# Baby Sign Language

Alex To | Aswin Thiruvengadam | Dan Ortiz | Jeffrey Laughman

# Why Baby Sign Language?

Baby sign language is comprised of a series of gestures (closely related to ASL) to improve communication between babies and caregivers prior to vocal vocabulary develops. The goal is to reduce crying and tantrums by providing a medium of communication and reducing caregiver guessing.

One of the members of our team is expecting, we wanted to theme our project around babies in celebration of his new family member!

Project objective

# Produce a deep learning model which accurately recognizes baby sign language gestures on the edge.

# Problems to solve

1. **Model Selection:** Gestures have temporal dependencies and require video classifiers

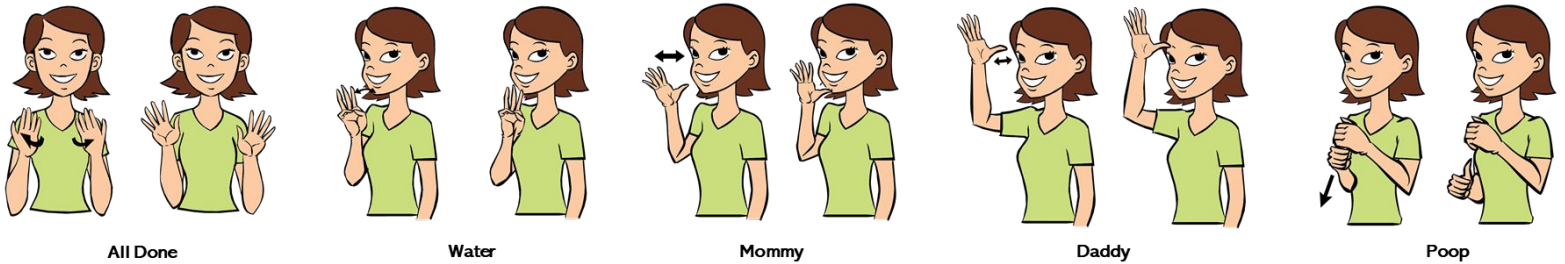
2. **Data Set:** Find and generate appropriate quantity of training and test data

3. **Edge Inference:** Video inference with limitations of Jetson Nano 4G

4. **Inference Ingestion:** Converting streaming input to model requirements

# Data Selection

Source: babysignlanguage.com




Number of Gestures

**5**

Gestures used from proof of concept

Median Training Videos per Gesture
(Source: Youtube and Home Made)

**45**

Transfer learning on model pre-trained on Kinetics-400 dataset

Resized Video

**224 x 224**

as required by Swin-T

# Model Selection

## Image Based

Considered Models:

- CNN (AlexNet/GoogleLeNet)
- Vision  Transformer
- Shifted Window Transformer

Area of Concerns:

- May miss gestures due to still images

## Video Based

Multiscale Vision Transformer

- Area of Concern: Computation and memory requirements

Video Shifted Window

- Selected Model

# **Cloud Training** Considerations

**Swin-T model**

Unwrapping the model code base

Container deployment

Wandb integration

**Data source**

Youtube data extraction

Data generation

Formatting to fit model

**Training time**

Single GPU: Tesla T4

Wall relative time: ~30mins

# **Edge Inference** Considerations

**Model Size**

Memory limitation on Nano 4G

GPU Computation Limitation

Dockerized environment

**GStreamer**

Within container functionality

Pipeline with preprocessing

DeepStream integration

**Integration**

Enabling training on cloud

Jetson inference

# Model Results

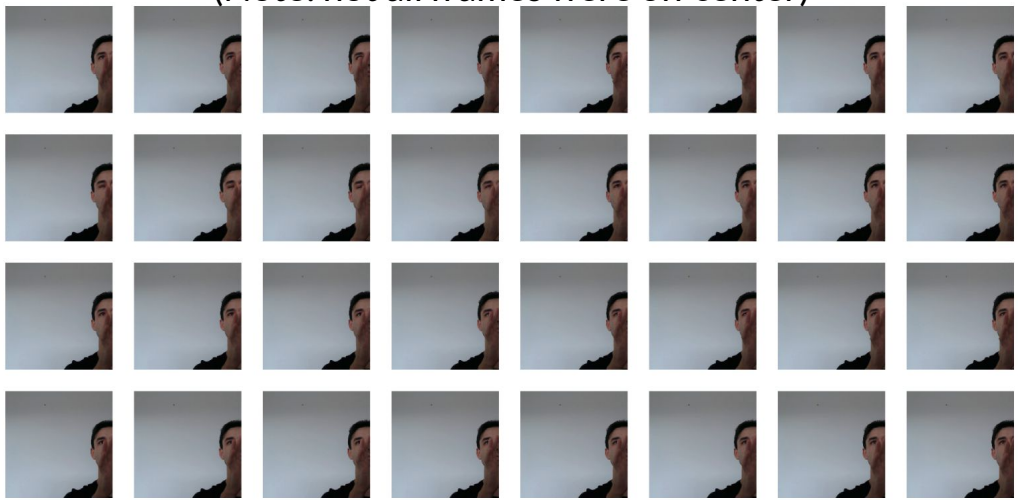
Top 1 accuracy scores

| Training | Validation | Test |
| --- | --- | --- |
| 0.94 | 0.94 | 0.90 |

on an independent set of videos

# Accuracy on test set

Sample frames from mislabeled clip
(Note: not all frames were off center)



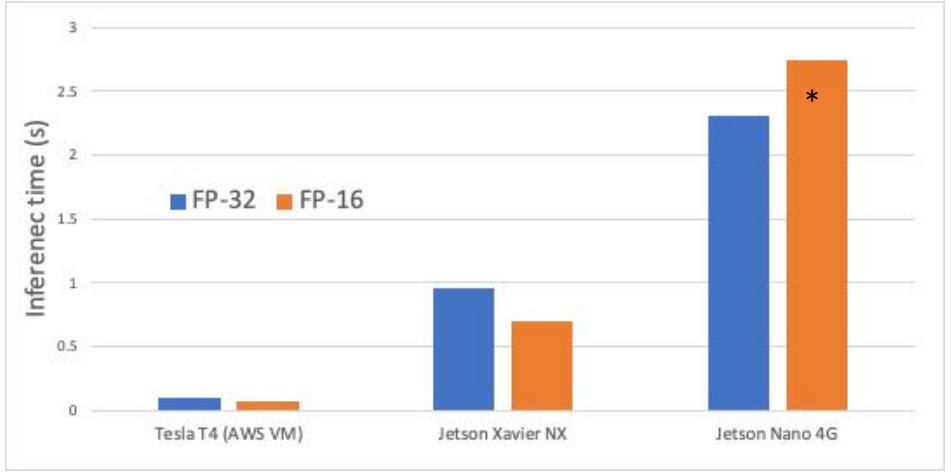
Top 1 accuracy: 90%

| | | Predicted label | | | | |
|---|---|---|---|---|---|---|
| | | all done | Water | Poop | Dad | Mom |
| True label | all done | 2 | 0 | 0 | 0 | 0 |
| | Water | 0 | 2 | 0 | 0 | 0 |
| | Poop | 0 | 0 | 2 | 0 | 0 |
| | Dad | 0 | 0 | 0 | 2 | 0 |
| | Mom | 0 | 1 | 0 | 0 | 1 |

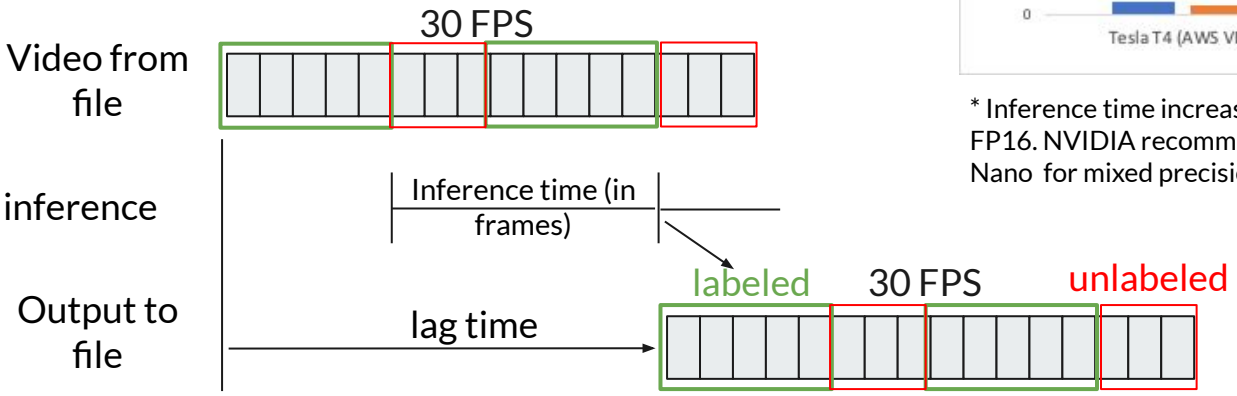
# Results: Frame rate

Model inference on the cloud and on the Jetson with a constant length video

- 1251 total frames
- Extracted at 30 FPS




* Inference time increased on the Jetson Nano when using FP16. NVIDIA recommendation is to use TensorRT on Jetson Nano for mixed precision



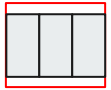
# Inference on video file

VM - FP16

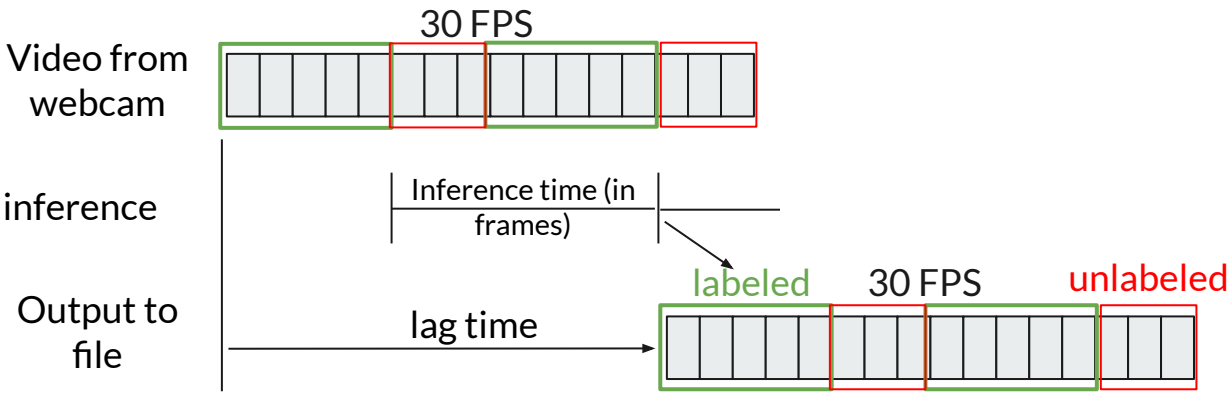

Xavier NX - FP16


Nano 4G - FP32


Unsure ➡ Low probability of inference hence actual label has been removed

! ➡ Frames that were missed during inference = 

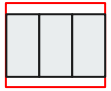
# Inference on webcam




Xavier NX - FP16


Q: Why does labeling look slightly predictive in video?
A: Due to "lag time," we can choose to either view video in lag (as done here, at the action)) or view label in lag

Unsure ➡ Low probability of inference hence actual label has been removed

! ➡ Frames that were missed during inference = 

# Potential Future work

### Further Optimization

Person signing ROI identification,
Hardware acceleration/streamlining,
GPU memory mapping for lower-memory edge devices (Ex: Baby Monitors)


### Explore AI For Good Foundation

Have received interest in further research and application for sponsored project

### Special Needs

Expand learning to other visual-based communication use cases (ex: Down Syndrome)

AI for Good
FOUNDATION

# Resources

**Original Source Material:**

- Paper: https://arxiv.org/pdf/2106.13230.pdf
- Repo: https://github.com/SwinTransformer/Video-Swin-Transformer

**Project Material:**

- Paper: https://docs.google.com/document/d/1kthw4s7_fLTN83xDKdST6W69WBIf2bkFEpByKBc-2sI/edit
- Repo: https://github.com/atox120/w251_fp

# Thank you.

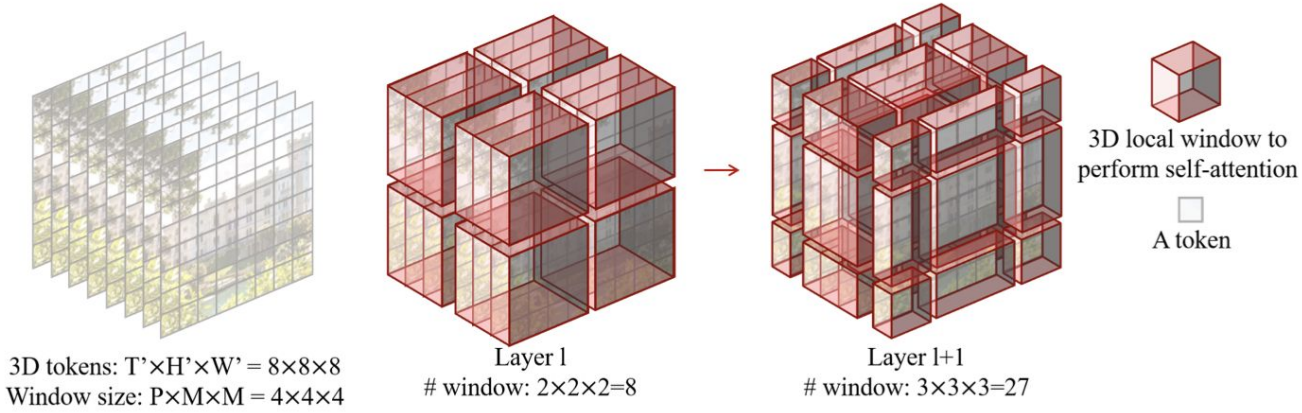
At this time we would like to open the floor for questions.

# About Video Swin transformer

Microsoft Research Asia

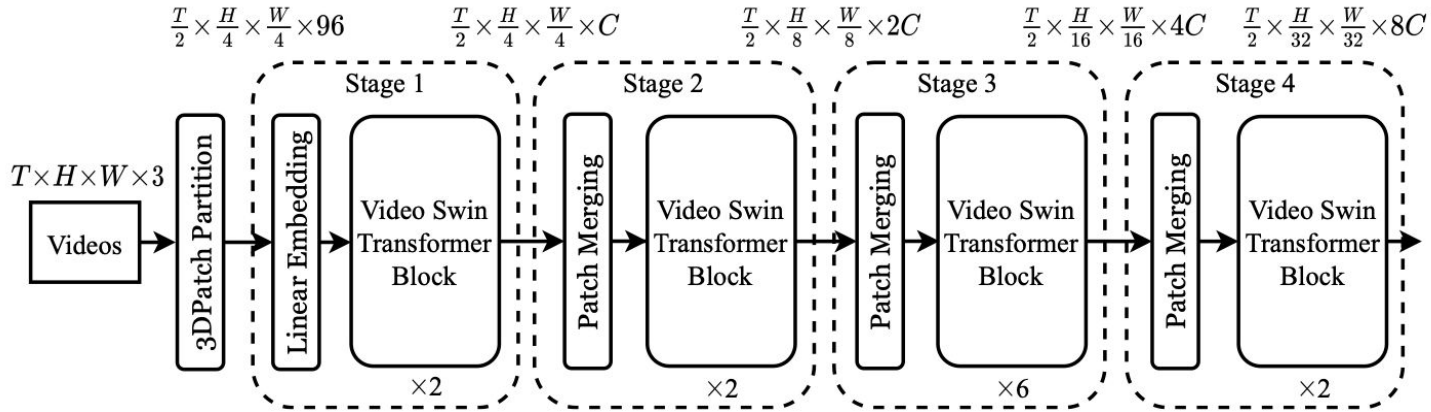
# Video Swin
## Shifted Widow

VSWIN starts by generating  non-overlapping 3D tokens (Time X Height X Width) from 32 frames.  Tokens are combined into non-overlapping windows. As the model progresses through each layer/stage the window is shifted which “introduces connections between neighboring non-overlapping windows in the previous layer and is found to be effective in image classification…”



# **Video Swin** Architecture

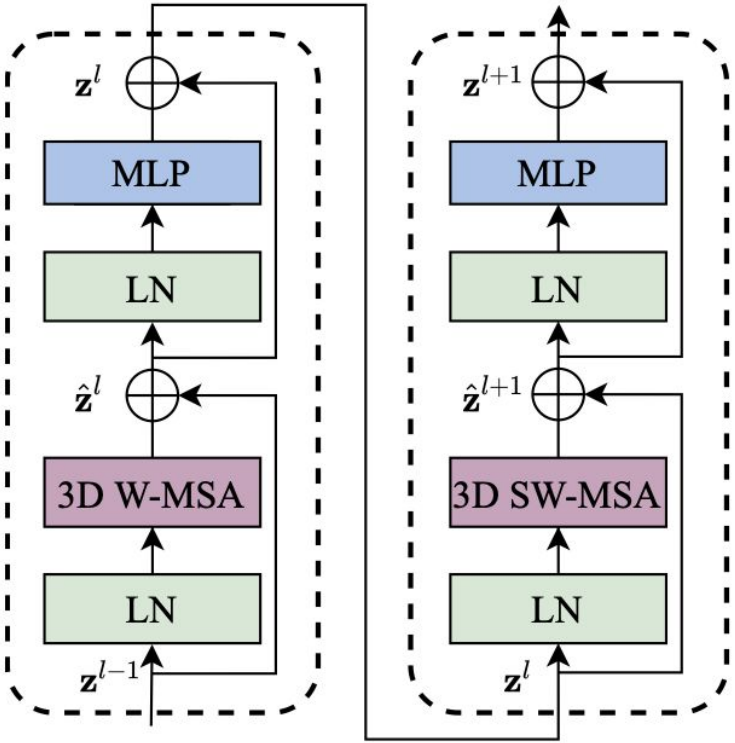
Below is the architecture for Video SWIN-T (tiny version) model. As the frame sets progress through the model……



# Video Swin
## Multi Self-Attention Head

Global attention is infeasible in video processing due to the quadratic scaling with respect to tokens. To resolve this, the VSWIN authors implement the Multi Self Attention Head with  3d Relative Position Bias.

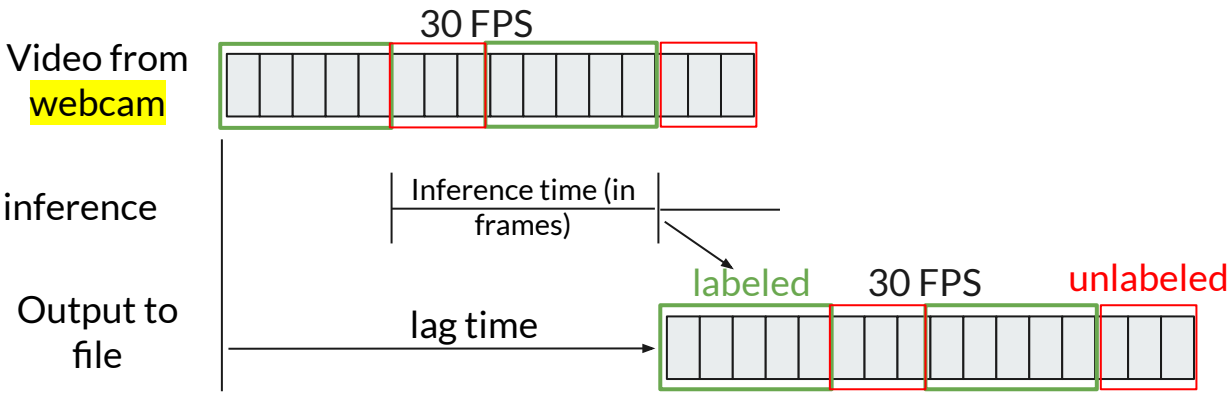
The position information is used as an input into the MSA



# Problems to solve

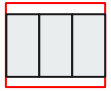
1. **Model Selection:** BSL signs include gestures resulting in a temporal element.
2. **Edge Inference:** Run video inference on the edge with the limitation of the 4 Gig Jetson Nano.
3. **Data Set:** Find and generate appropriate quantity of training and test data.
4. **Inference Ingestion:** How to connect the model to the streaming input.

# Inference on webcam



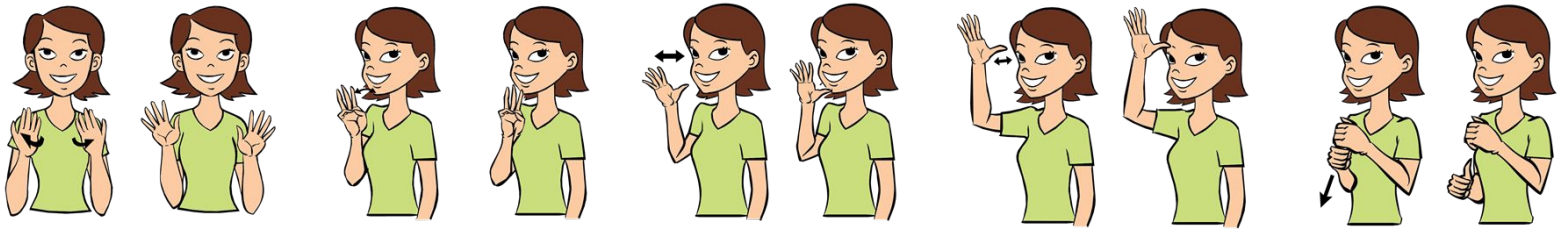

Xavier NX - FP16


Unsure ➡ Low probability of inference hence actual label has been removed

! ➡ Frames that were missed during inference = 

# Data Selection

Source: babysignlanguage.com



All Done | Water | Mommy | Daddy | Poop

Model pre-trained on Kinetic 400

BSL data sources:

1. Youtube
2. Generated by team

| | Water | Mom | Dad | Poop | All done |
|---|---|---|---|---|---|
| **Train** | 37 | 50 | 45 | 42 | 46 |
| **Val** | 9 | 12 | 11 | 10 | 11 |
| **Test*** | 2 | 2 | 2 | 2 | 2 |

*on video clips of person not seen during training or validation

Resized Video

**224 x 224**

as required by Swin-T