LREC 2022 Marseille

7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual

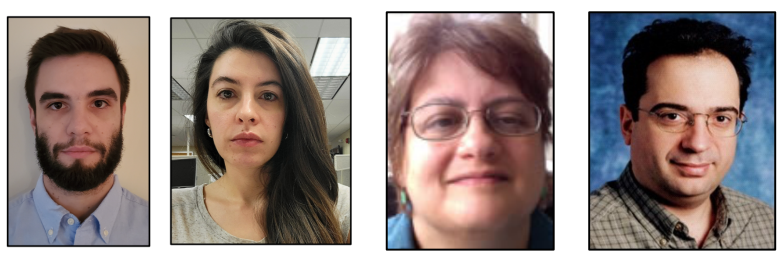
Isolated Sign Recognition using ASL Datasets with Consistent Text-based Gloss Labeling and Curriculum Learning

Rutgers, The State University of New Jersey; cbim; Boston University Linguistics

Konstantinos M. Dafnis*[1], Evgenia Chroni *[1], Carol Neidle[2], Dimitris N. Metaxas[1]

[1]Rutgers University; [2]Boston University

ks703@cs.rutgers.edu, etc44@cs.rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu



Datasets

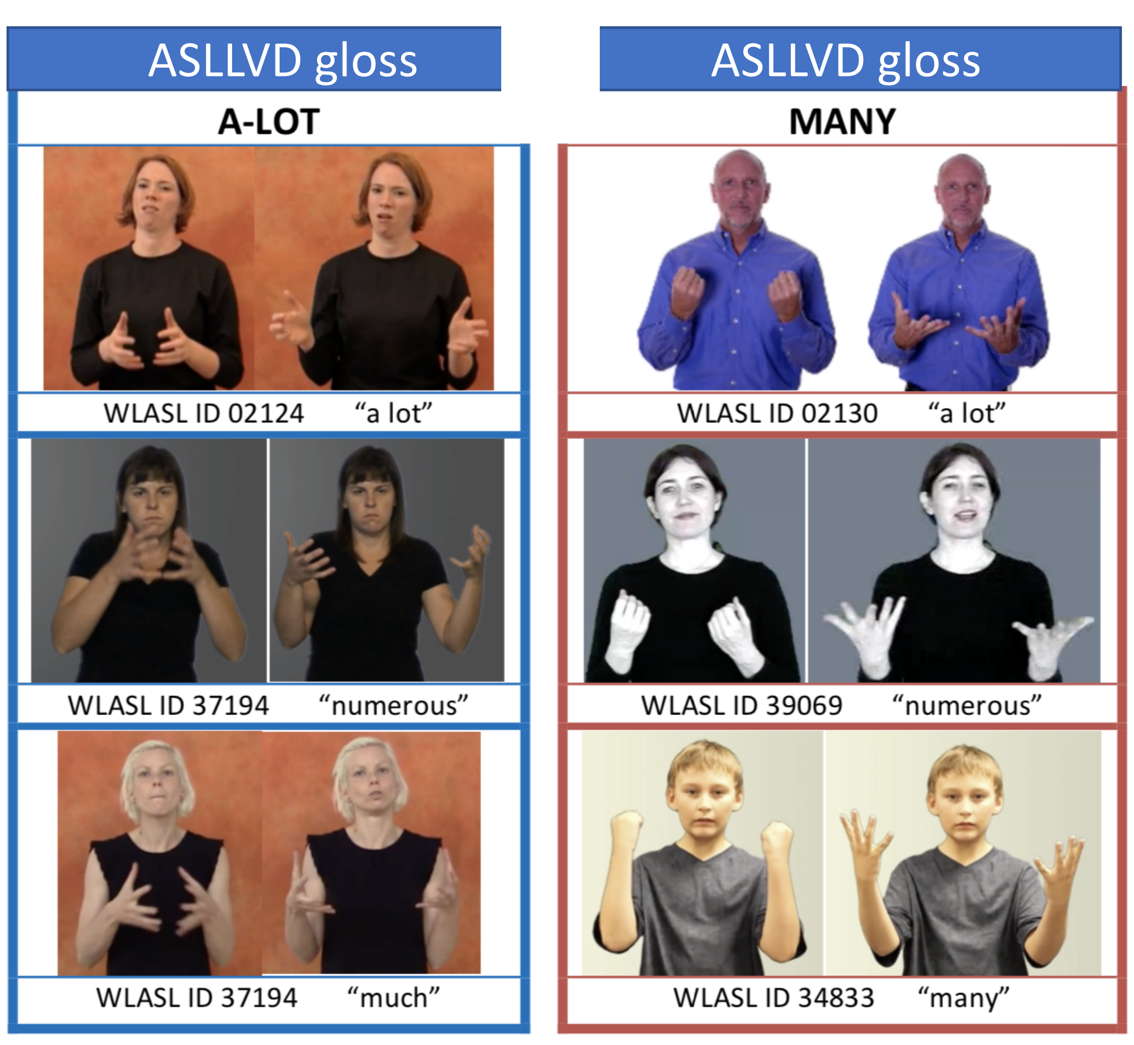
1. **ASLLVD** https://dai.cs.rutgers.edu/dai/s/signbank

Citation-form signs, available for download

- Boston University American Sign Language Lexicon Video Dataset (ASLLVD): **9,748** sign tokens; **6** signers

2. **WLASL** https://dai.cs.rutgers.edu/dai/s/wlasl

- Valuable collection of videos: Many publicly shared video collections, with > 100 signers
- Widely used for sign recognition research
- **BUT** serious problem for machine learning: no 1-1 correspondence between text-based gloss labels and signs



Examples of pervasive gloss inconsistencies in the WLASL

We modified the gloss labels for the WLASL videos so that they are consistent with those for the ASLLVD:

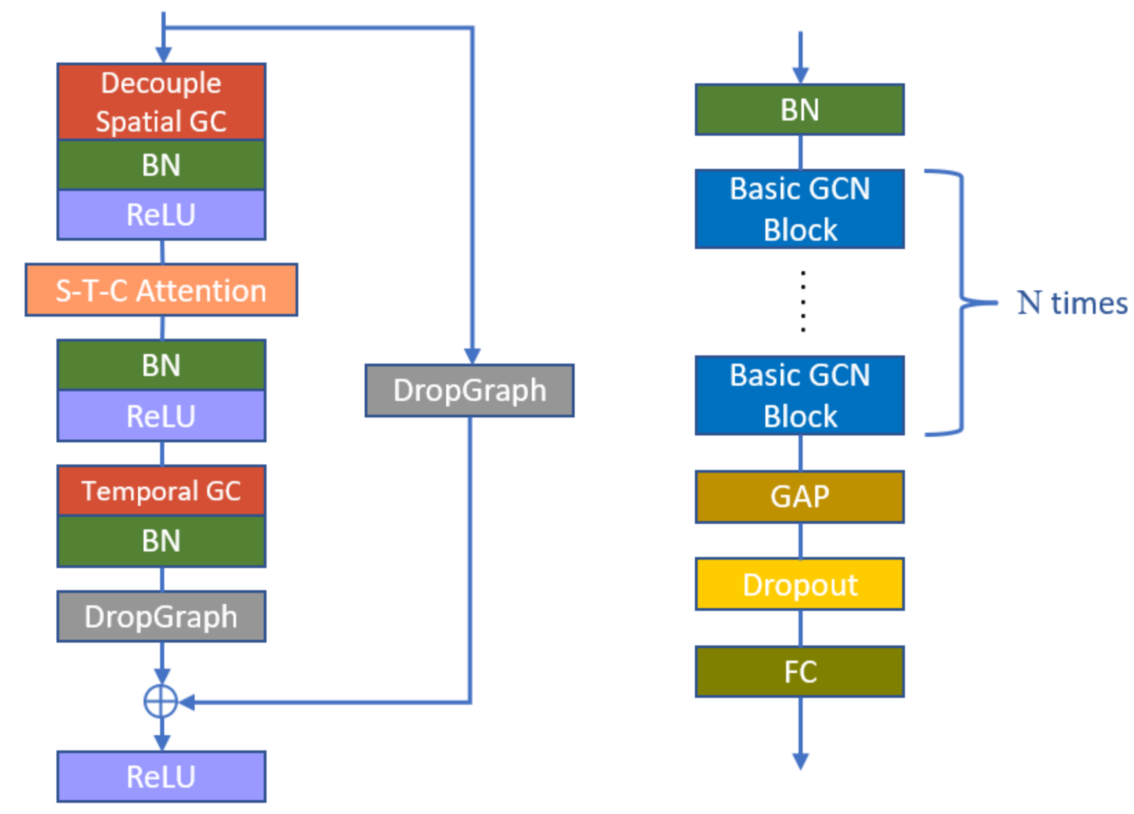
1. Ensuring internally consistent gloss labels for the WLASL &
2. Making it possible to merge these datasets. **We used both!**

We tested sign recognition on the set of signs for which we had a minimum of **6** or **12** examples per sign; see below.

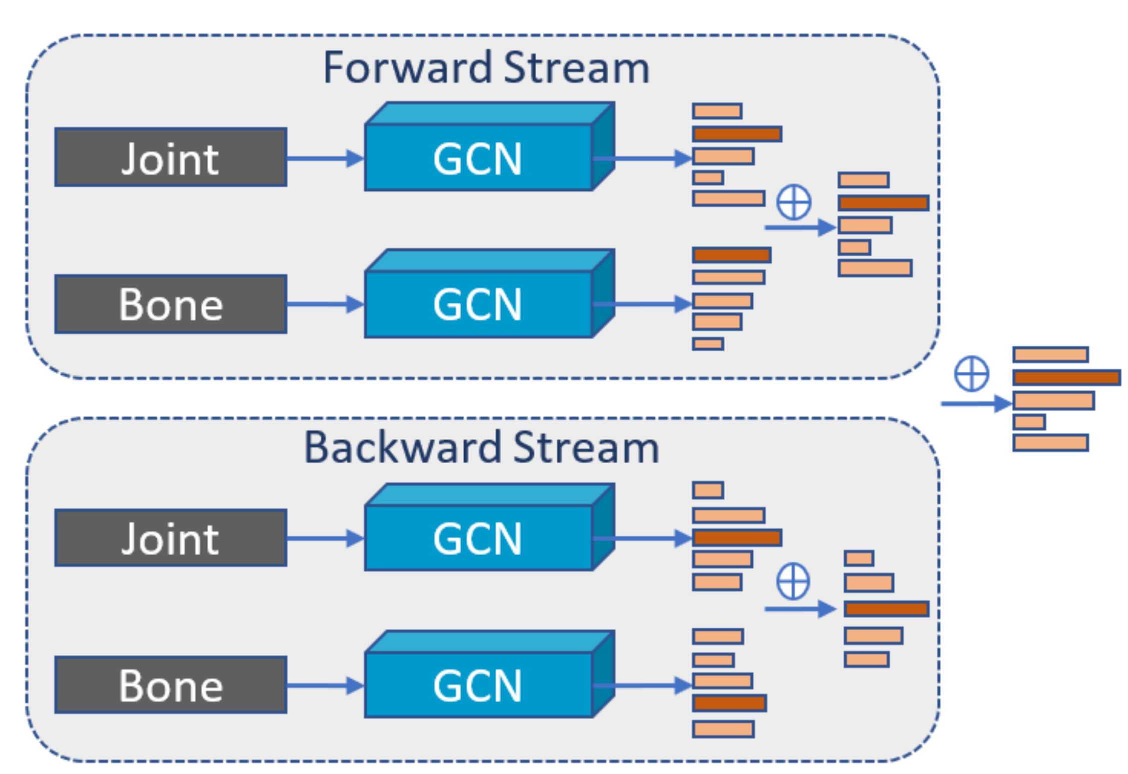
Definition: **Curriculum learning is a "strategy that trains a machine learning model from easier data to harder data, which imitates the meaningful learning order in human curricula."**
However, deciding which samples to categorize as easy or hard is not trivial. For this reason, we use a type of curriculum learning introduced in Saxena et al. (2019), which dynamically estimates, during training, the order of difficulty of each input video for sign recognition by using a new family of trainable parameters for deep neural networks, called data parameters.

Our Approach

- We use a Skeleton-based Spatial-Temporal Graph Convolutional Network (GCN).
- Compared with the sequence-based methods and image-based methods, graph-based methods are more intuitive, since the human body is naturally organized as a graph rather than a sequence or an image.
- Both the forward and backward directions of the video data are used for isolated sign recognition. The forward and backward scores are fused using weighted summation to obtain the final prediction.
- In each direction, we use two types of data streams as input: the human skeleton keypoint (joint) coordinates, and the bone vector (distance between keypoints).



(a) Basic GCN Block (b) Network Architecture



(c) Multi-stream Architecture

Illustration of the GCN pipeline: **(a)** The basic GCN block architecture. **(b)** The GCN architecture. There are 10 basic GCN blocks in all. GAP represents the global average pooling layer, and FC the fully connected layer. **(c)** The overall architecture of the Multi-stream GCN.

Curriculum Learning (CL)

Method Description

$\{(x^i, y^i)\}_{n=1}^{N}$, x^i is a data sample, y^i is the label of x^i

Instance parameter $\phi_i^{instance}$

Class parameter $\phi_{y_i}^{class}$

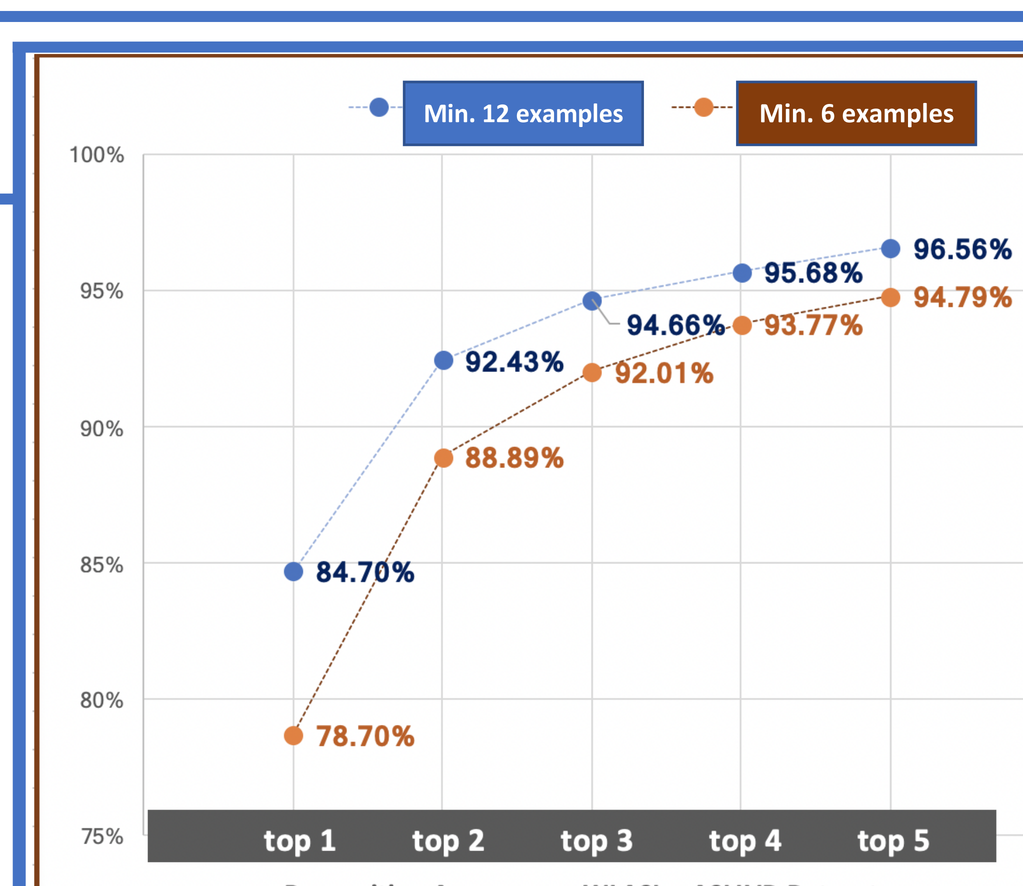
+ Data parameter $\phi_i^* = \phi_{y_i}^{class} + \phi_i^{instance}$

Both instance parameters and class parameters are learnable, so we use two optimizers: one for each.

The modified cross entropy loss becomes: $L^i = -log(p_{y^i}^i)$, $p_{y^i}^i = \frac{exp(z_{y^i}^i/\phi_i^*)}{\Sigma_j exp(z_j^i/\phi_i^*)}$.
where z^i are the logits.

Isolated Sign Recognition

Min. # examples per sign	6				12			
Total # class labels	1,502				990			
Total # examples	23,016				18,482			
Streams	Forward	Forward	Backward	Backward	Forward	Forward	Backward	Backward
	Top-1	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*	*Top-1*	*Top-5*
Joint	72.96	91.42	74.19	91.14	79.18	94.09	78.24	93.78
Bones	72.63	91.47	72.09	91.09	76.31	93.51	76.49	93.30
Multi-stream	77.58	94.21	77.65	94.24	83.07	95.87	82.26	95.87
Forward Multi-stream w/ CL	77.63	94.34			82.59	96.23		
	Top-1		*Top-5*		*Top-1*		*Top-5*	
Fusion (no CL)	78.70		94.79		84.70		96.56	



*Equal Contribution

This work was partially funded by grants from the National Science Foundation.