Exploration of TPU Architectures for Transformer Performance Optimization in Drought Monitor Classifications

TEAM 1

Michael, Dipesh & Sushil

Project Description

- **Drought: A Complex Natural Hazard**
  - Definition: Drought is a prolonged period of abnormally low rainfall, leading to water scarcity.
  - Complexity: Despite its common occurrence, drought remains one of the least understood natural hazards.

- **Impacts**
  - Environmental: Disruption of ecosystems, loss of biodiversity.
  - Societal: Water scarcity, food insecurity, displacement of communities.
  - Economic: Crop failures, livestock losses, impacts on industries reliant on water.

Types of Drought

- **Meteorological Drought:**
  - Cause: Less rainfall than usual.
  - Impact: Reduced water in the air.

- **Agricultural Drought:**
  - Cause: Not enough water for plants.
  - Impact: Crop failures, food shortages.

- **Hydrological Drought:**
  - Cause: Not enough water in rivers and underground.
  - Impact: Less water for drinking, irrigation, etc.

- **Socioeconomic Drought:**
  - Cause: Not enough water for economic needs.
  - Impact: Water scarcity affecting businesses and livelihoods

How is Drought monitored?

The U.S. Drought Monitor:

- Collaborative effort by several government agencies and academic institutions.
- Provides weekly updates on drought conditions across the United States.

How It Works:

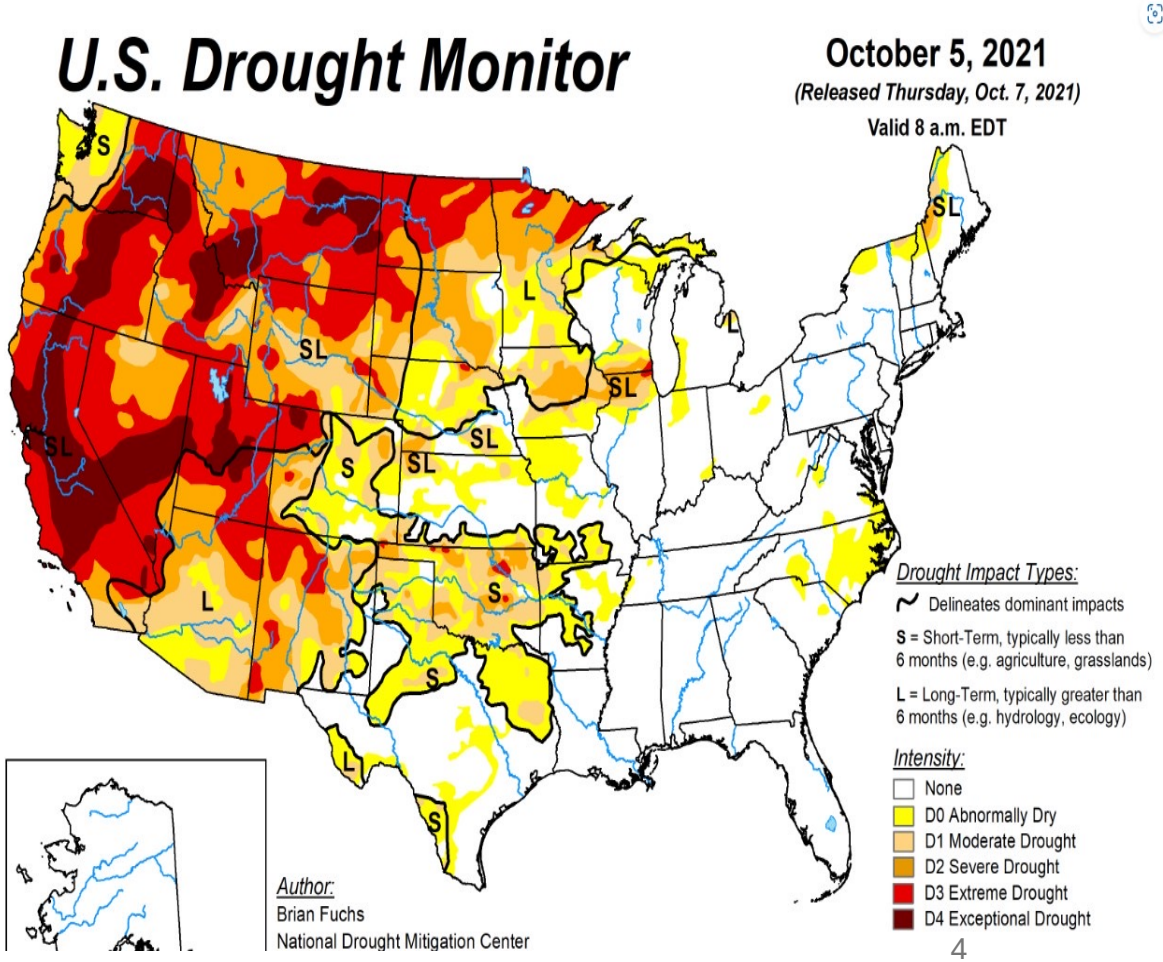
- Integrates various data sources, including satellite observations and ground-based measurements.

Importance:

- Helps policymakers, farmers, and communities make informed decisions.
- Guides allocation of resources for drought response and mitigation efforts.

Accessibility:

- Available online for public access and use.
- Provides detailed information at national, regional, and local levels.



Project Objectives

- Enhance drought monitoring with remote sensing methods for efficient decision-making.
- Investigate TPU architectures with SCALE-Sim to reduce latency in Transformer models.
- Enhance processing efficiency for USDMs by implementing optimized configurations identified in Phase 2.

Study Area and Drought Indicators

- **Colorado River Basin**
  - The Colorado River Basin is a vital water source serving seven U.S. states and Mexico.
  - Facing its worst drought in recorded history, impacting ecosystems, agriculture, and indigenous communities.
  - 246,000 square miles

Drought Indicators

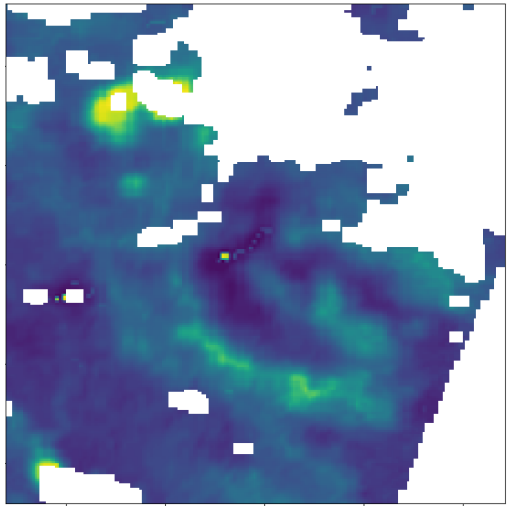
- Remote Sensing Variables
- Collected using Goggle Earth Engine
- July 2015 – July 2019(5 years)

Data	Source	Temporal Resolution	Spatial Resolution	Description
Precipitation	ERA5	1	27830 meters	Total precipitation (daily sums)
Temperature	ERA5	1	27830 meters	Average air temperature at 2m height (daily average)
Soil Moisture	NASA	1	9K	Retrieved soil moisture estimate from the disaggregated/downscaled vertical polarization brightness temperature at 9-km grid cell;
Snow Cover	MODIS	1	500m	Provide Normalized Difference Snow Index (NDSI) values from 0 – 100
NDVI	MODIS	16	500M	Normalized Difference Vegetation Index
US Drought Monitor Raster	USDM	7	-	

Task 1

- Use Transformer model architecture for spatiotemporal meteorological forecasting.

- **Dataset Used:**
  - Satellite images of SMAP (Soil Moisture Active Passive) by NASA.
  - SMAP measures the amount of water in the top 5cm of soil.
  - Satellite images taken daily from June 1, 2015 to July 29, 2019.



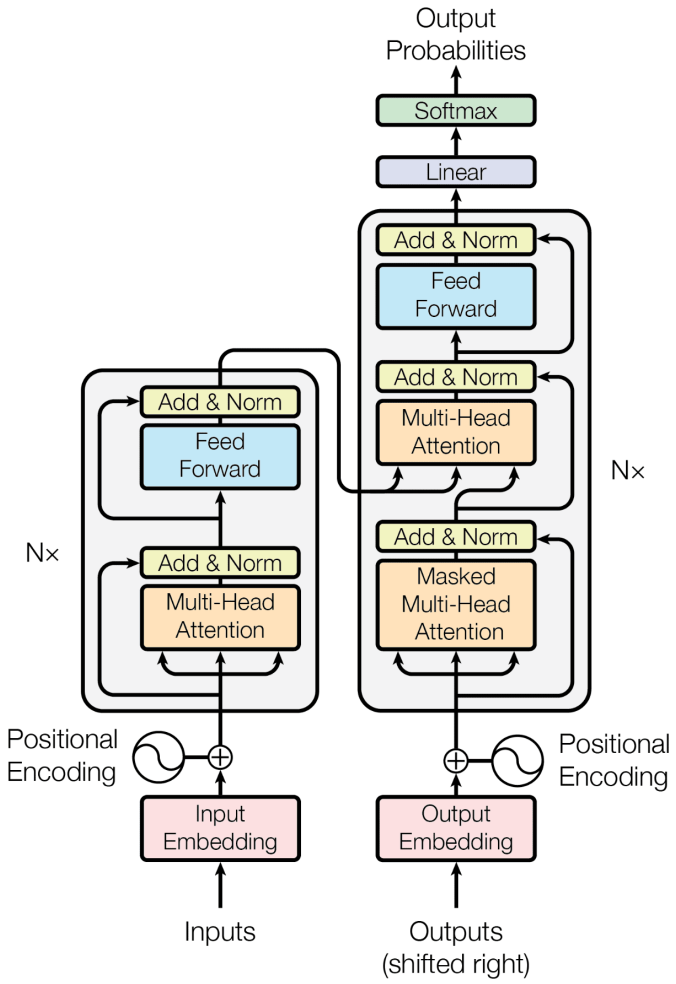
Transformer

- Neural network architecture designed to handle sequential data, primarily used for NLP tasks.

- **Core Components:**
  - Self Attention Mechanism
  - Positional Encoding
  - Multi-Head Attention

- **Challenges with image data:**
  - High dimensionality
  - Complexity in positional encoding
  - Does not capture hierarchical features

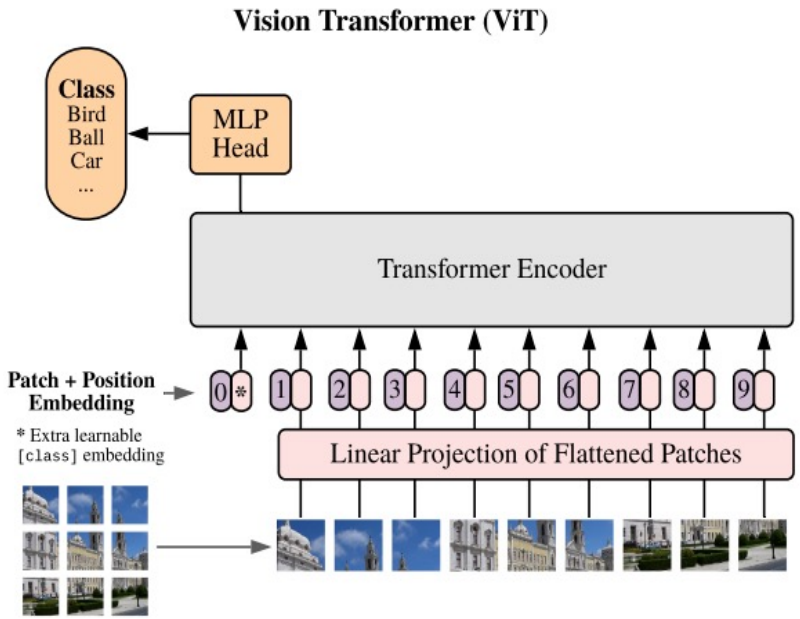
- Adaptation for Image data: **Vision Transformer**



Vision Transformer

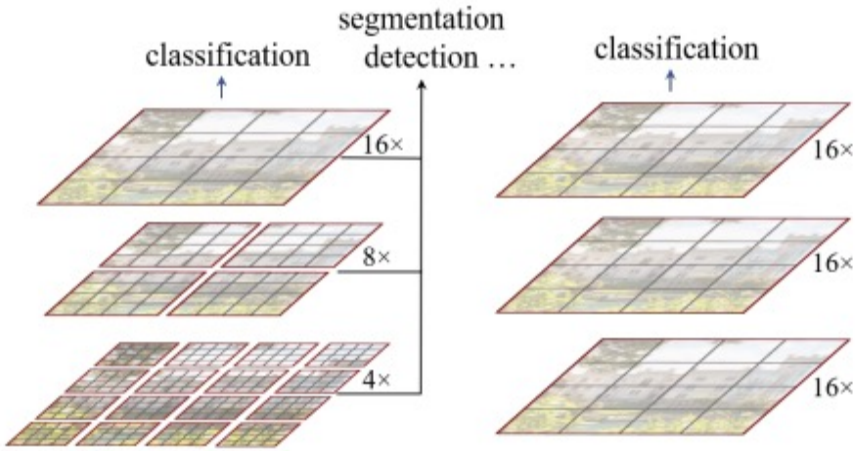
- Employs a transformer like architecture over patches of image
- Each patch is treated as token in sequence
- Patch Embedding and positional encoding
- MSA – Captures relationships between all patches, which enables model to capture global context.

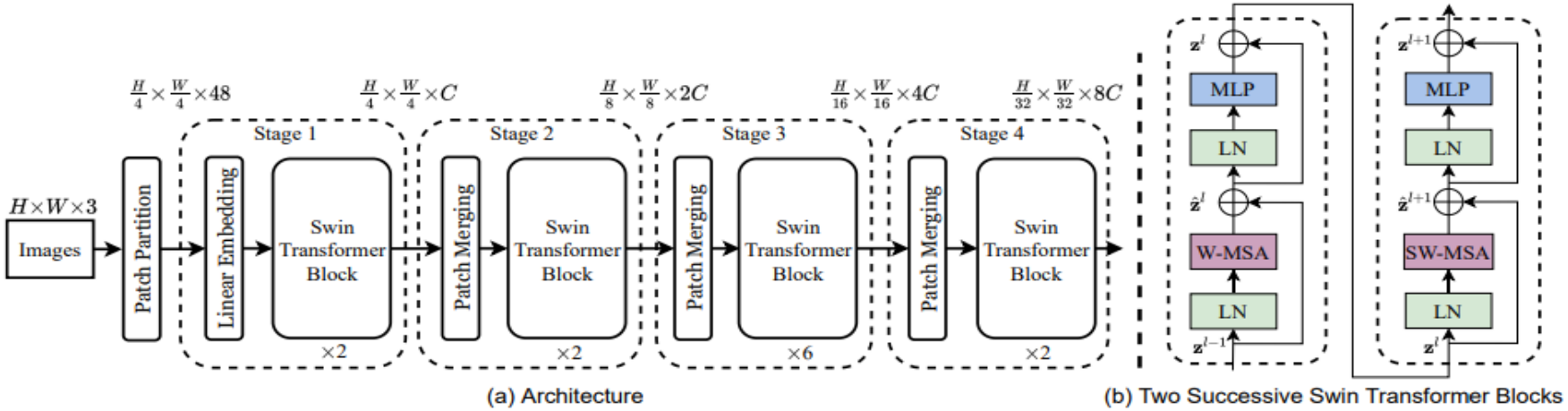
- **Challenges:**
  - High computational cost: Quadratic complexity w.r.t number of image patches
  - Difficulty in handling high resolution images



Swin Transformer

- Introduces hierarchical and efficient local attention mechanism
- Linear computational complexity w.r.t image size
- Shifted windows Attention mechanism
  - Facilitates interaction between patches across different windows



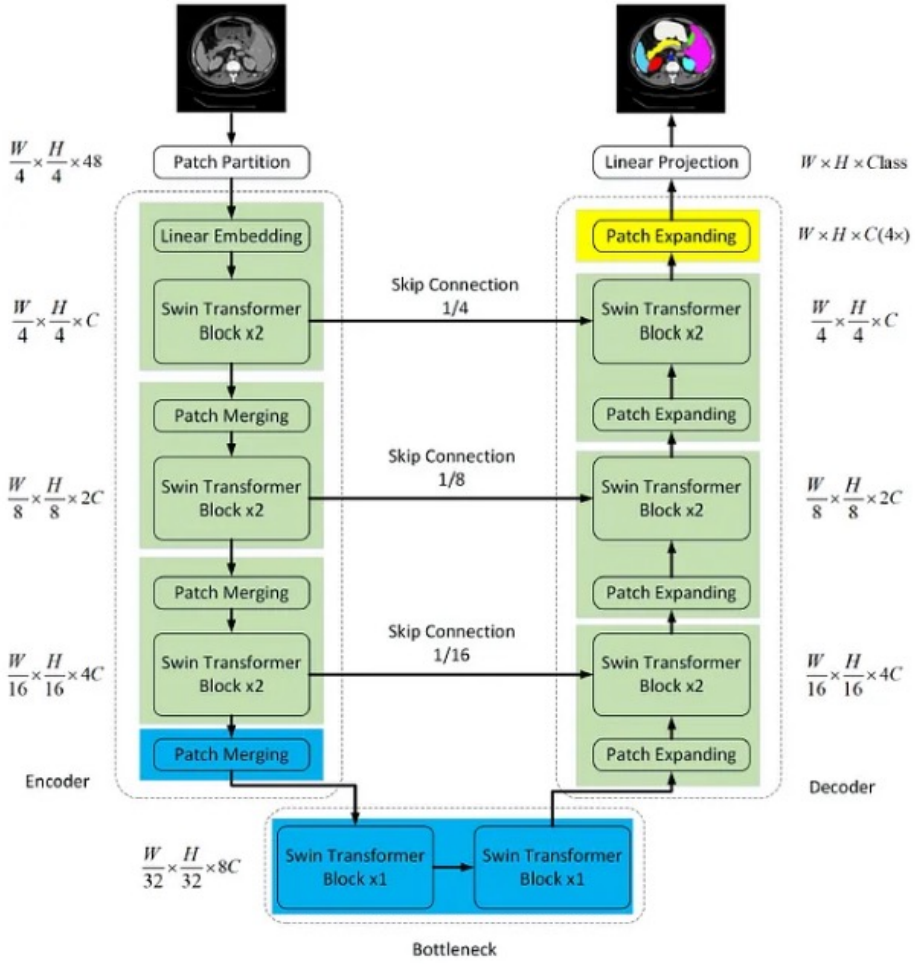


(a) Architecture

(b) Two Successive Swin Transformer Blocks

Swin-Unet

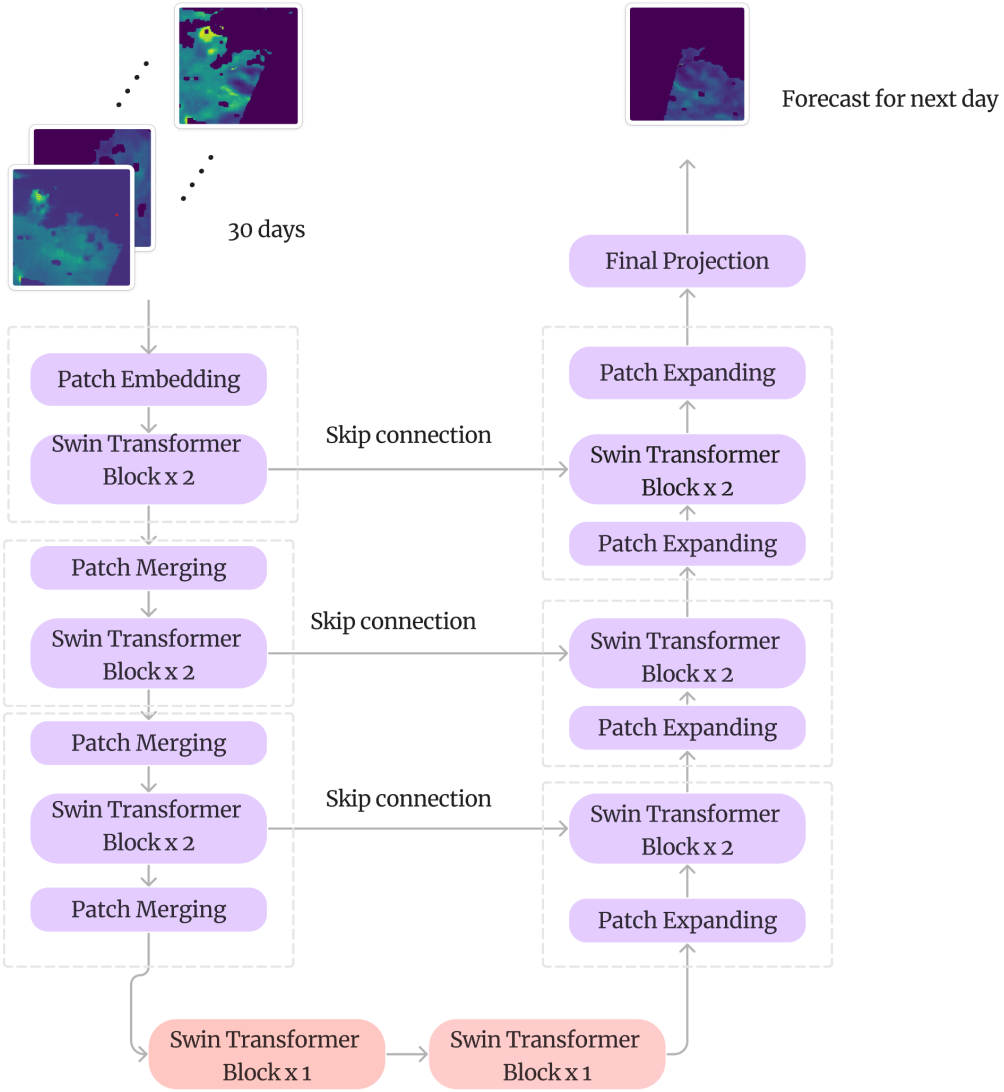
- Integrates swin transformer block with U-shaped architecture
- Designed for medical image segmentation
- Symmetric decoder block is added
- Patch expanding layer is used to upsample the extracted deep features
- Skip connection are used to fuse the multiscale features from the encoder with the up sampled features.



Swin-Unet Architecture

Swin-Unet : Our Model

- Image size = 224 x 224
- Patch size = 4 x 4



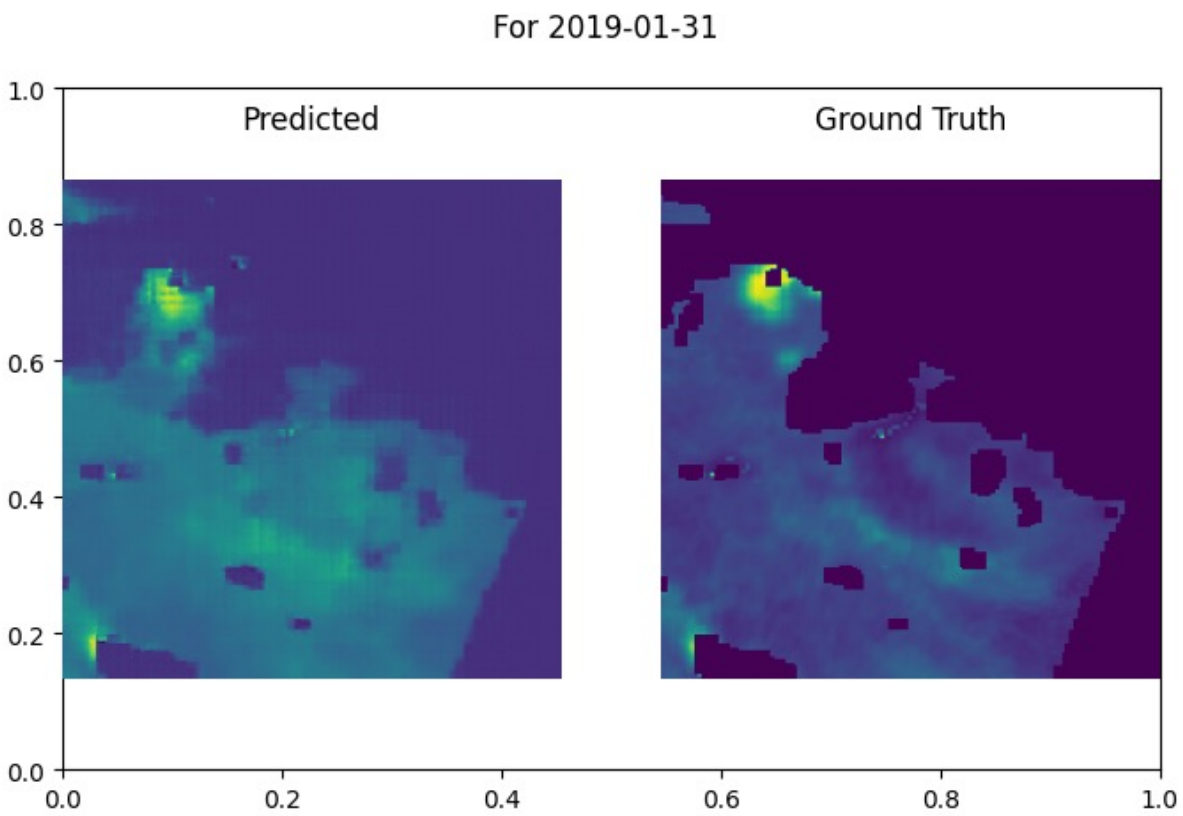
Implementations

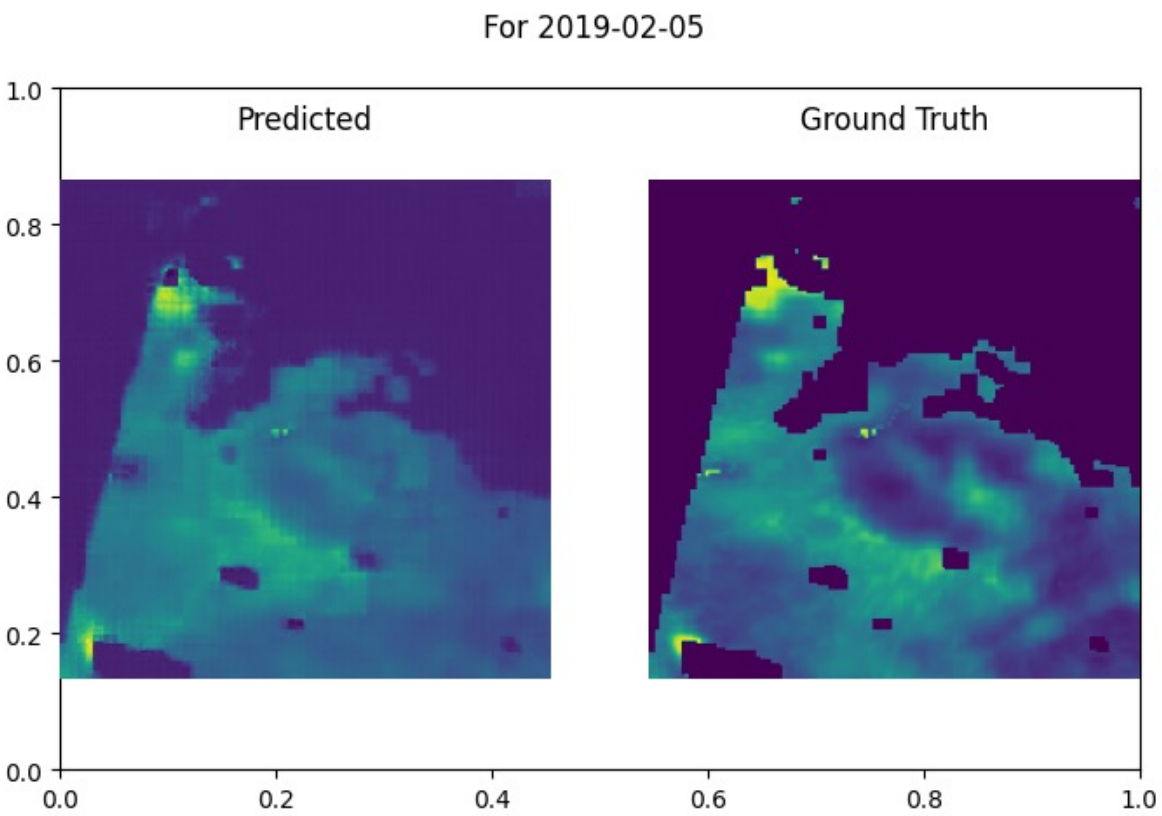
- **Implementation Detail:**
  - PyTorch
  - NVIDIA GeoForce RTX 3070, 8 GB

- **Dataset**
  - Train - June 1, 2015 to 2017
  - Validation – 2018
  - Test – 2019

- Epochs – 100
- Batch size – 4
- Optimizer – Adam
- Weight Initialization – Kaiming uniform
- Loss function – Mean Squared Error (MSE)
- Training time – 40 minutes
- Train loss – 0.00239, validation loss = 0.00189

Results


For 2019-01-31
Predicted
Ground Truth
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


For 2019-02-05
Predicted
Ground Truth
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Systolic CNN AcceLErator Simulator (SCALE-Sim)

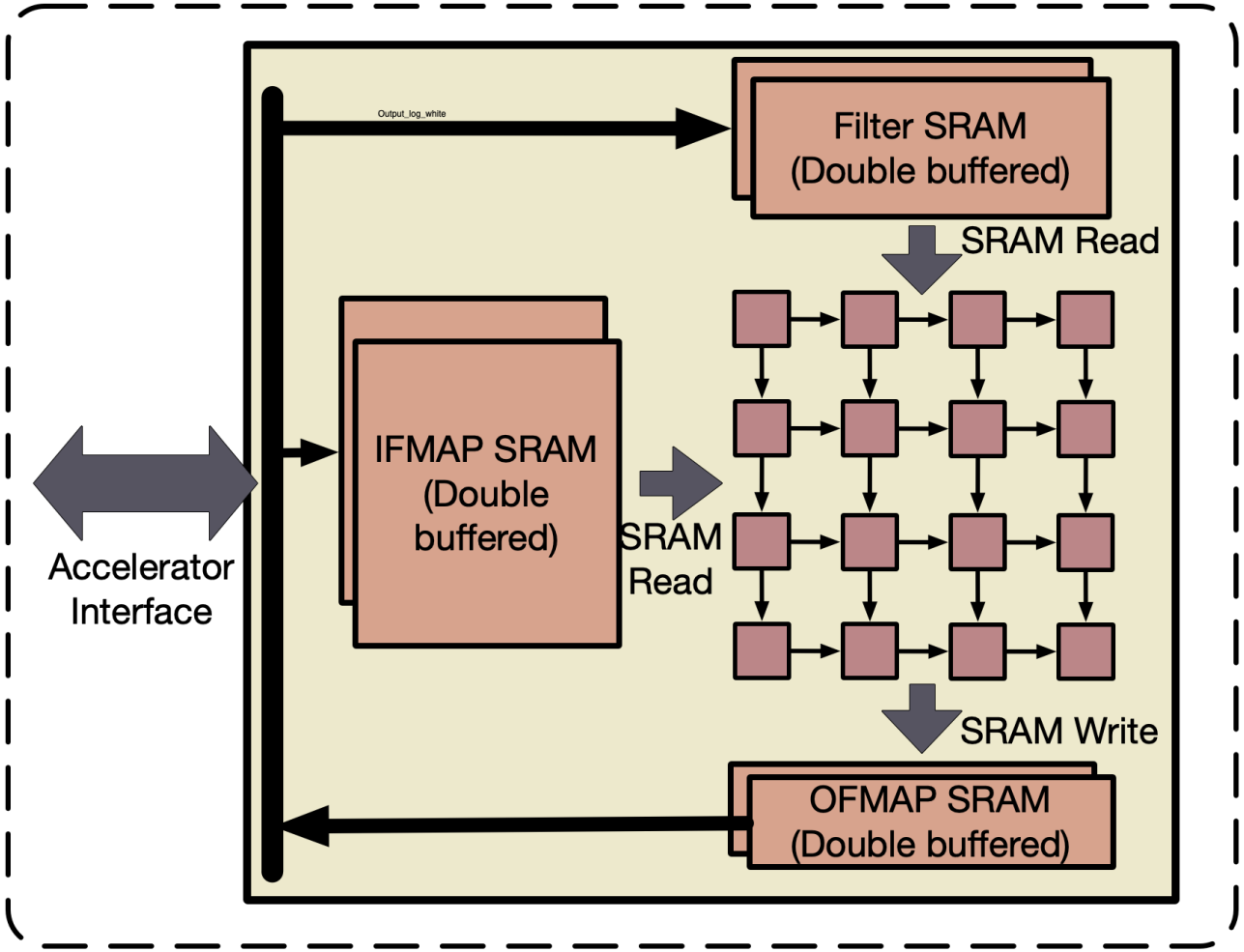
- Simulator for systolic array-based accelerators
- Provides cycle accurate timing, memory bandwidth and traces
- Perform analysis of design space of systolic array architectures
- **Note:** Optimized for CNN layers; attention mechanisms and specific transformer operations not supported

SCALE-Sim Architecture

Parameter	Value
Array Height	32
Array Width	32
IFMAP SRAM Size	1024
Filter SRAM Size	1024
OFRAM SRAM Size	128
Dataflow	WS

Config file with architecture specs

Topology.csv



SCALE SIM

Cycle and IF BW

Traces
SRAM Read/Write
DRAM Read/Write

Input Files

```
[general]
run_name = scalesim_default

[architecture_presets]
ArrayHeight: 32
ArrayWidth: 32
IfmapSramSzkB: 64
FilterSramSzkB: 64
OfmapSramSzkB: 64
IfmapOffset: 0
FilterOffset: 10000000
OfmapOffset: 20000000
Bandwidth: 10
Dataflow: ws
MemoryBanks: 1

[run_presets]
InterfaceBandwidth: USER
```

Config File

```
Layer name,IFMAP Height,IFMAP Width,Filter Height,Filter Width,Channels,Num Filter,Strides,
PatchEmbedding,224,224,4,4,30,32,4,
SwinTransformerBlock1_Attn1,56,56,1,1,32,96,1,
SwinTransformerBlock1_Attn2,56,56,1,1,32,32,1,
SwinTransformerBlock1_MLP1,56,56,1,1,32,128,1,
SwinTransformerBlock1_MLP2,56,56,1,1,128,32,1,
SwinTransformerBlock2_Attn1,56,56,1,1,32,96,1,
SwinTransformerBlock2_Attn2,56,56,1,1,32,32,1,
SwinTransformerBlock2_MLP1,56,56,1,1,32,128,1,
SwinTransformerBlock2_MLP2,56,56,1,1,128,32,1,
SwinTransformerBlock3_Attn1,28,28,1,1,64,192,1,
SwinTransformerBlock3_Attn2,28,28,1,1,64,64,1,
SwinTransformerBlock3_MLP1,28,28,1,1,64,256,1,
SwinTransformerBlock3_MLP2,28,28,1,1,256,64,1,
SwinTransformerBlock4_Attn1,28,28,1,1,64,192,1,
SwinTransformerBlock4_Attn2,28,28,1,1,64,64,1,
SwinTransformerBlock4_MLP1,28,28,1,1,64,256,1,
SwinTransformerBlock4_MLP2,28,28,1,1,256,64,1,
```

Topology File

Terminal Output

```
====================================================
****************** SCALE SIM **********************
====================================================
Array Size:     32x32
SRAM IFMAP (kB):        64
SRAM Filter (kB):       64
SRAM OFMAP (kB):        64
Dataflow:       Weight Stationary
CSV file path:  topology.csv
Bandwidth:      10
Working in USE USER BANDWIDTH mode.
====================================================

Running Layer 0
100%|                                                  | 48450/48450 [00:16<00:00, 2902.06it/s]
Compute cycles: 148469
Stall cycles: 100020
Overall utilization: 31.68%
Mapping efficiency: 100.00%
Average IFMAP DRAM BW: 9.997 words/cycle
Average Filter DRAM BW: 10.000 words/cycle
Average OFMAP DRAM BW: 10.155 words/cycle
Saving traces: Done!
```

Future Implementations

- Drought monitor classification
- Efficient data preprocessing and augmentation
- Incorporating multiple meteorological indicators
- Experiment with different TPU architectures

Thank you.