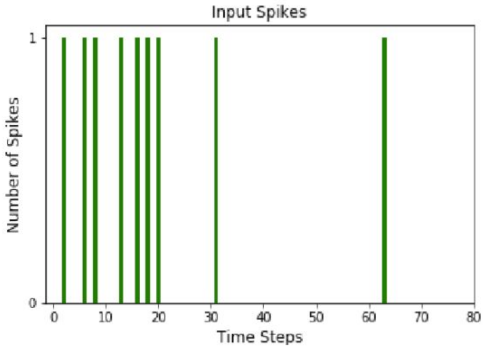
- No random sampling on FPGA
  - Making it much easier to deploy
- It has minimal effects on performance

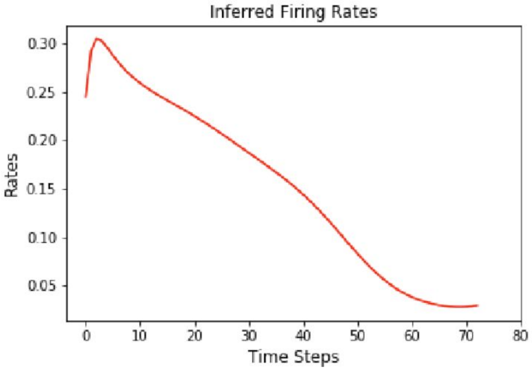


Autoencoder-based LFADS

Autoencoder architecture with

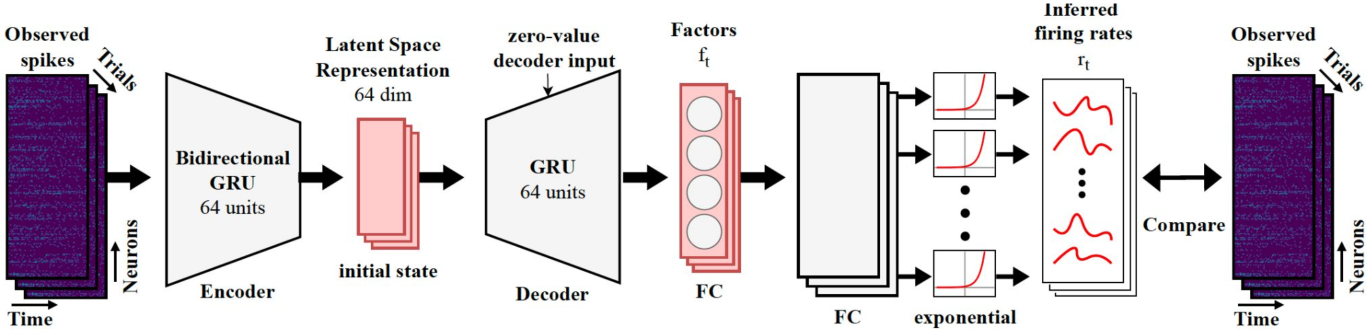
- Bidirectional GRU Encoder
- GRU Decoder





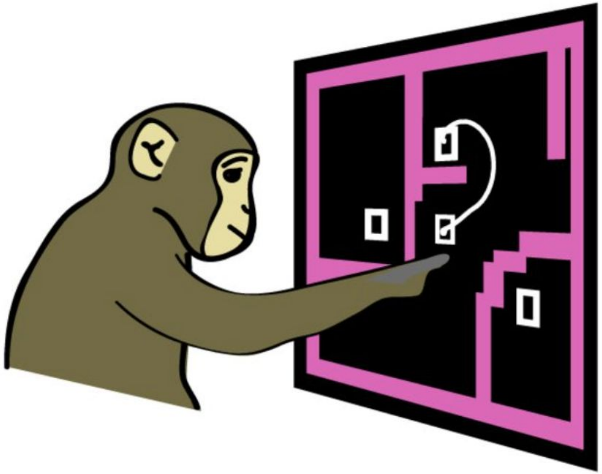
Key features:

- Input: Sequential spiking data
- Output: Firing rate



Experimental Data

- **Monkey reaching tasks****
  - Perform a center-out reaching task with eight outer targets
  - Spiking activity from the primary motor cortex (M1) along with the 2D hand position are recorded during each trial



* **Gallego Juan A, Perich Matthew G, Chowdhury Raeed H, Solla Sara A, Miller Lee E. Long-term stability of cortical population dynamics underlying consistent behavior // Nature Neuroscience.* 2020. 23, 2. 260–270.

Experimental Data

- **Monkey reaching tasks****
  - Perform a center-out reaching task with eight outer targets
  - Spiking activity from the primary motor cortex (M1) along with the 2D hand position are recorded during each trial

- **Dataset***
  - total of 170 trials.
  - 136 trials (80%) for training
  - 17 trials for validation and testing
  - Each trial with shape (1,73,70): 70 recording channels, with 73 discrete time steps per channel

*Cole Hurwitz, Akash Srivastava, Kai Xu, Justin Jude, Matthew G. Perich, Lee E. Miller, and Matthias H. Hennig. 2021. Targeted neural dynamical modeling.(2021). arXiv: 2110.14853 [q-bio.NC].

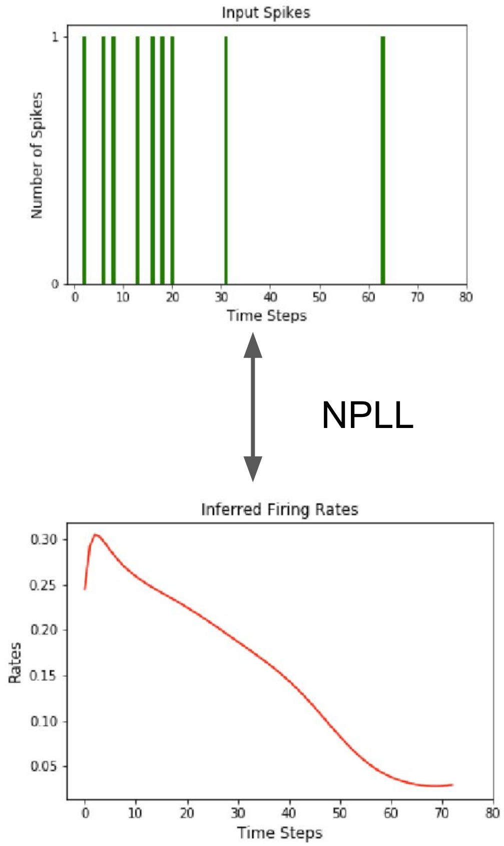
* **Gallego Juan A, Perich Matthew G, Chowdhury Raeed H, Solla Sara A, Miller Lee E. Long-term stability of cortical population dynamics underlying consistent behavior // Nature Neuroscience.* 2020. 23, 2. 260–270.

Model Performance Evaluation

Two Metrics are used for this study

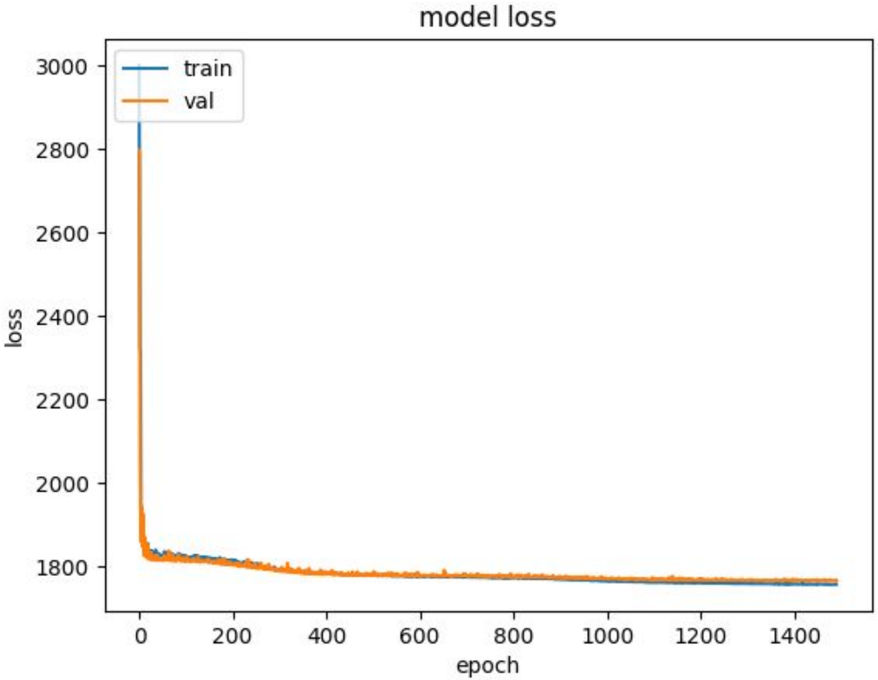
- Negative Poisson log-likelihood (NPLL)
  - Between the predicted log firing rates and input spikes
  - LFADS assumes spiking variability follows a Poisson distribution

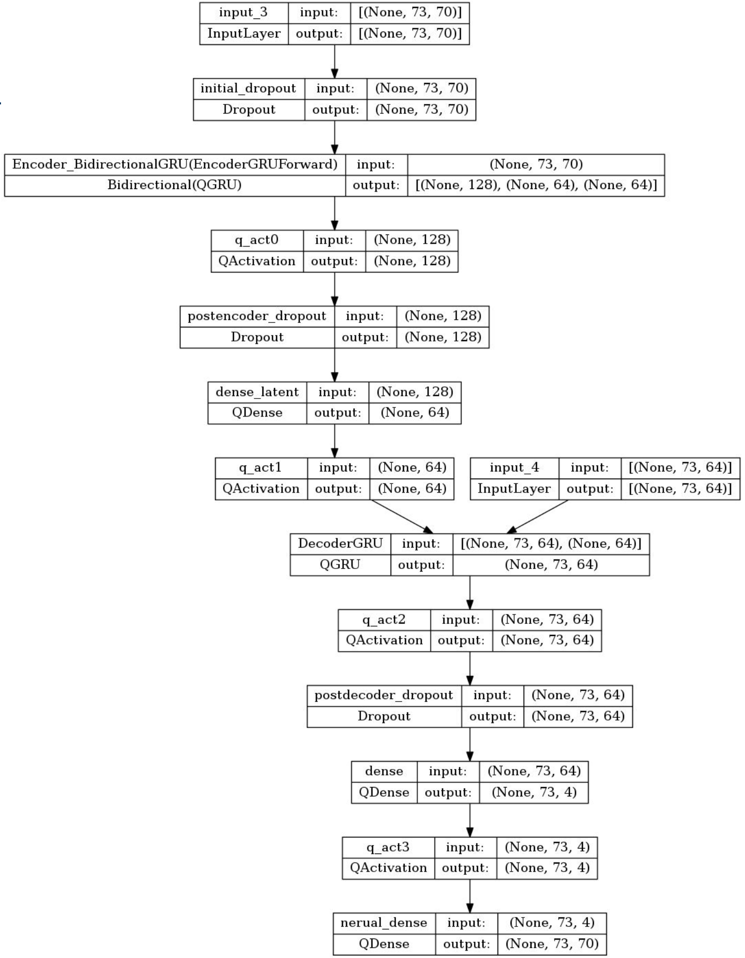
- Coefficient of determination (R2 score)
  - Fitting the reconstructed temporal factors ft to the measured behavioral data (hand position)
  - Training set for the linear regression model, fit on test set
  - A score closer to 1: stronger alignment of the factors with the behavioral data



Show TF model training

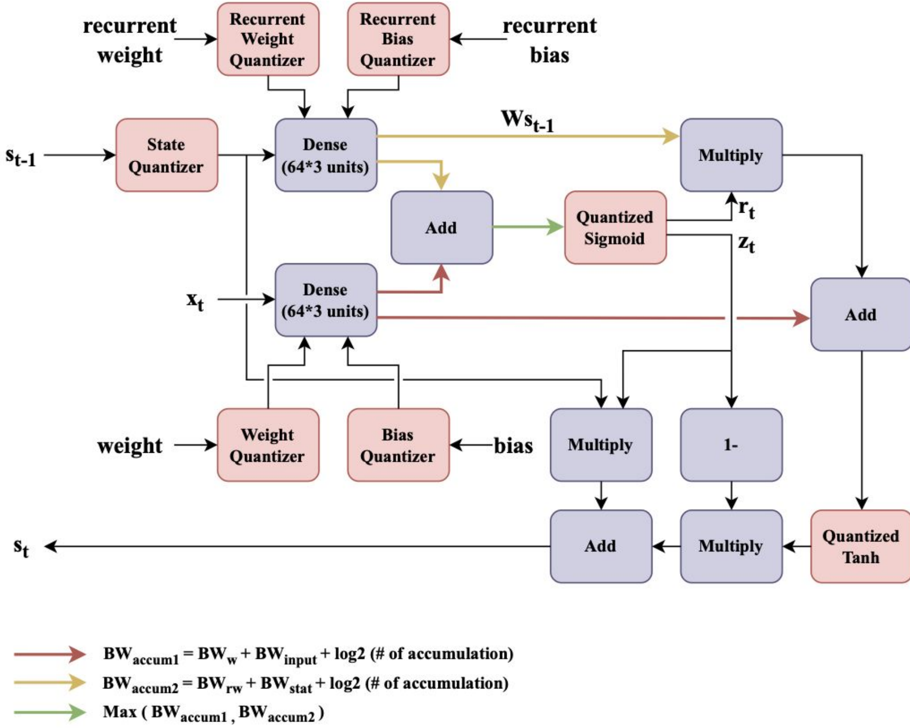
Loss = Poisson Log-likelihood loss

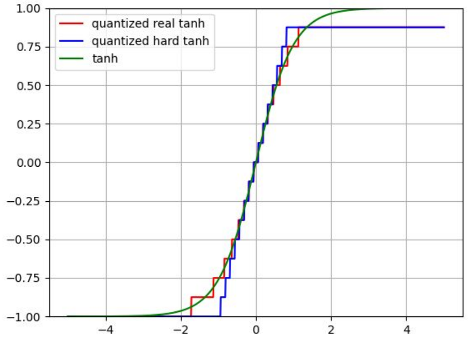


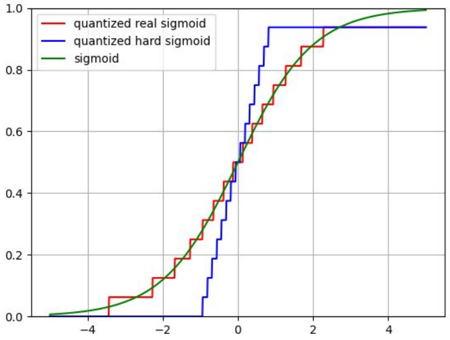


Quantization -Aware Training

- QAT using QKeras
- To minimize quantization error and accuracy drop
  - State, weight and bias quantizer
  - Adopting piecewise linear hard activation to eliminate quantization error
  - Automatic adjustment for bitwidth on accumulator to avoid overflow

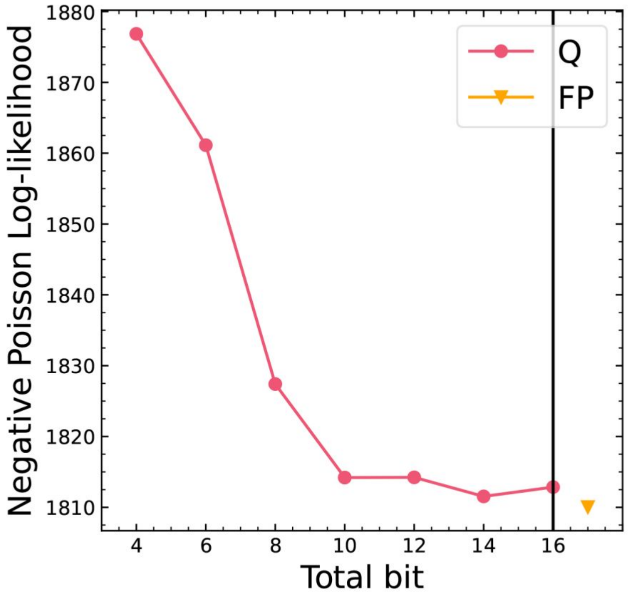




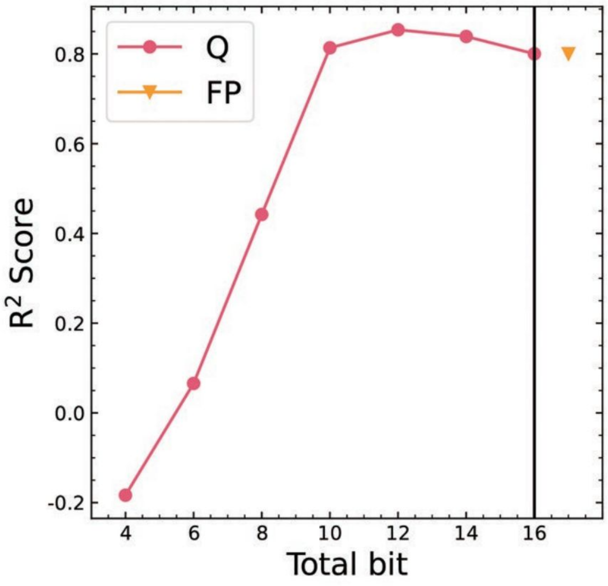


QAT Results: Total bit-width scan

Noticeable degradation in performance **below total width of 10 bits** in both NPLL and R2 Score



(a) QAT NPLL

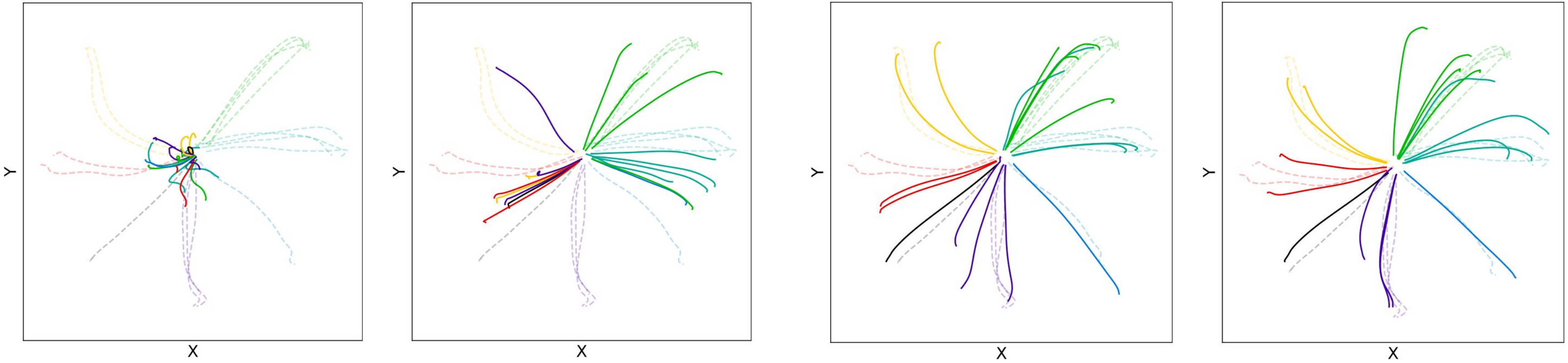


(b) QAT R^2

Behavioural Reconstruction

- Similar degradation in behavior reconstruction
  - The hand movement trajectories in the 2D $x - y$ plan
  - Same direction are grouped together and denoted by the same color

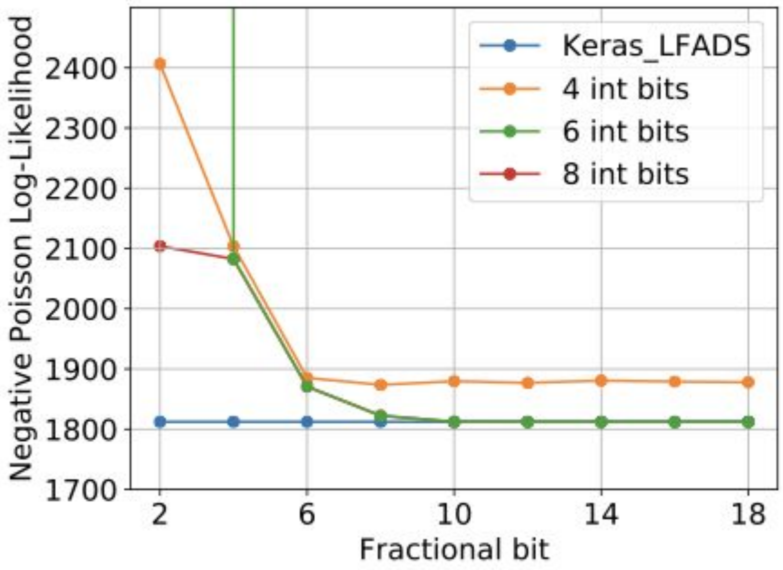
Dotted = target, Solid = predicted

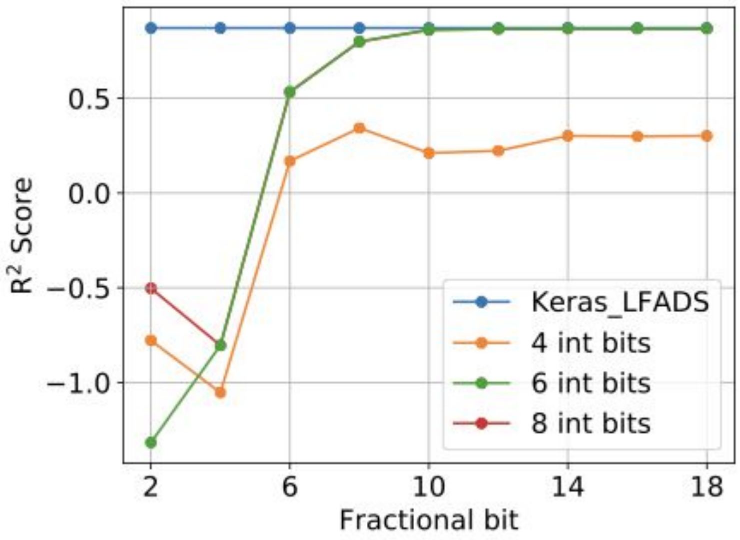


(a) 4 bits (b) 8 bits (c) 12 bits (d) floating point

Post-training Quantization

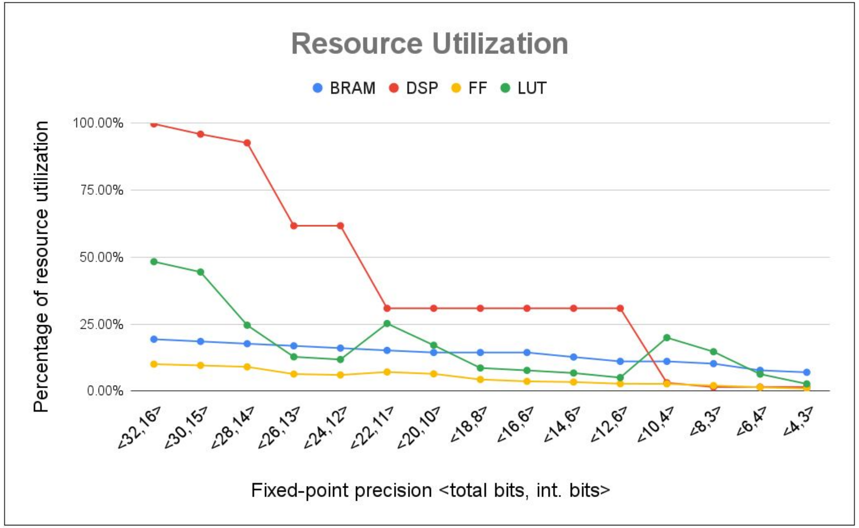
- At least **6 integer bits and 10 fractional bits**, <16,6>, are needed to achieve a similar performance as the floating-point model.





Resource Unitization

- Post training quantization (PTQ)
- Logic synthesis result
- Target platform : Alveo **U250**
- The limitation of FPGA inference for higher bit width is DSPs



LFADS on FPGA

- Target platform : U55C (NRP)
- Precision: ap_fixed<16,6> , Frequency=200 MHZ, apply dataflow scheme
- Average latency : 41.97 us

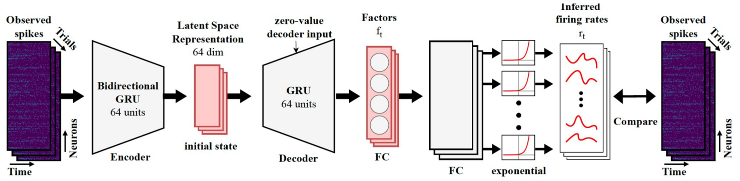
  (Run 1000 times and calculate the average)

V synthesis	U55C (NRP)
HLS version	2022
BRAM	474 (23.51%)
DSP	1,869 (20.71%)
FF	150,882 (5.79%)
LUT	164,726 (12.64%)

Summary and Outlook

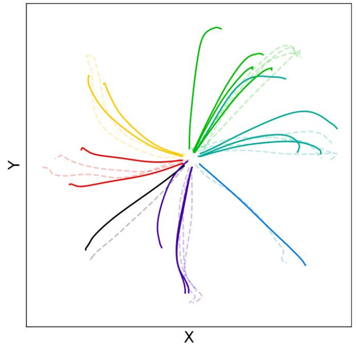
One of the first FPGA deployment of LFADS model

- Shown results of a simplified LFADS model with Autoencoder structure
- Quantization of GRU layers are implement and optimized
- We are able to fit the best model within a board
  - We can fit the model in the Alveo U55C
- Improve inference latency by 1000 times
  - Observed latency is ~42 micro-seconds



Next steps:

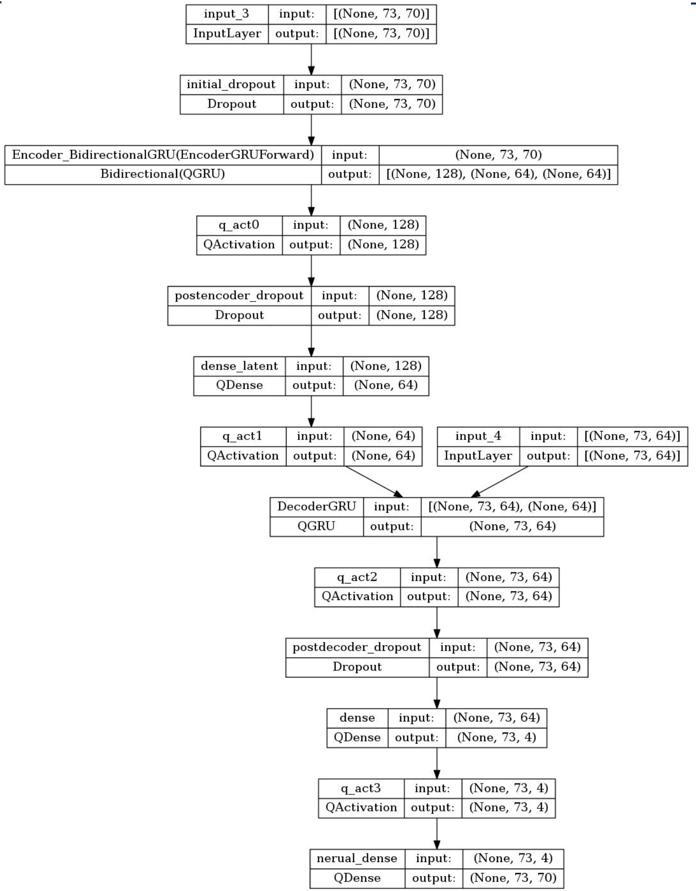
Deployment of the original Variational Autoencoder-based LFADS



Backup

Model Architecture

```
Model: "lfads"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dropout (Dropout)           multiple                  0

 EncoderRNN (Bidirectional)  multiple                  52224

 dropout_1 (Dropout)         multiple                  0

 dropout_2 (Dropout)         multiple                  0

 DenseMean (Dense)           multiple                  8256

 DenseLogVar (Dense)         multiple                  0 (unused)

 activation (Activation)     multiple                  0 (unused)

 DecoderGRU (GRU)            multiple                  24960

 Dense (Dense)               multiple                  256

 NeuralDense (Dense)         multiple                  350

=================================================================
Total params: 86,059
Trainable params: 86,046
Non-trainable params: 13
_________________________________________________________________
```



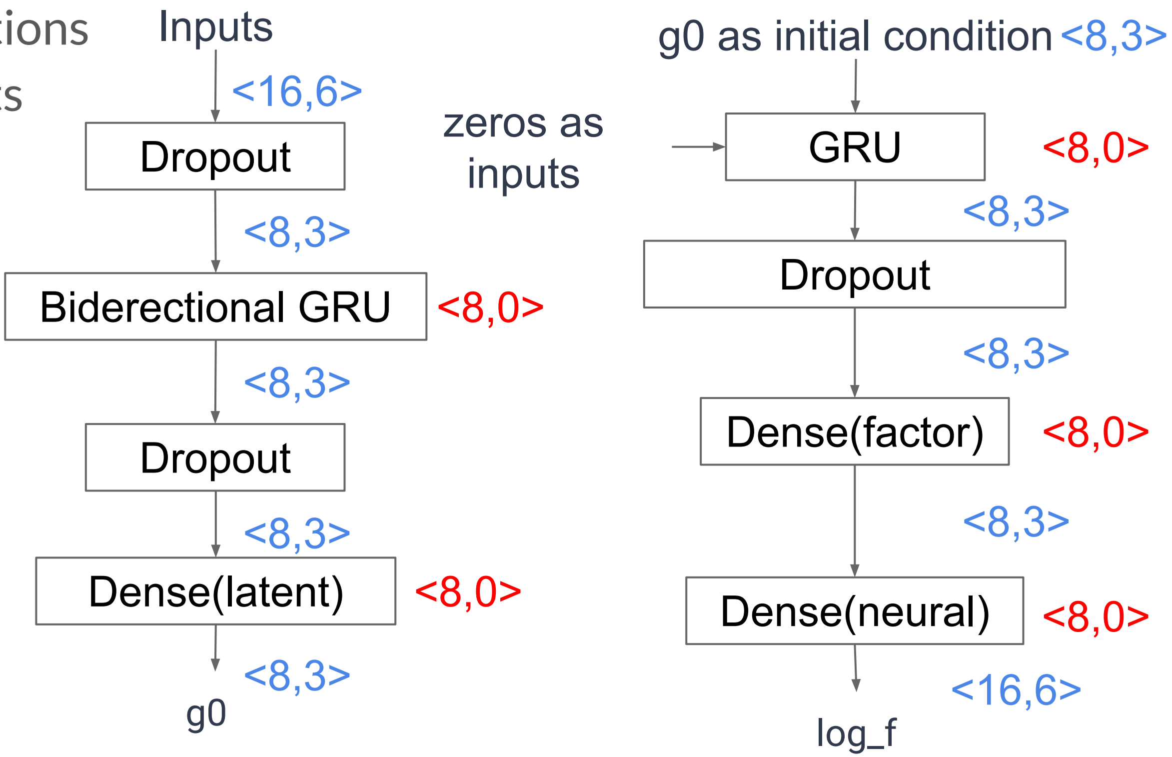
Different Data Transmission Scheme

- **IO_parallel**
  - In order to access all input (output) at a cycle, it needs to do array_reshape
  - Doing array_reshape complete in GRU layer will beyond the limit 65536.
  - Total bits width in Input : 73x64x16bits = 74752bits > 65536
  - It can’t be synthesized.
- **IO_stream**
  - Input (Output) is transmitted sequentially.
  - It doesn’t need to do array_reshape to access the input (output) at a cycle.
  - It can be synthesized and even apply for larger size.

QAT Precision

- Input, output <16,6>
- 3 integer bits for activations
- 0 integer bits for weights

8-bit model:



LFADS

Latent Factor Analysis via Dynamical Systems (LFADS)

- LFADS models the complex brain activities
    - Brain are extremely complex, which is extremely hard to model
- LFADS combines feed forward processing and sequential processing
    - Feed forward processing purely depends on input
    - Sequential processing mainly depends on dynamic
- Low latency processing provides the possibility of real-time data processing