Toward an Expressive Semantic Map for Sign Language Processing

Richard Lim, Class of 2027

In the last decade, neural machine translation systems have revolutionized how computers understand language by representing words as vectors that capture their meanings and relationships. Simultaneously, breakthroughs in signal processing and natural language processing (NLP) have expanded these methods beyond text to include information-rich media like music and speech. This project explores whether similar techniques can be applied to American Sign Language (ASL) to produce general-purpose vector representations of signing which also capture the expressive richness of ASL.

Much of the **existing literature** in sign language NLP (SL-NLP) has focused on either sign recognition or translation between text and transcriptions of ASL [3], [4], [5]. In creating vector embeddings, SignBERT [4] adapted Google's BERT [7] architecture to encode sequences of ASL signing into semantic vectors, then decode them to predict words. However, SignBERT focuses entirely on hand keypoints, neglecting important aspects of ASL signing like facial expressions and the relative position of the hand and body. This makes it unsuitable for applications of ASL beyond simple translation.

Therefore, the **goal** of our present research is to assess the suitability of different architectures that produce general-purpose vector embeddings that (1) capture semantic and expressive depth, and (2) account for the role of the body and face in ASL signing. To achieve this, we sought to find or develop an architecture that improved upon SignBERT, adapt it to extract vector embeddings, and then train it on the ASL dataset.

The **architecture** we decided upon was that of the spatial temporal graph convolutional network (ST-GCN) [6]. Originally developed for action recognition, the ST-GCN takes as input a graph representing a video of human actions. As Figure 1 illustrates, nodes function as joints across different time points. Edges, conversely, represent limbs in the spatial dimension and connect adjacent time points for the same joint in the temporal dimension. The input for each node is interpreted by taking a weighted sum of its neighbors' inputs. In our case, this can be understood as having the effect of each joint be determined by the effect of neighboring joints and time points immediately before and after. By allowing us to visualize the importance of specific joints for predicting each sign, this process improves interpretability over SignBERT. To adapt the architecture to our data, we also expanded the graph to include representations of face and hand keypoints.

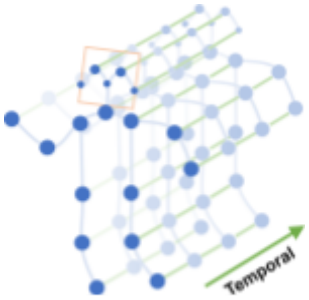

Figure 1: A spatial temporal graph representation of a video

The **data** from this project originates from How2Sign [1], a multimodal ASL dataset in multiple parallel forms, most crucially full-sentence text and skeletal keypoint locations extracted using OpenPose [2].

Training the model involved 3 key steps. Firstly, we pre-computed word embeddings using an existing language model. We picked the BERT-based sentence encoder *all-MiniLM-L6-v2* for its effectiveness in encoding full sentences. Secondly, we extracted a vector embedding from an inner state of the model. Finally, we compared the output to these pre-computed embeddings, using their similarity score to train the model.

Our **preliminary results** demonstrate that more training data improves the model significantly when attempting to predict previously unseen data, demonstrating that the model generalizes and is not merely "memorizing" the dataset.

Next steps for our research include training the model on the entire dataset and qualitatively evaluating the produced embeddings. Do different instances of the same phrase yield similar embeddings, indicating robustness to style and signer differences? Do emotionally marked signs produce systematic "offsets" in embedding space relative to neutral signs, analogous to how semantic offsets (e.g., king – man + woman ≈ queen) emerge in word embeddings? I will pursue these questions over the course of my independent study which extends my summer work. We hope to uncover the semantic and expressive richness of these vectors that make them suitable for applications in sign language processing, translation, and music interpretation.

Faculty Mentor: Jeová Farias
Funded by the Bowdoin Research Fellowship

References

[1] A. Duarte *et al.*, “How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language,” *DIGITAL.CSIC (Spanish National Research Council (CSIC))*, Jun. 2021, doi: https://doi.org/10.1109/cvpr46437.2021.00276.

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, doi: https://doi.org/10.1109/tpami.2019.2929257.

[3] B. Saunders, N. Cihan Camgoz, and R. Bowden, “Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video,” Nov. 2020.

[4] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, “SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, doi: https://doi.org/10.1109/iccv48922.2021.01090.

[5] A. Moryossef, “sign.mt: Real-Time Multilingual Sign Language Translation Application,” 2024.

[6] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” in *AAAI Conference on Artificial Intelligence*, 2018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.