American Sign Language Translation: An Approach Combining MoViNets and T5

N. Cabral & D. Emery
University of California, Berkeley
School of Information
{cabralnc96, deanna.emery}@berkeley.edu

Abstract

This paper addresses the challenge of developing a machine translation solution for American Sign Language (ASL) to overcome barriers in accessibility for deaf individuals. Recognizing and translating signs requires a comprehensive understanding of hand gestures, body poses, as well as facial features to capture the complete meaning of each sign, presenting a formidable challenge for sign language processing. We present a new architecture for translation of ASL comprised of a fine-tuned MoViNets CNN model and a T5 encoder-decoder model to generate translations from the video embeddings, achieving a BLEU score of 1.98 and an average cosine similarity score of 0.21. Additionally, we found that fine-tuning a pre-trained language model on single words first, and then further fine-tuning on complete captions resulted in superior performance. With more data and training time, this model architecture shows promise to achieve results comparable to state-of-the-art models.

1. Introduction

This paper addresses the challenge of developing a machine translation solution for American Sign Language (ASL). The need for automated ASL translation is evident as deaf individuals, particularly children, face barriers in accessibility and interaction [2]. This issue is exacerbated by the lack of available interpreters, as well as individuals fluent in ASL, resulting in reduced independence, expressiveness, and a gap in inclusive technology [1]. The initiative to create a translation capability plays a critical role in mitigating the adverse effects on communication, comprehension, and participation experienced by individuals with hearing loss. Additionally, it would aid in fostering inclusivity and equal opportunities in education, employment, and daily life [2].

The development of an automated ASL translation capability that is independent of interpreter availability and audience fluency requires a nuanced approach [6]. ASL differs from being a signed form of English in that it is a visual language with a unique multimodal nature [2]. It employs manual features such as finger placement, palm orientation, movement, and location, as well as non-manual markers like head movement, mouth actions, facial expressions, and eye gaze to convey information [15]. Recognizing and translating signs requires a comprehensive understanding of all these features to capture the complete meaning of each sign, presenting a formidable challenge for computer vision in sign language processing [15]. In sentence-level translation, concatenation of word-level translations is insufficient for accurately recognizing, translating, and generating complete sentences. This approach neglects the complex grammatical and linguistic structures inherent in sign language, which differ significantly from spoken language [2]. To address these complexities and nuances, the use of advanced transformer models is essential for accurate translation and caption generation of ASL [15].

Additionally, the task of advancing ASL translation faces challenges rooted in the scarcity of reliable and publicly available ASL datasets that feature both signed words and sentences, as well as their textual translations [16]. This scarcity becomes further pronounced at the sentence level [16]. Most available datasets contain word-level data, which is not a strong representation of real-world use cases such as those that may occur in natural conversations [15]. As the current state of the art in automated ASL translation grapples with this scarcity, there is a critical need for developing more representative datasets to enhance the effectiveness and practicality of any proposed translation models [16].

Currently, the state of the art in automated ASL

translation falls considerably short of solving the problem [2, 6, 16]. In this paper, we take a novel approach to the machine translation of ASL, aiming to develop a TensorFlow model capable of interpreting ASL signs from video data. Utilizing computer vision and natural language processing techniques in combination with diverse video datasets, our proposed model attempts to perform machine translation for sentence-level ASL.

1.1. Background

The task of sign language recognition and translation has been approached through a variety of proposed methods [2, 6, 16]. Li et al. [6] applied a Two-Stream Inflated 3D ConvNets (I3D) classification model to their dataset of 26K word-level videos spanning a vocabulary of 2,000 unique words [6]. Although the model was able to achieve reasonable accuracy, such approaches do not scale well as the vocabulary size grows.

Tarrés et al. [15] approached sentence-level translation building off of prior work to train an I3D model to generate embeddings that could be used in a transformer architecture for translation. They developed a curated dataset of over 2500 sentence-level ASL videos, broken into 45K sub-captions, with a vocabulary of 16K words for of training and testing. Their generated I3D features [2] were passed to a Fairseq model to generate captions for each video [15].

Uthus et al. [16] contributes the most expansive sentence-level ASL video dataset yet with 60K unique words across 11K videos. The methodology in their research was based around using MediaPipe [8] to capture human pose data, and then passing this embedding representation to a pre-trained T5 v1.1 encoder-decoder model. This approach achieved a state-of-the-art BLEU score of 12.39 when applied on the How2Sign dataset after additional fine-tuning [16]. Our approach draws inspiration from the YouTube-ASL paper, in that we will use a pre-trained encoder-decoder for caption generation.

For our approach to sentence-level translation, we attempt to merge the two approaches from Tarrés et al. [15] and Uthus et al. [16]. Similar to Tarrés et al. [15], we opt to use a CNN model instead of MediaPipe, as MediaPipe is generally geared towards detecting body poses and may be less suitable for capturing subtle information such as facial expressions and eye gaze (CITATION). We follow Uthus et al. [16] decision to use a pre-trained language model as it was shown to improve the model performance considerably.

Dataset	Vocab	# Signers	# Hours	# Videos
WLASL	2000	119	14	25,513
MS-ASL	1000	22	25	21,083
YouTube-ASL	60,000	2519	984	11,093

Table 1. **Data Sources**. Our models were trained on three datasets, including WLASL, MS-ASL, and YouTube-ASL. Both WLASL and MS-ASL are word-level datasets, while YouTube-ASL is sentence-level. The YouTube-ASL dataset is the most comprehensive and diverse publicly available data source.

2. Data

We made use of three different data sources, detailed in Table 1. The Microsoft ASL Dataset [4] (MS-ASL) and Word-Level ASL Dataset [6] (WLASL) contain videos of individual words, while the YouTube-ASL dataset [16] contains full sentences or sequences of words.

For our experiment, we worked with videos from the YouTube-ASL dataset, downloaded along with their respective English captions. We downloaded the full WLASL dataset and a subset of the MS-ASL dataset containing only the words already present in the WLASL dataset. The processing consisted of passing the videos through OpenCV and converting the frames into RGB numpy arrays. Each video was then divided into clips corresponding to each caption. The video frames for each caption were then cropped to the center square and resized to 224 x 224, such that the final dimensions for a video of length N would be $N \times 224 \times 224 \times 3$.

The captions corresponding to the YouTube-ASL dataset were also cleaned by converting them to lowercase and removing special characters (keeping punctuation) and empty spacing. We also removed any captions that consisted of only bracketed words (eg. [music], or [demonstrates sign]). While nearly all captions in the dataset occurred uniquely, there were a few slogans that occurred up to 30 times. To avoid biasing the model on these phrases and terms, we sampled up to 5 occurrences for each caption.

3. Methods

Our model architecture consisted of a two-step approach. First, we used a convolutional network model to extract video features in the form of embeddings from our data. Next, we passed the output video embeddings into a large language model for translation.

To generate the video embeddings, we used MoViNets, an action-recognition classification model that makes use of 3D convolutions, illustrated in Fig-

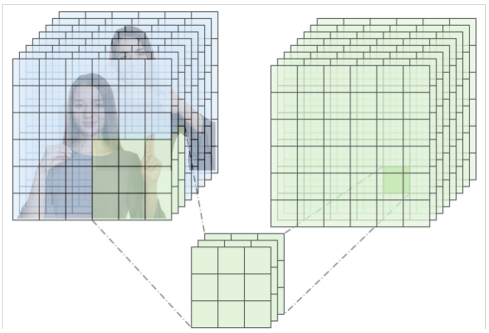

Figure 1. **A 3-Dimensional Convolution** illustrated.

ure 3, to extract not only image features from each video frame, but also temporal features across the video frames [5]. The MoViNets model has been pre-trained on the Kinetics600 dataset, and outperforms many of the state-of-the-art models while also being lighter weight with only 3 million parameters [5].

For the large language model, we used T5, an encoder-decoder model that has been pre-trained on translation tasks and is relatively lightweight compared to other large language models with 222 million parameters [10].

3.1. Action Recognition Classifier

We used the MoViNets A2-base model architecture, shown in Table 2, which consists of six 3D convolutional blocks [5]. Because the video embeddings needed to match the shape of the input embeddings for the T5 model, we had to modify MoViNets's architecture. To do this, we removed the final four layers from the model and replaced them with two 2D convolutions followed by a flattening layer to reduce the dimensionality down to 768. We then designed the model to have two outputs for both the classification and the video embedding. The output video embedding preserves the number of frames, N, in the video such that the final shape is $N \times 768$. The modified MoViNets architecture is shown in Table 3.

With the modified architecture defined, we fine-tuned the MoViNets model as a classifier on the ASL word-level datasets, including both WLASL and MS-ASL. Due to sparse coverage of words and considerably slow training times, we reduced the dataset to only words with at least 27 occurrences, resulting in the 107 most frequent words. Since the model was heavily swayed by imbalanced classes, we dropped data such that no class had more than 50 occurrences, resulting in a total of 2,518 data points in the training set. To further augment the dataset, we re-sampled 2 videos from each class in the training dataset and applied horizontal flip, increasing our training dataset

Stage	Operation	Output size
data	stride 5, RGB	N×224^2
$conv_1$	1×3^2, 16	N×112^2
$block_2$	$\begin{bmatrix} 1\times5^2, 16, 40 \\ 3\times3^2, 16, 40 \\ 3\times3^2, 16, 64 \end{bmatrix}$	N×56^2
$block_3$	$\begin{bmatrix} 3\times3^2, 40, 96 \\ 3\times3^2, 40, 120 \\ 3\times3^2, 40, 96 \\ 3\times3^2, 40, 96 \\ 3\times3^2, 40, 120 \end{bmatrix}$	N×28^2
$block_4$	$\begin{bmatrix} 5\times3^2, 72, 240 \\ 3\times3^2, 72, 160 \\ 3\times3^2, 72, 240 \\ 3\times3^2, 72, 192 \\ 3\times3^2, 72, 240 \end{bmatrix}$	N×14^2
$block_5$	$\begin{bmatrix} 5\times3^2, 72, 240 \\ 3\times3^2, 72, 240 \\ 3\times3^2, 72, 240 \\ 3\times3^2, 72, 240 \\ 1\times5^2, 72, 144 \\ 3\times3^2, 72, 240 \end{bmatrix}$	N×14^2
$block_6$	$\begin{bmatrix} 5\times3^2, 144, 480 \\ 1\times5^2, 144, 384 \\ 1\times5^2, 144, 384 \\ 1\times5^2, 144, 480 \\ 1\times5^2, 144, 480 \\ 3\times3^2, 144, 480 \\ 1\times3^2, 144, 576 \end{bmatrix}$	N×7^2
$conv_7$	1×1^2, 640	$N \times 7^2$
$pool_8$	$N \times 7^2$	1×1^2
$dense_9$	1×1^2, 2048	1×1^2
$dense_{10}$	1×1^2, 600	1×1^2

Table 2. **Movinets A2 Original Architecture**, where N represents the number of input frames.

to 2732 samples. This not only helped to increase the number of samples per class, but also helped to address right vs. left-handed ASL signers [3].

We used the RMSprop optimizer [11], Categorical-Crossentropy loss function, and a CosineDecay learning rate scheduler for fine-tuning. Both the optimizer and loss function were only applied to the classifier output; the video embedding output was not trained. Due to memory limitations, we had to unfreeze one 3D convolution block at a time, train for 2 epochs, and then progress to the next block. We trained the model for 48 hours over 9 epochs on an NVIDIA A10G Tensor Core GPU. Although the model continued to show improvement over each epoch, we prematurely ended the training due to resource constraints.

3.2. Language Model

For the language modeling component, we used the T5-base model from Hugging Face [10]. We additionally tested the T5-large, T5 v1.1-base/large, and FLAN-base/large models, but found the original T5 models to work best. We opted to use the base model due to GPU memory constraints. The T5-base model has an embedding size of 768.

For training, since we did not have tokenized text

Stage	Operation	Output size
data	stride 5, RGB	N×224^2
$conv_1$	1×3^2, 16	N×112^2
$block_2$	$\begin{bmatrix} 1\times5^2, 16, 40 \\ 3\times3^2, 16, 40 \\ 3\times3^2, 16, 64 \end{bmatrix}$	N×56^2
$block_3$	$\begin{bmatrix} 3\times3^2, 40, 96 \\ 3\times3^2, 40, 120 \\ 3\times3^2, 40, 96 \\ 3\times3^2, 40, 96 \\ 3\times3^2, 40, 120 \end{bmatrix}$	N×28^2
$block_4$	$\begin{bmatrix} 5\times3^2, 72, 240 \\ 3\times3^2, 72, 160 \\ 3\times3^2, 72, 240 \\ 3\times3^2, 72, 192 \\ 3\times3^2, 72, 240 \end{bmatrix}$	N×14^2
$block_5$	$\begin{bmatrix} 5\times3^2, 72, 240 \\ 3\times3^2, 72, 240 \\ 3\times3^2, 72, 240 \\ 3\times3^2, 72, 240 \\ 1\times5^2, 72, 144 \\ 3\times3^2, 72, 240 \end{bmatrix}$	N×14^2
$block_6$	$\begin{bmatrix} 5\times3^2, 144, 480 \\ 1\times5^2, 144, 384 \\ 1\times5^2, 144, 384 \\ 1\times5^2, 144, 480 \\ 1\times5^2, 144, 480 \\ 3\times3^2, 144, 480 \\ 1\times3^2, 144, 576 \end{bmatrix}$	N×7^2
$conv_7$	2×1^2, 96	$N \times 6^2$
$conv_8$	3×1^2, 48	$N \times 48^2$
$flatten_9$	1×1^2, 768	$N \times 768$
$dense_{10}$	1×1^2, 107	1×1^2

Table 3. **Movinets A2 Modified Architecture**, where N represents the number of input frames. The final four layers are modified to contain two 2D convolutions, followed by a flattening layer to output an intermediate embedding shape of $N \times 768$

for the input to the model, we had to bypass the embedding layer and instead input our video embeddings (with dimensions $N \times 768$, and padded to a fixed sequence length) directly into the encoder. We passed the tokenized caption corresponding to the video as the label, setting the maximum token length to 128 and padding or truncating the captions as needed.

We found that the model struggled to learn when trained on the captions. The captions did not always form complete sentences, sometimes beginning in the middle of a sentence or consisting of only one or two words. This lack of context or grammatical structure in the captions presented an additional challenge to the model. For this reason, we fine-tuned T5 in a two step approach, first training the model to generate individual words, and then further fine-tuning it to generate full captions.

3.2.1 Word-Level Generation

Before training on complete sentences, we started by training T5 to generate predictions for single words at a time. We passed a dataset of 25K video embeddings of words from the WLASL and MS-ASL datasets, containing 2000 unique words, and allowed the model to generate up to 128 tokens. For the text generation parameters, we set the temperature to 0.01, top_k to 50, and top_p to 0.90.

For fine-tuning, we used Adafactor [13] for the optimizer, SparseCategoricalCrossentropy for the loss, and a CosineDecay learning rate scheduler. We limited the input sequence length to 265 frames, padding or truncating all video embeddings to this length and providing its corresponding attention mask. We set the batch size to 32 (the largest our GPU memory could afford), used an initial learning rate of 0.0001 with a warm-up of 50 steps, and trained the model for 15 epochs.

3.2.2 Sentence-Level Generation

After fine-tuning the T5 model on individual words, we did further fine-tuning to generate full captions or sentences. Due to resource constraints, we were limited to a dataset of 20K captions from YouTube-ASL, containing roughly 13K unique words.

We used a similar training configuration as described in Section 3.2.1 for the word-level fine-tuning but set the initial learning rate to 0.0005 with a warm-up of 600 steps. The maximum input sequence length was set to 320 frames (the 95th percentile for frame counts across our captions). All videos were either padded or cropped to this length and input to the model along with an attention mask. Because of the larger input data size, our batch size was reduced to 24. We trained the model for 10 epochs.

4. Results

4.1. Modified MoViNets Classifier

Our modified MoViNets model achieved a final top-1 validation accuracy of 0.17, and a top-5 validation accuracy of 0.29. Although these are low scores, we did not need a strong classifier since our interest was purely in the intermediate embeddings.

With the model trained, we iterated through all our data in WLASL, MS-ASL, and YouTube-ASL and generated our MoViNets video embeddings, preserving the original number of frames.

4.2. T5 Word-Level Generator

Our T5 model, fine-tuned on individual words, achieved a top-1 validation accuracy score of 0.56, an improvement of 0.39 compared to the 0.17 accuracy from the MoViNets classifier. It is worth noting that the MoViNets classifier was trained on only 107 words,

while T5 was now able to classify 2000 unique words correctly more than half the time.

Looking deeper into the model predictions, we found that the model would often predict synonyms or related terms instead, as shown in Table 6. Common themes, like animals, numbers, or locations would often get confused. Upon manual inspection of a few examples like this, we found that in some cases, the same video file would be listed twice in the dataset under two translations (eg. picture and image, cop and police, or eyeglasses and glasses). In other cases, we found that related terms would have similar root hand gestures. For example, the signs for some colors follow a similar pattern, involving a twisting wrist motion of the hand near the chin, but with different shapes of the fingers or hand. There are further instances, such as the signs for care and for careful, which have identical arm and hand gestures but differ only in facial expression. The model likely mislabeled cases like these due to limitations in the MoViNets model. Perhaps with a significantly larger dataset, both in size and in number of classes, the CNN model could learn to identify these features. Alternatively, the model may require multiple input embeddings, potentially including multiple CNN models each trained to focus on one subpart of a gesture (eg. hands, facial expression, body pose).

In order to more fairly evaluate the model performance given the similarly themed terms that were mislabeled, we used the SentenceTransformers *all-mpnet-base-v2* model [12] to compute the cosine similarity between predictions and labels. We found an average cosine similarity of 0.65.

Although this model is a text generator rather than a classifier, we compare our accuracy against other state-of-the-art ASL word classifiers. To date, there are highly accurate word-level models that make use of transformer architectures [14], but these models are mostly limited to small numbers of classes and tend to degrade in performance as the number of classes increases. Table 4 lists some of the top-performing classifier models alongside our MoViNets classifier model and our T5 word-generator. Although the T5 word-level accuracy is only 0.56, it is an improvement over the other 2000-word WLASL I3D model.

4.3. T5 Sentence-Level Generator

The further fine-tuned sentence-level model achieved a validation BLEU score of 1.98 (calculated using SacreBLEU version 2 [9]) and an average Sentence-Transformers cosine similarity of 0.21.

Table 5 compares our model's performance against the top-performing models for ASL translation from Google and How2Sign. It is important to note that in both their cases, they tested their models on a dataset of how-to videos, which have more focused language than the general videos we used in our evaluation, so these metrics are not exactly one-to-one.

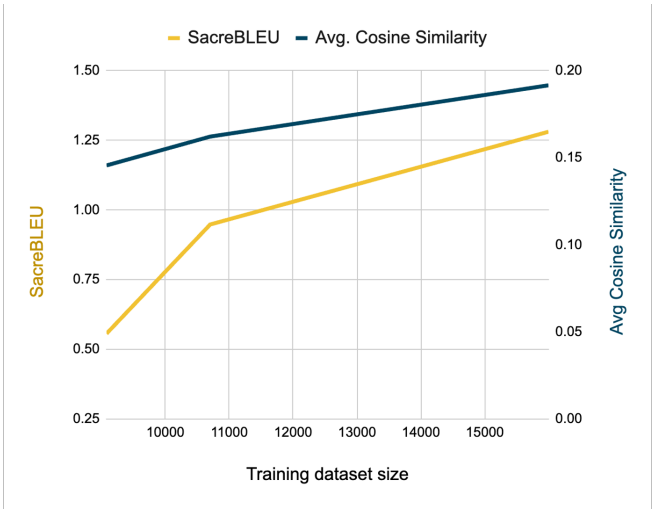

Figure 2. **Impact of increasing training sample size.** The plot shows the model's SacreBLEU score (on the left y-axis) and average SentenceTransformers cosine similarity (on the right y-axis) when trained on increasing larger datasets. For each increment of 2000 datapoints, the model performance improves considerably.

We would also like to highlight that while our model falls short of the best-achieved metrics, it has been trained on significantly fewer data points, and evaluated on a broader test dataset of general YouTube-ASL videos. The top performing model, created by the YouTube-ASL team, achieved a top BLEU score of 12.39 when trained on over 600K captions; but when trained on a dataset of 45K captions, its BLEU score was 1.22 when evaluated on the How2Sign dataset. In comparison, our model achieved a BLEU score of 1.98 when evaluated on YouTube-ASL data despite being trained on only 20K captions.

To better understand how the dataset size impacted our model, we performed an experiment, training for just 5 epochs on increasingly larger training sets. Figure 4.3 demonstrates that the model improves considerably with each addition of 2000 data points, suggesting that with additional resources and a larger dataset, our model architecture could achieve competitive results.

5. Conclusion

Machine translation of American Sign Language (ASL) remains an ongoing challenge. This paper presents a new architecture for the translation of ASL comprised of a fine-tuned MoViNets CNN model

Model	100 Words	200 Words	2000 Words
WLASL I3D [6]	0.66	-	0.32
WLASL Kaggle [14] (Mediapipe + Transformer)	-	0.89	-
Ours (MoViNets classifier)	0.17	-	-
Ours (MoViNets + T5)	-	-	0.56

Table 4. **ASL Word-Level Model Accuracy Scores**. We compare our T5 Word-Level model against ASL word classifiers. For each model, its number of classes and accuracy scores are provided.

Approach	Train Data	Eval Data	Total Caption Count	BLEU
GloFE-VN [7]	H2S	H2S	45K	2.24
Tarrés et al.[15]	H2S	H2S	45K	8.03
YouTube-ASL [16]	H2S	H2S	45K	1.22
	YT-ASL	H2S	600K	3.95
	YT-ASL → H2S	H2S	¿600K	12.39
Ours	WLASL + MS-ASL → YT-ASL	YT-ASL	20K	1.98

Table 5. **Metrics for ASL to English translation.** H2S refers to the How2Sign dataset, YT-ASL refers to the YouTube-ASL dataset. YT-ASL → H2S refers to training on YouTube-ASL and then further fine-tuning on How2Sign. Finally, in our case, we trained on the WLASL and MS-ASL datasets, and further fine-tuned on YouTube-ASL. The final SacreBLEU scores as well as the approximate data sizes for each of these models are provided.

to produce video embeddings and a fine-tuned T5 encoder-decoder model to generate translations from the video embeddings. Using this architecture, our model achieved results comparable to state-of-the-art solutions despite operating with a smaller dataset. Further training of the model presented here on a larger data set would be a natural progression to evaluate its efficacy. A significant limitation of this architecture is the large size of the MoViNets model, which runs slowly and therefore may not be practical for live-translation. Exploration of smaller architectures for the CNN component in future work would help alleviate this shortcoming.

Acknowledgements

The research in this paper benefited considerably from several discussions with our instructors. We gratefully acknowledge Mark Butler, Cornelia Ilin, and Zona Kostic, for their unwavering guidance both during and outside of lecture. We also thank the providers of the YouTube-ASL, WLASL, MS-ASL, and How2Sign datasets, which were used in our modeling efforts. Additionally, we would like to give our heartfelt thanks to Amanda Duarte, Laia Tarrés Benet, Dan Kondratyuk, Jenny Buechner, Kira Wetzel, and Danie Theron for their support during the various stages of this project.

References

[1] Jenny Buechner. Personal interview, November 16 2023. President of the National Association of the Deaf.

[2] Amanda Duarte, Samuel Albanie, Xavier Giró i Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries, 2022.

[3] Handspeak. Asl signing for left-handed individuals, Accessed: December 6, 2023.

[4] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language, 2019.

[5] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition, 2021.

[6] Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, 2020.

[7] Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. Gloss-free end-to-end sign language translation, 2023.

[8] Google LLC. Mediapipe: Cross-platform, customizable ml solutions for live and streaming media, Accessed: December 6, 2023.

[9] Matt Post. A call for clarity in reporting bleu scores, 2018.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[11] Saragada Reddy, K. Reddy, and V. Vallikumari. Optimization of deep learning using various optimizers, loss functions and dropout. *International Journal of Recent Technology and Engineering*, 7: 448–455, 01 2018.

[12] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[13] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost, 2018.

[14] Hoyeol Sohn. First-place solution for google - isolated sign language recognition kaggle competition, 2023. URL `https://www.kaggle.com/competitions/asl-signs/discussion/406684`.

[15] Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos, 2023.

[16] David Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus, 2023.

6. Appendix

Label	Prediction	Cosine Similarity
jail	prison	0.925136507
downstairs	upstairs	0.906025946
dorm	dormitory	0.890939891
mom	mother	0.885809124
awful	terrible	0.883836389
cop	policeman	0.878503382
dad	father	0.877186656
physician	doctor	0.872051835
many	numerous	0.860269904
choose	choice	0.853119254
nineteen	eighteen	0.852182686
smell	odor	0.839136362
sixteen	eighteen	0.833990157
image	picture	0.831962466
two	three	0.829272568
gas	gasoline	0.826644242
odd	weird	0.819896817
one	two	0.818584502
yourself	myself	0.815259039
boots	shoes	0.812349737
my	mine	0.809549153
four	three	0.808388174
november	december	0.804248929
glasses	eyeglasse	0.796240926
eyeglasse	glasses	0.796240866
basketball	volleyball	0.785786808
police	cop	0.785434961
quick	fast	0.784223914
seven	eight	0.77988565
girlfriend	boyfriend	0.775026083
nephew	niece	0.766748369
mother	father	0.755923152
necklace	jewelry	0.753350377
one	three	0.746693313
psychologist	psychology	0.74485153
correct	right	0.743494213

Table 6. **Related Term Predictions.** The model often predicts terms that are synonyms or related terms.

Label	Prediction	Cosine Similarity
but	chair	0.021876482
biology	hanuk	0.02349641
preach	europe	0.038973857
still	camping	0.040663257
until	scot	0.040889744
soap	able	0.043406848
chemistry	hanuk	0.043694299
guide	most	0.043999713
camera	retire	0.047875576
ready	people	0.048999496
flirt	screwdriver	0.056134962
thank you	education	0.056145877
fast	certificate	0.056815136
then	practice	0.057618942
drawer	believe	0.057674389
september	ro	0.058547124
always	israel	0.058576863
team	anatomy	0.065127857
preach	summer	0.067174107
study	rehe	0.067501262
stand	monthly	0.069805004
puzzled	gallaude	0.069984712
possible	apos	0.071164407
wow	any	0.073385336
teach	house	0.073960878
compare	director	0.074195437
dormitory	listen	0.074779183
free	pray	0.076081157
bite	physics	0.076826833
except	hanuk	0.077750593
much	russia	0.077941388
because	gira	0.078685805
tournament	or	0.078722745
sell	cuba	0.080816194
soon	dollar	0.084375046
bracelet	move	0.084493935

Table 7. **Unrelated Predictions.**

Caption	Prediction	Cosine Similarity
praise the lord	praise the lord	1.0000
fox	fox	1.0000
delicious	delicious	1.0000
rainbows rainbows high up in the sky,	rainbows rainbow high up in the sky.	0.9709
a d grade	a c grade	0.9204
scrub your hands for at least 20 seconds.	dry your hands using a clean towel.	0.8944
school performances for deaf children	encouraging deaf performers to participate.	0.8635
raindrops, raindrops falling to the ground.	raindrops.	0.8338
he called himself, "i am"	jesus called himself, "i am"	0.8307
a radical life of discipleship to those around him!	a radical life of obedience to god.	0.8248
27	25	0.8190
welcome to sign1news.	sign1news.	0.8180
you are my sunshine my only sunshine	you are my sunshine	0.8100
welcome to sign1news.	thank you for watching sign1news.	0.7968
baby and dad!	baby and boy!	0.7553
my second asl professor's name was will white.	my second asl professor's name was robert sanderson.	0.7457
again, the number is (866) 246 0133	the phone number is: +61 866 328 8933	0.7403
snowflake, right? slowly, slowly. right?	snowflakes snowflake?	0.6992
pink	white	0.6518
the interpreting and deaf communities may look at these messages, empathize with their similar experiences,	social media has become a tool of empowerment for deaf people.	0.6275
n: it's hot, we've been out at sea for so long.	a: i miss fish. m: it's too fast.	0.6238
language delays are from a lack of exposure to early language development.	language deprivation is when a child is unable to understand spoken language.	0.6184
the boy gets sad when he realizes his gum is no longer there	the little boy sees a bench where he puts his gum	0.6154
see that? i sign what's west and east for me, not for you.	if you're going to be the "right" direction, you sign as you would see it. you go right.	0.5914
the deaf studies department emailed me the application forms and instructions.	so i asked them to help me with my application and if they would be interested in it. they said no. esl is for deaf people.	0.5834
the board has also been working with iad.	the board has also been working to provide leadership training	0.5821
over time, qabil's personality became more	qabil is adam's twin sister.	0.5720
he uses a scary situation to humble jacob when he's about to see esau.	jacob often goes through trials and sanctification,	0.5680
juice	milk	0.5596
so i'm really interested in seeing how this all turns out	so i think this is going to be really exciting	0.5524
hey! all you had do past stay	past can easy, all you have to do past stay	0.5524
not only that, if i'm eating, i can sign at the same time.	i sign because if a friend is deaf,	0.5507

Table 8. **Top model translations.**

Caption	Prediction	Cosine Similarity
his power to smear him.	four months ago, a cat named yvonne swann came to the rescue.	-0.0057
now she lives with her grandson in chechnya after losing her husband just 10 years ago.	the smithsonian museums and galleries are also offering a wide variety of exhibitions.	-0.0042
which is her husband's ministry	and i am willing to accept it but he is not willing	-0.0036
same!	i want to remind you of some rules	-0.0030
implementation of federal accessible canada act,	[anselmo]: hello!	-0.0012
oreo is offering a sweet prize. anyone who	at least 20 people have died.	-0.0010
the camp will run from july 14 to july 20	this is a fad.	-0.0007
they later went to an apartment	so if you have a snort on your phone, text the doctor,	-0.0002
amount of time, like on snapchat and instagram stories.	the national transportation safety board is working closely with the state and federal agencies	0.0000
sign language is important to preserving	the playwright was the one who wrote the plays.	0.0019
if you are still writing and let go when the time runs out,	you can give us your answer before we start discussing.	0.0024
fox?	i'm not sure why esl is important.	0.0025
after this, my clients came back and could easily sign	language	0.0027
and she likes it.	i have a patreon and e mail account.	0.0028
and that's with most people staying home!	it means that people who have antibodies, antibodies and antibodies	0.0046
wait	the paper is clean and dry.	0.0050
i feel rejuvenated by this beautiful weather.	it is not a problem for me because my back is stiff and my shoulder isn't flexible.	0.0080
hello, ryan!	yes, i know that the seattle seaport is beautiful	0.0087
in the near future.	the fcc says if you're going to try to avoid this, you will lose your chance of winning.	0.0102
north dakota, they don't need	you can register for either the online or in person	0.0103
now, it's 2016! friday evening, june 17th is the 10 anniversary celebration!	i'm a deaf actor.	0.0110
this story is one of the most shocking,	to receive their degrees at ivy league	0.0123
champ = "the best of the best"	deaf people feel the same way.	0.0146
and what are your deaf kids doing for the summer?	this is edward shaw.	0.0177
the best way to contact our team is to e mail	and to be supported by their deaf and hard of hearing peers.	0.0184
if you prefer to talk with a real person.	the country's president has declared emergency.	0.0193

Table 9. **Poorly translated sentences.**