SignSense

American Sign Language Translation

Synthetic Capstone
12/14/2023

Nashat Cabral
Deanna Emery
Deepak Krishnamurthy

Nashat Cabral

Business Intelligence Analyst

Deanna Emery

Senior Data Science Manager

Deepak Krishnamurthy

Data Engineer

1.8
Million

PEOPLE IN THE U.S. WITH SEVERE HEARING LOSS

500
Thousand

USERS OF AMERICAN SIGN LANGUAGE IN THE U.S.

10
Thousand

TOTAL CERTIFIED AMERICAN SIGN LANGUAGE INTERPRETERS

Problem Statement

ASL speakers face obstacles due to the absence of real-time translation, leading to limited accessibility, reliance on interpreters, barriers in dynamic settings, reduced independence, and an inclusive technology gap.

“ … with less time to develop [language] from youth, [the deaf community] prefer ASL because their English is not strong. ”

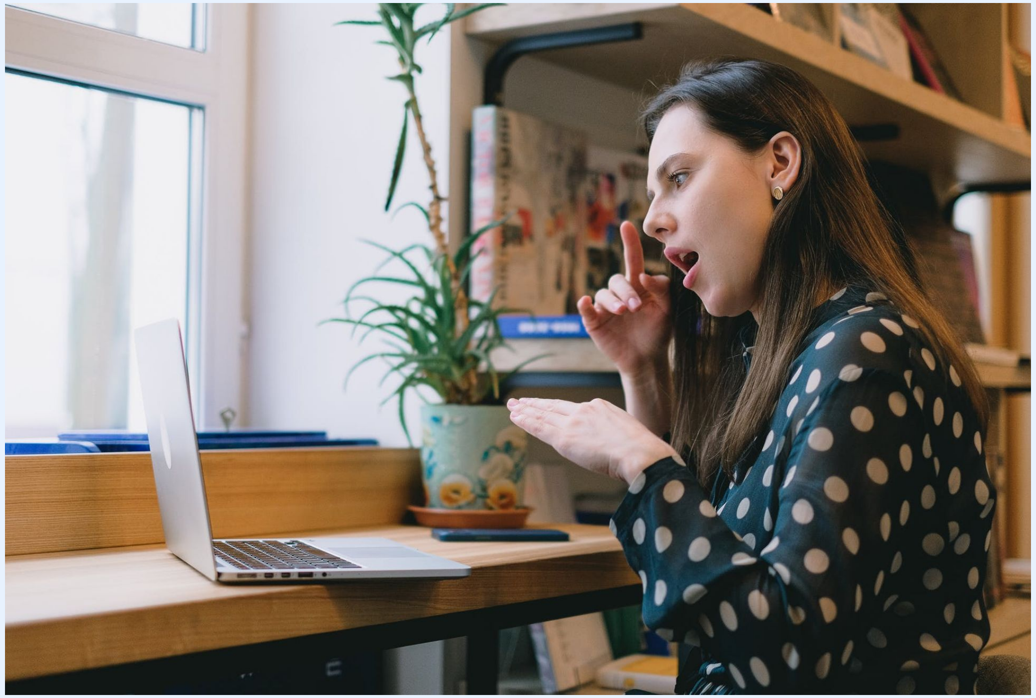
- Jenny Buechner,
  President of the National
  Association of the Deaf

Research Objectives

1 Develop a **model infrastructure** for the translation of **sentence -level** sign language videos to English

2 Train the model to translate given sign language to a desired level of **accuracy**

3 Improve **efficiency** of model to potentially be applied to future attempts at live translation



Impact

- Impact Areas

  Social Inclusion

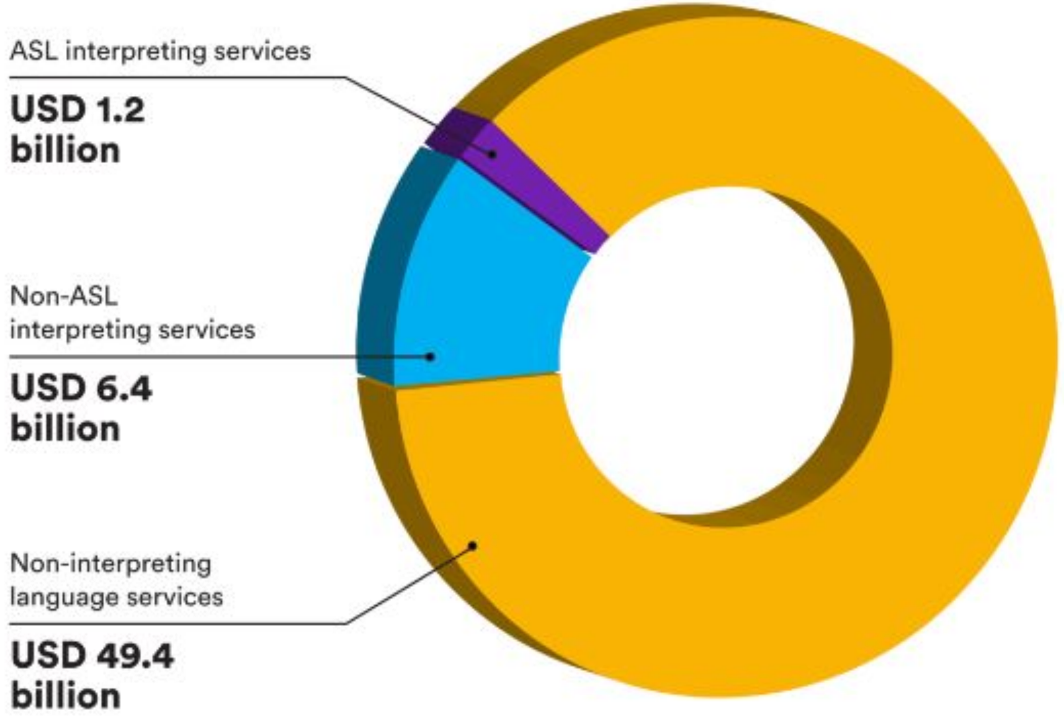
  Empowerment

- Social Impact

  Social inclusion and empowerment in education, healthcare and everyday interactions

- Monetary Impact

  ASL Interpretation market estimated at $1.2B in 2021



https://www.nimdzi.com/asl-interpreting/

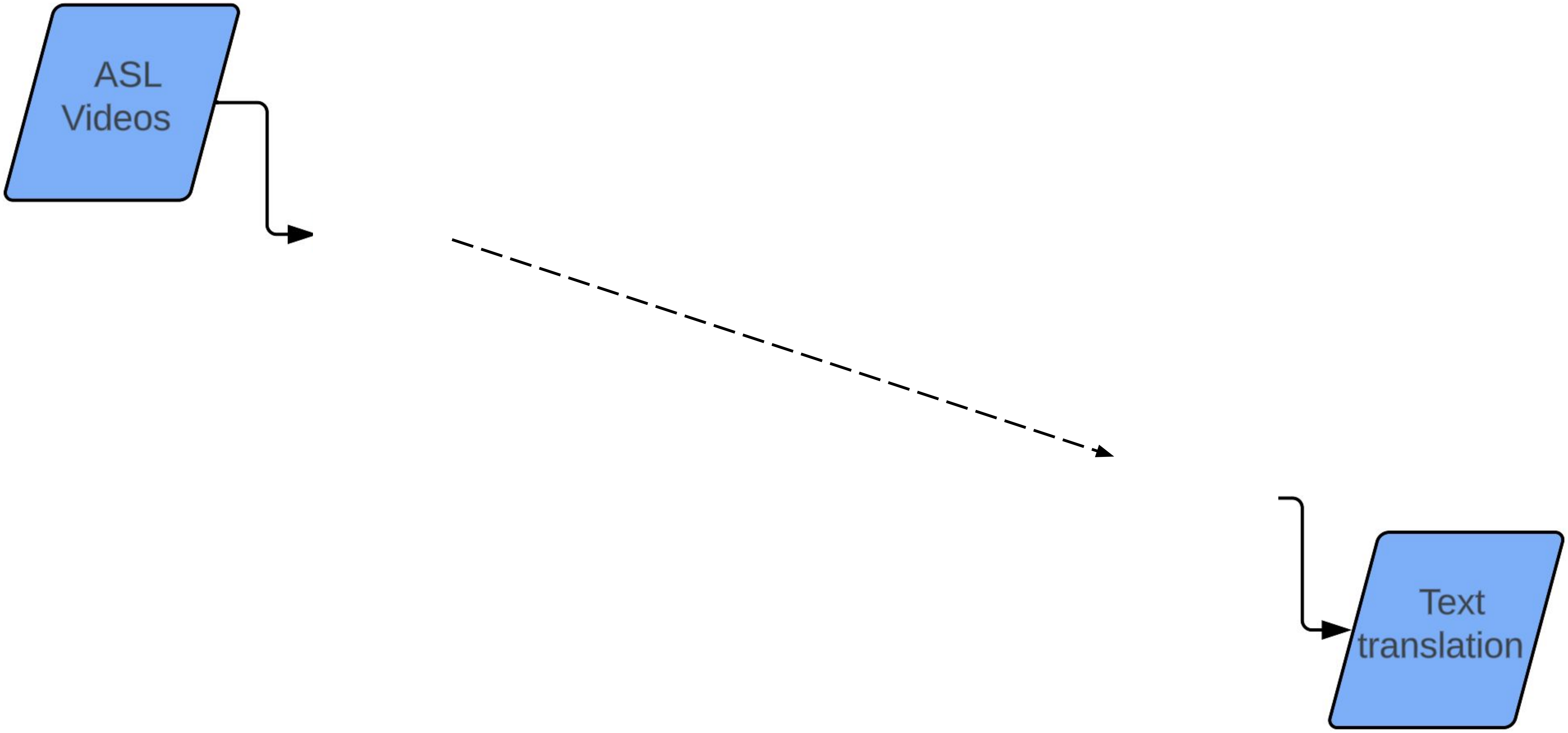
Target Users

- Target Users:
  - Research-Focused Community
  - Automated ASL Translation Developers
  - Data Contributors
- Model shared on Hugging Face

Data Sources

Dataset	Type	Vocabulary	# Signers	# Hours	# Videos
WLASL	Words	2000	119	14	25513
MS-ASL	Words	1000	22	25	21083
YouTube-ASL	Sentences	60000	2519	984	11093

Modeling Approach



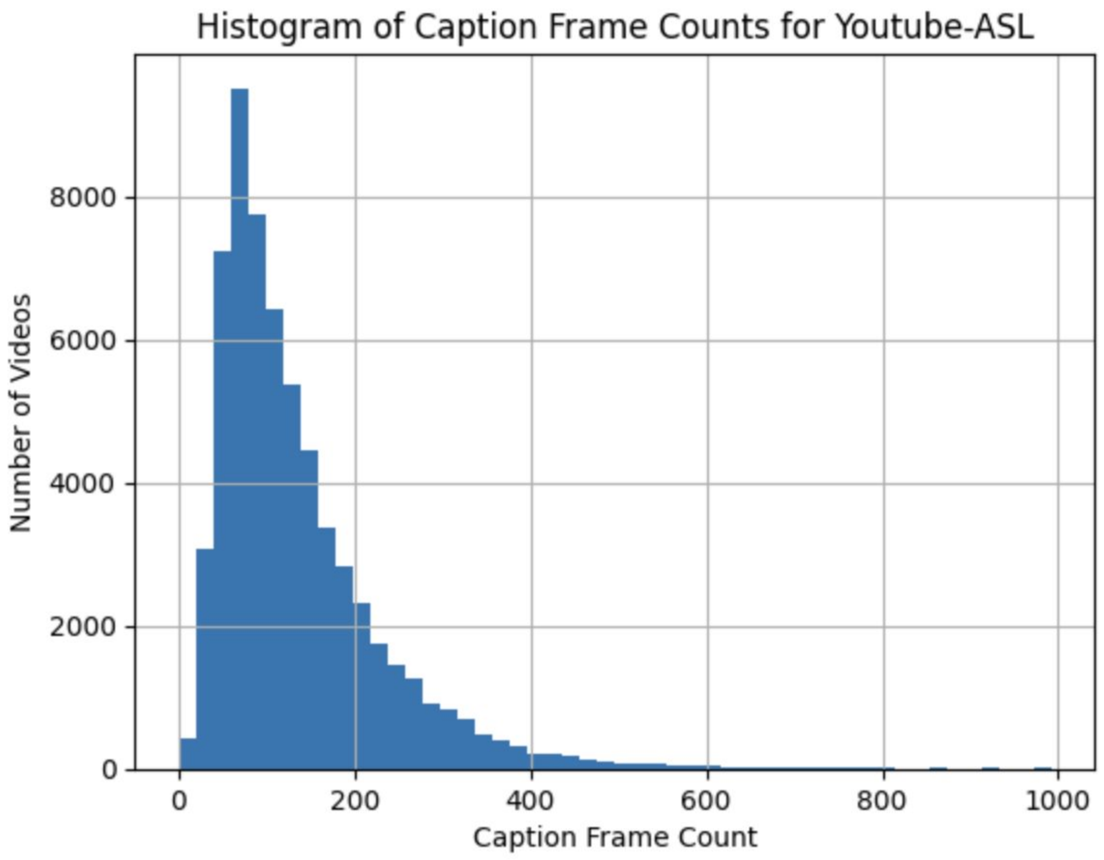
Data Processing

- Sample of YouTube-ASL used along with WL-ASL and MS-ASL
- OpenCV used to convert videos to Numpy arrays and stored with captions
- Captions cleaned – removing special characters and spacing
- Numpy arrays converted to Float32 for quicker processing

Caption	frame_rate	start_time_seconds	end_time_seconds
Hello everyone.	29.97003	6.320	7.440
Welcome to Sign1News.	29.97003	7.440	10.020
I'm Candace Jones.	29.97003	10.020	11.220
Here are your top stories for today.	29.97003	11.220	14.500

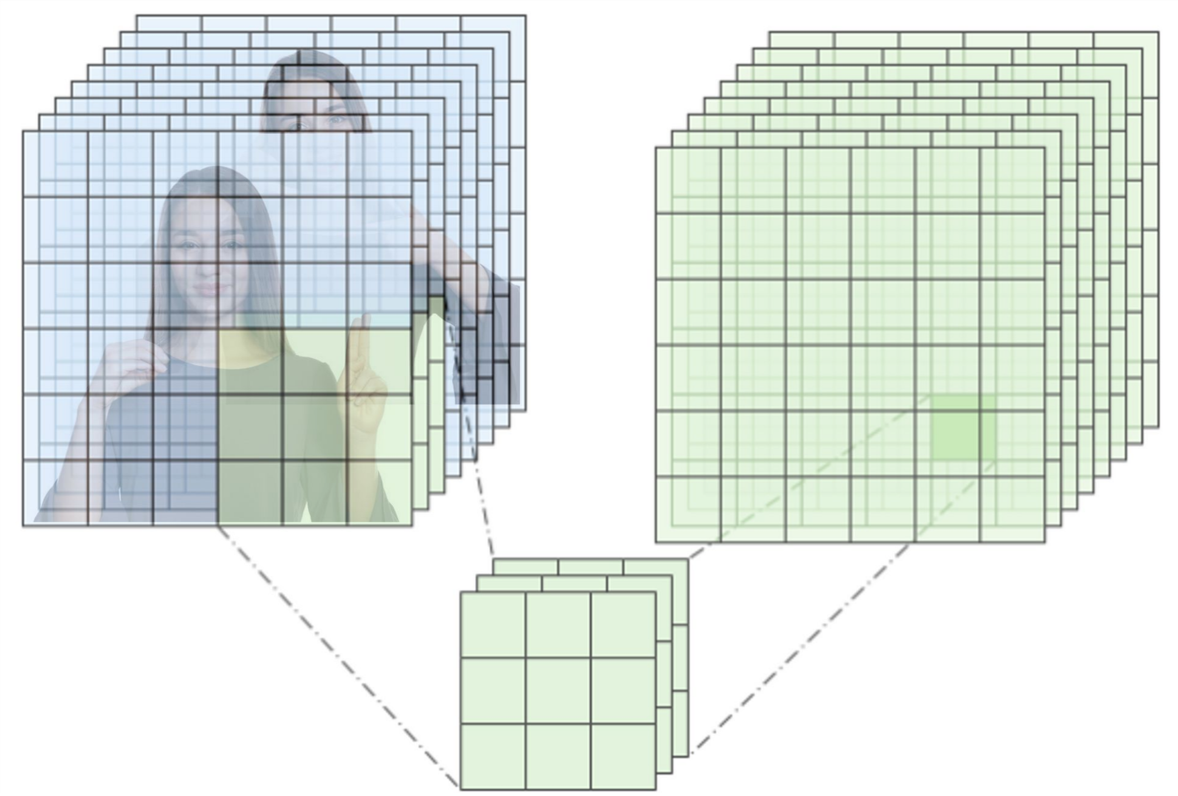
Data Processing

- Sample of YouTube-ASL used along with WL-ASL and MS-ASL
- OpenCV used to convert videos to Numpy arrays and stored with captions
- Captions cleaned – removing special characters and spacing
- Numpy arrays converted to Float32 for quicker processing



Video Embeddings

- **MoViNet**: action-recognition model pre-trained on the Kinetics600 dataset
- Uses a series of **3D convolutions** to capture temporal features across video frames
- Outperforms other modern action recognition models on Kinetics datasets
  - e.g. I3D, ViVit, VATT, X3D, MobileNetV3
- Comparatively light-weight
  - 3M trainable parameters vs 10-100M for other models



MoViNet: Modified Architecture

Stage	Operation	Output size
data	stride 5, RGB	frames $\times$ 224^2
conv$_1$	1×3^2, 16	frames $\times$ 112^2
block$_2$	1×5^2, 16, 40 3×3^2, 16, 40	frames $\times$ 56^2

- Sparse word coverage – filtered to 107 most frequent words
- Unfreeze 5 layers at a time, train for 2 epochs
- 48 hours of training over 9 epochs
- Validation Accuracy –
    - Top-1 accuracy: 0.17
    - Top-5 accuracy: 0.29

	3×3^2, 144, 480 1×3^2, 144, 576	
conv$_7$	1×1^2, 640	frames $\times$ 7^2
pool$_8$	frames $\times$ 7^2	1×1^2
dense$_9$	1×1^2, 2048	1×1^2
dense$_{10}$	1×1^2, 600	1×1^2

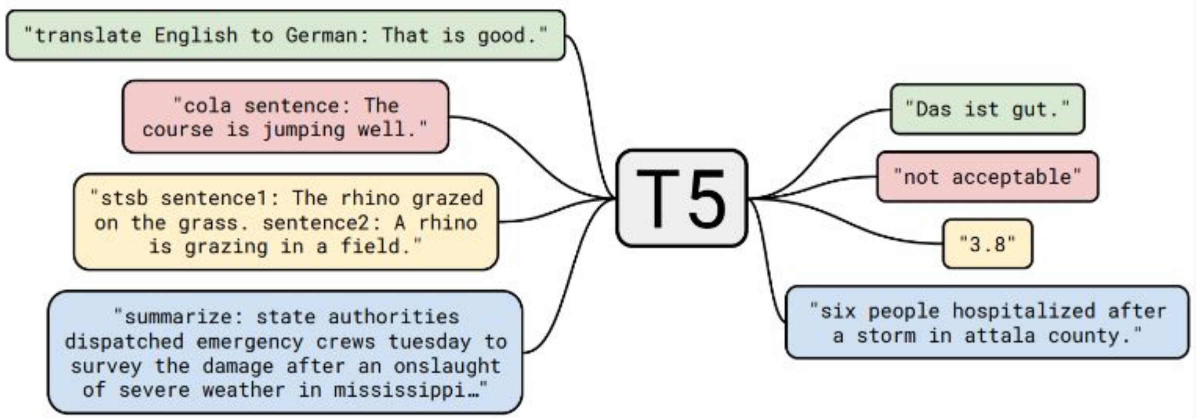
modified

Stage	Operation	Output size
data	stride 5, RGB	frames $\times$ 224^2
conv$_1$	1×3^2, 16	frames $\times$ 112^2
block$_2$	1×5^2, 16, 40 3×3^2, 16, 40 3×3^2, 16, 64	frames $\times$ 56^2
block$_3$	3×3^2, 40, 96 3×3^2, 40, 120 3×3^2, 40, 96 3×3^2, 40, 96 3×3^2, 40, 120	frames $\times$ 28^2
block$_4$	5×3^2, 72, 240 3×3^2, 72, 160 3×3^2, 72, 240 3×3^2, 72, 192 3×3^2, 72, 240	frames $\times$ 14^2
block$_5$	5×3^2, 72, 240 3×3^2, 72, 240 3×3^2, 72, 240 3×3^2, 72, 240 1×5^2, 72, 144 3×3^2, 72, 240	frames $\times$ 14^2
block$_6$	5×3^2, 144, 480 1×5^2, 144, 384 1×5^2, 144, 384 1×5^2, 144, 480 1×5^2, 144, 480 3×3^2, 144, 480 1×3^2, 144, 576	frames $\times$ 7^2
conv$_7$	2×1^2, 96	frames $\times$ $6^2 \times 96$
conv$_8$	3×1^2, 4	frames $\times$ $4^2 \times 48$
flatten$_9$	1×1^2, 768	frames $\times$ 768
dense$_{10}$	1×1^2, 107	1×107

Language Model

T5: Encoder-Decoder Model

- Bypass the tokenization step and embedding layer in the model
- Input video embeddings as if they are text embeddings
- Tokenized caption is the label for fine-tuning



Layer (type)	Output Shape	Param #
shared (Embedding)	multiple × 768	24674304
encoder (TFT5MainLayer)	multiple	109628544
decoder (TFT5MainLayer)	multiple	137949312

```
Total params: 222903552 (850.31 MB)
Trainable params: 222903552 (850.31 MB)
Non-trainable params: 0 (0.00 Byte)
```

T5 Fine-Tuning

Word-Level Generation

- 25k files from WLASL and MS-ASL
- 2000 unique words
- Accuracy score: 0.56
- Average cosine similarity: 0.65 (SentenceTransformers)

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

T5 Fine-Tuning

Sentence-Level Generation

- Further fine-tuned the word-level model on complete sentences
  - 20k files from Youtube-ASL
  - 13k unique words
- SacreBLEU score: 1.98
- Average cosine similarity: 0.21 (SentenceTransformers)

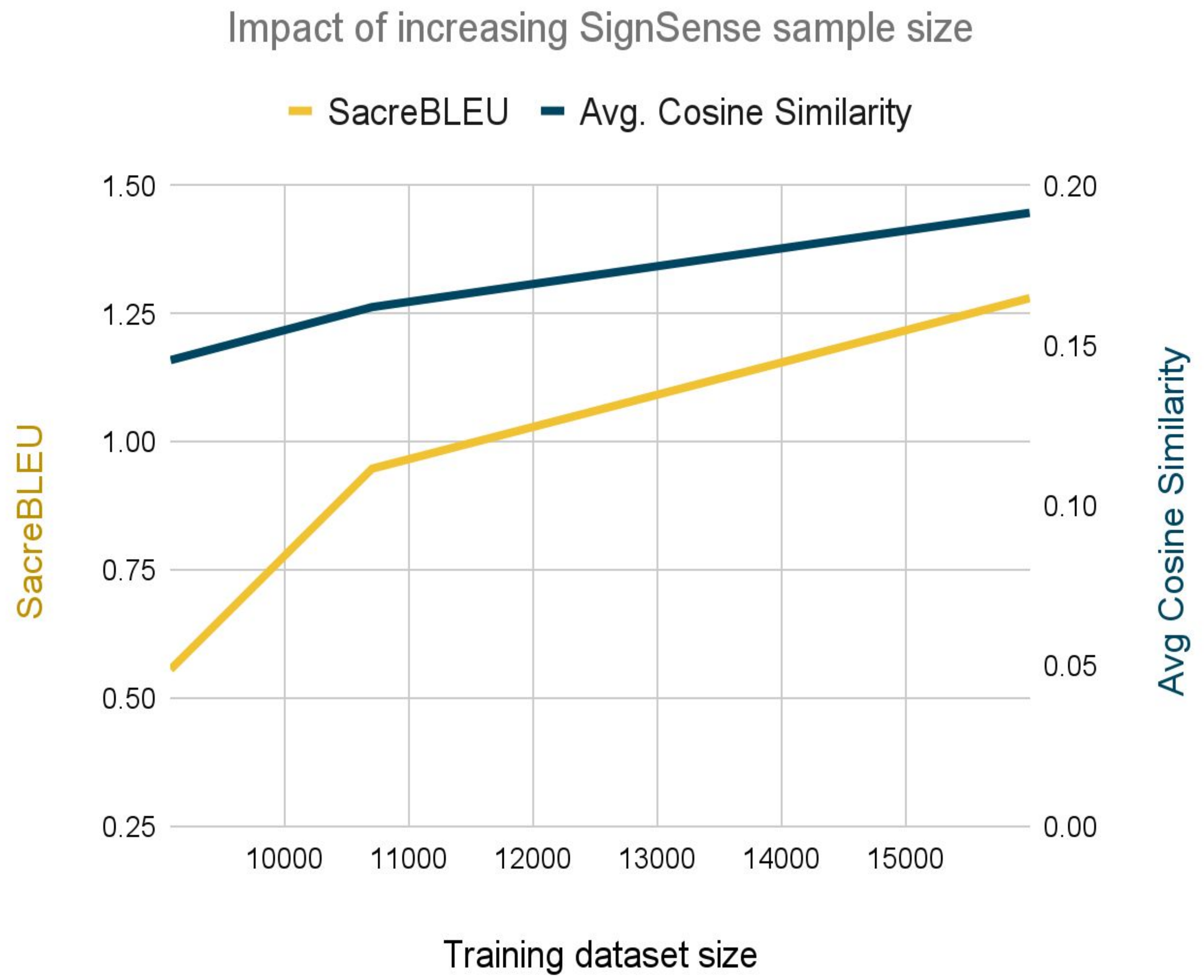
Caption	Prediction
praise the lord	praise the lord
fox	fox
delicious	delicious
rainbows rainbows high up in the sky,	rainbows rainbow high up in the sky.
a d grade	a c grade
scrub your hands for at least 20 seconds.	dry your hands using a clean towel.
school performances for deaf children	encouraging deaf performers to participate.
raindrops, raindrops falling to the ground.	raindrops.

now, it's 2016! friday evening, june 17th is the 10 anniversary celebration!	i'm a deaf actor.
this story is one of the most shocking,	to receive their degrees at ivy league
champ = "the best of the best"	deaf people feel the same way.
and what are your deaf kids doing for the summer?	this is edward shaw.
the best way to contact our team is to e mail	and to be supported by their deaf and hard of hearing peers.
if you prefer to talk with a real person.	the country's president has declared emergency.

Evaluation

Model	Data Size	SacreBLEU
How2Sign	45k captions	2.21 / 8.03
Google	45k captions	1.22
Google	610k captions	3.95 / 12.39
SignSense (Ours)	20k captions	1.98

Evaluation



Impact of increasing SignSense sample size
SacreBLEU
Avg. Cosine Similarity
1.50
1.25
1.00
0.75
0.50
0.25
0.20
0.15
0.10
0.05
0.00
SacreBLEU
Avg Cosine Similarity
10000
11000
12000
13000
14000
15000
Training dataset size

Key Learnings

Challenges:

- Memory & Modeling Time
- Modifying CNN architecture for custom purpose
- Modifying Transformers library to handle non-textual inputs

Technical takeaways:

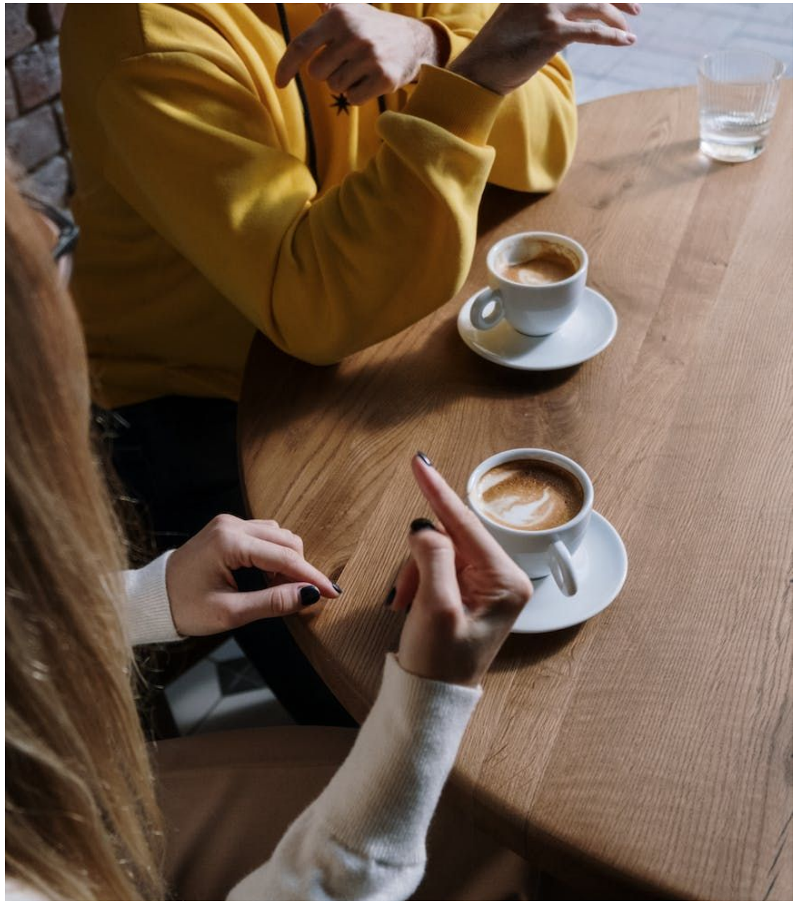
- Promising architecture achieved with limited time and cost
- CNN architecture's choice will have a large impact on inference speed

Demo

Hugging Face [Link]

Future Work

- Larger dataset + compute resources
- Quantization
- Lighter weight CNN architecture for faster inference
- Examine use of ASL classifiers and its effect on the model performance
- Compare performance across demographic groups



Conclusion

SignSense is dedicated to empowering the deaf community through the pursuit of an automated American Sign Language translation capability.

We actively encourage the ongoing collection of ASL video data to advance the creation of a truly automated translation system.

Acknowledgements

- Cornelia Ilin, Zona Kostic, and Mark Butler for their unwavering guidance both during and outside of lecture.

- The providers of the YouTube-ASL, WLASL, MS-ASL, and How2Sign datasets, which were used in our modeling efforts.

- 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 Gir´o

  i Nieto, and G¨ul 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. Sentencebert: Sentence embeddings using siamese bertnetworks. 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 Tarr´es, Gerard I. G´allego, Amanda Duarte, Jordi Torres, and Xavier Gir´o 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.