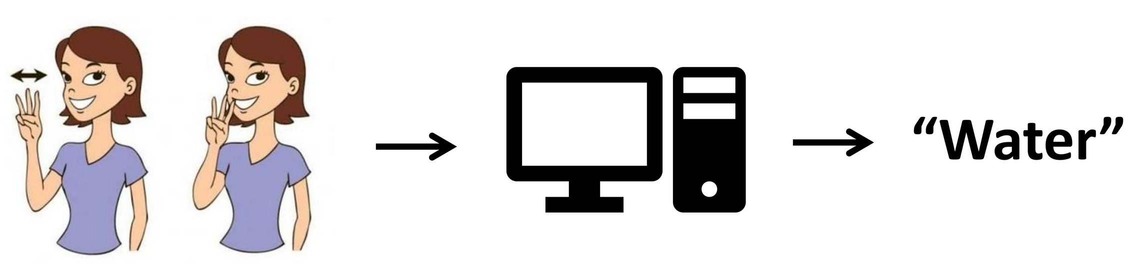
Optimizing Deep Learning Models for Sign Language Recognition and Translation

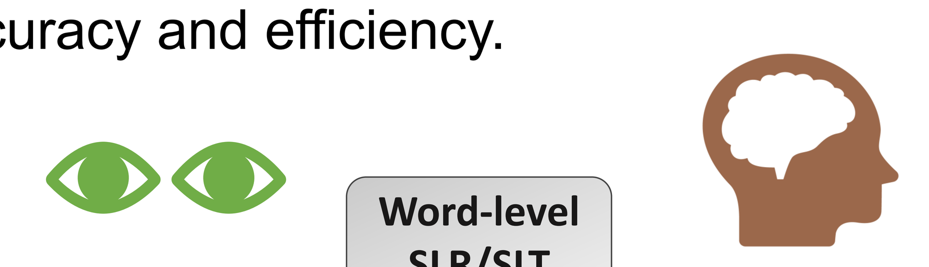
Neeharika Kompella[1], Maryam Dehghani[2,3], Nazanin Alizadeh[2,3], Tom Chau[2,3]
[1]Electrical and Biomedical Engineering, McMaster University, [2]PRISM Lab, Bloorview Research Institute, [3]Institute of Biomedical Engineering, University of Toronto

Background

Sign Language Recognition (SLR) and Translation (SLR) systems can help **bridge the communication gap** between Deaf and Hearing individuals.



Traditional models **struggle to capture hand gesture, facial expression, and body language simultaneously.** **CNN-Transformer models** provide a strong approach to improve accuracy and efficiency.



Objective

To **optimize** a CNN-Transformer model for SLR and SLT via **hyperparameter tuning** and **data processing.**

Methods

Part 1: Effects of Hyperparameter Tuning

Model was trained on the **ASLLVD dataset**; performance of various configurations were evaluated via **testing accuracy, F1 score, and observing the loss curve** over **20 epochs.**

Hyperparameter	Candidate Values
Attention heads	2, 4, 16
Transformer layers	2, 4
Learning rate	1e-5, 8e-5, 1e-4, 2e-4
Optimizer	Adam, AdamW
Data split	60/20/20, 70/15/15

Part 2: ASL Dataset Preprocessing

Two large, publicly available ASL datasets—**WLASL** and **SEM-LEX**—were processed, filtered, and combined for the purposes of training forthcoming models.

Thorough evaluation of hyperparameter values yield an accurate and efficient CNN-Transformer model for Sign Language Recognition and Translation.

Holland Bloorview Kids Rehabilitation Hospital
Bloorview RESEARCH INSTITUTE

Results

Best Overall Configuration

8 heads, 2 layers, Adam, 1e-4 learning rate, 70/15/15 split
Final testing accuracy: **73.6%**
Final F1 score: **0.66**

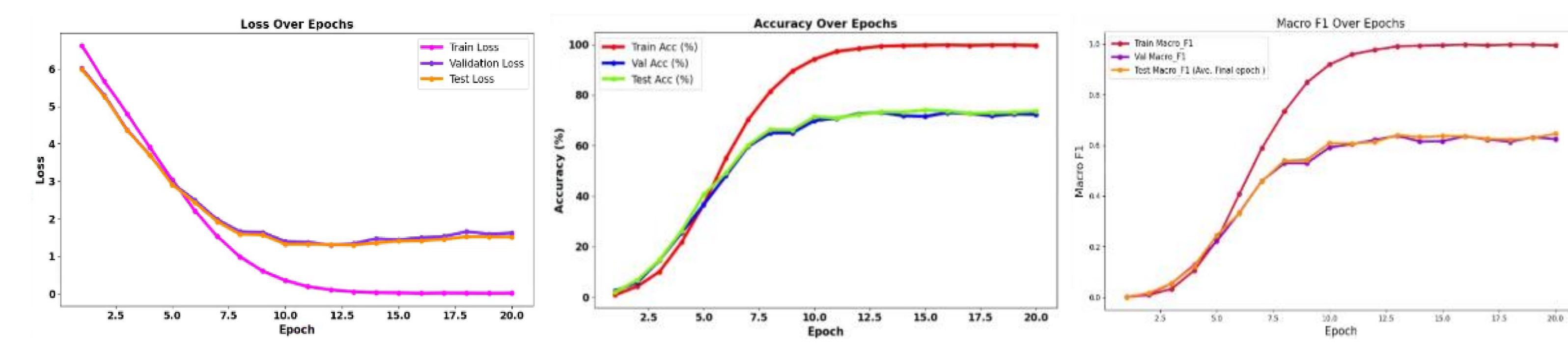


Fig. 1: Loss, accuracy and F1 score plots of 8/2/1e-4 configuration

Greater network depth resulted in **overfitting** and **poorer model performance.**

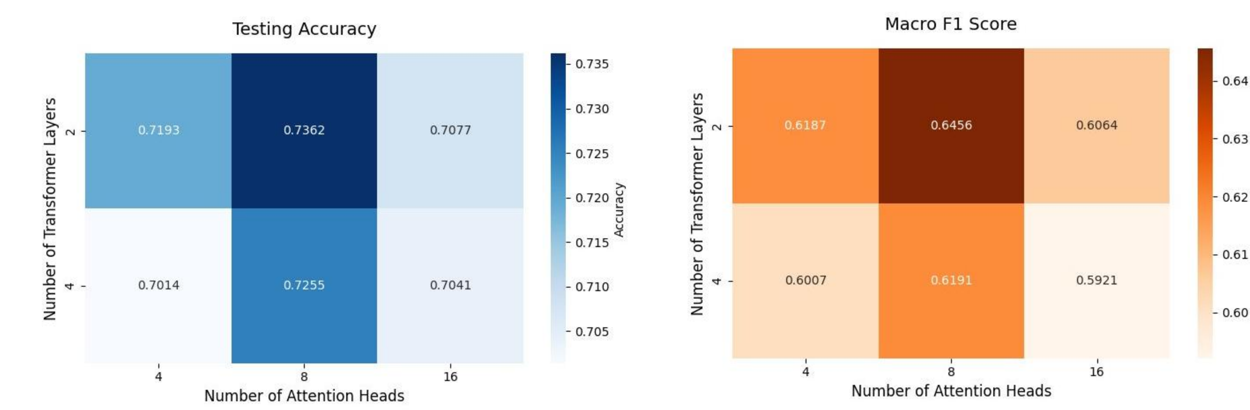


Fig. 2: Accuracy and F1 score for all head/layer configurations.

Optimal learning rate for all configurations was 1x10^{-4}.

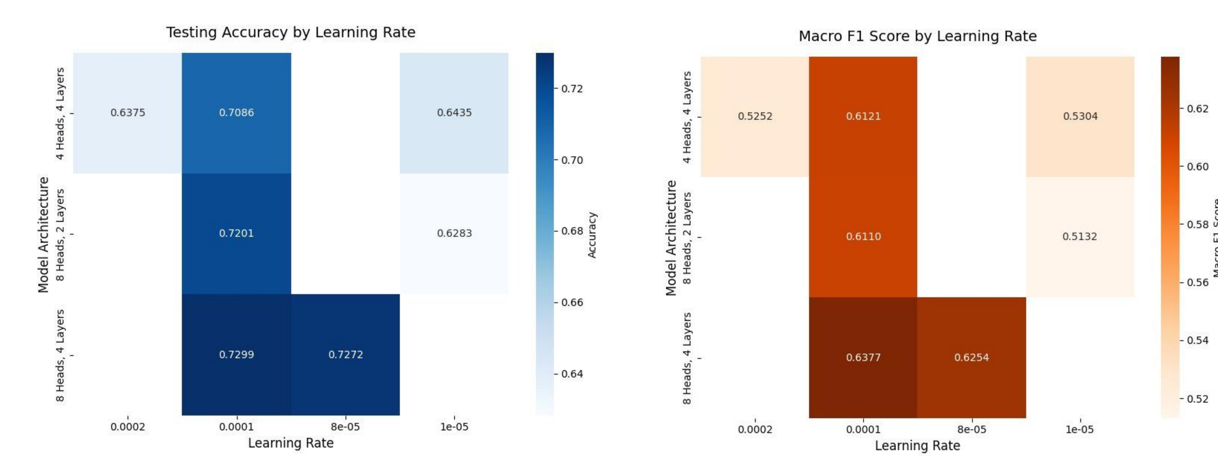


Fig. 2: Accuracy and F1 score for head/layer configurations vs. learning rate

Weight decay with alternative optimizer **AdamW** **minimally improved overfitting for deeper networks but hurt lightweight configurations.**

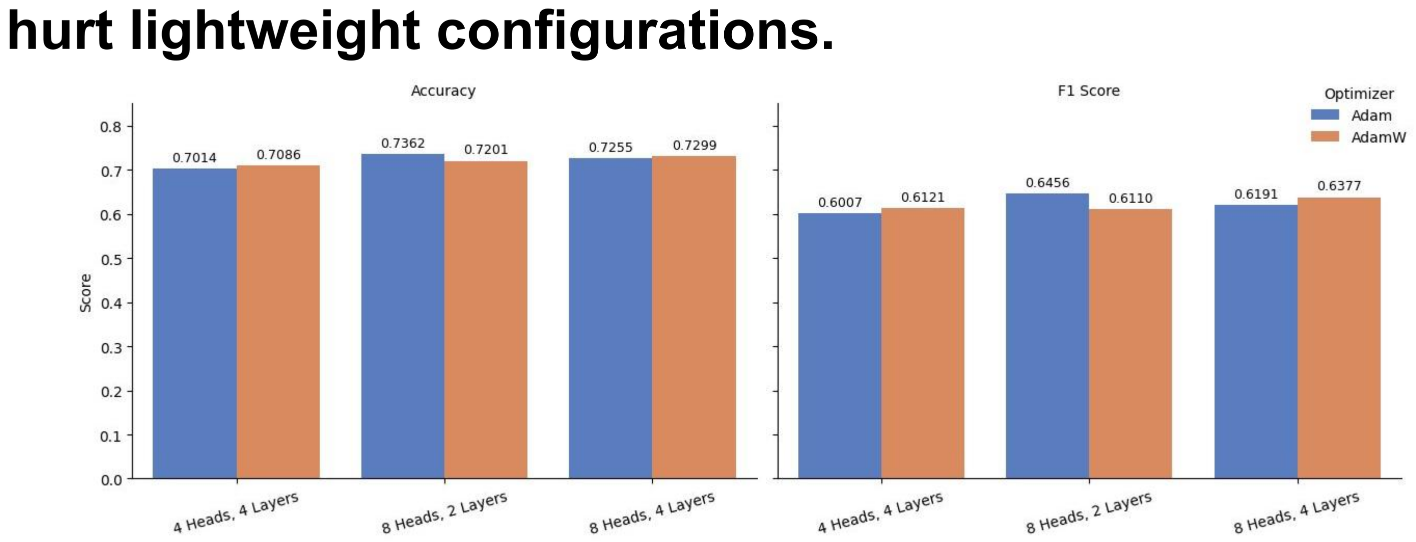


Fig. 3: Performance of various configurations before and after use of AdamW

Conclusion

Lightweight model is most stable and accurate with current dataset; performance is **bottlenecked by dataset size.**

Future work should train on **larger datasets** (e.g., WLASL+SEM-LEX), **filter core vocab** for young signers, and develop a system for **real-time deployment.**

Relevance to Holland Bloorview

SLR/SLT can aid in **early sign language acquisition** and significantly **reduce communication barriers** between young patients, their families, and non-signing clinicians, ultimately fostering a **more inclusive and effective** rehabilitative environment.