DS598 - Deep Learning
Prof. Thomas Gardos

BOSTON UNIVERSITY

Video ASL Recognition

April 30, 2024

Jasmine (Anh) Pham
Farid Karimli

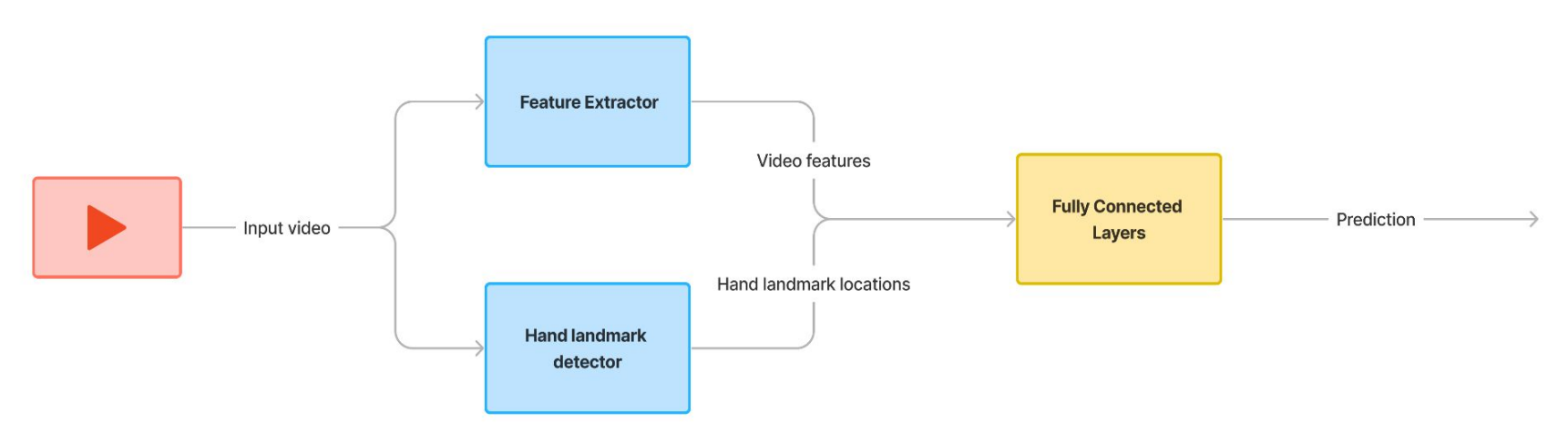
I. Research Background & Motivation

- American sign language (**ASL**): vital mode of **communication** for the **hearing impaired**

- Isolated sign language recognition (ISLR) from videos is **challenging**
  - **Dynamic** nature of signs
  - **Variability** among signers
  - Background clutter

BOSTON UNIVERSITY

II. Proposed Approach



Initial focus - determine **best pre-trained** model for **feature extraction**

III. Dataset

Microsoft ASL Citizen

	Train	Val	Test
# Signers	35	6	11
# Videos	40,154	10,304	32,941
Signer Distribution	60% F, 40% M	83% F, 17% M	55% F, 45% M

BOSTON UNIVERSITY

IV. Models

Feature Extractors:

- **VideoMAE** (Video Masked Autoencoding)
  - Masked auto-encoding - learning video representation from unmasked portions of video
  - Uses ViT with joint space-time attention (ViVit) as backbone

- Video Vision Transformer (ViVit) and Timesformer (Time-Space Transformer)
  - Compute self-attention temporal and spatial components of video

Hand Landmark Detector

- Mediapipe - Google
  - 21 hand landmarks - x, y, z coordinates

V. **Evaluation & Comparison -** classification

	I3D	VideoMAE			
# Signs	2731	40	100	500	1000
# Training	40,154	654	1,611*	15,081	40,154
# Validation	10,304	121	302	3,430	10,304
# Test	32,941	477	1,191	12,025	32,941
Test Accuracy	0.631	**0.832**	**0.853**	**0.88**	TBD
Test Top 5	0.861	**0.966**	**0.970**	**0.963**	TBD
Test Top 10	0.908	**0.989**	**0.972**	**0.980**	TBD

*15-24 videos per sign in the training

Challenges

- Video downsampling
- Video Vision Transformer is terrible on its own
- Resource Memory

Ongoing and further work…

- Training on all signs
- Exploring incremental learning
- Supplement video features with hand landmark locations

Thank you for watching!