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

[Diagram showing monetary impact with categories and values:]
- ASL interpreting services: USD 1.2 billion
- Non-ASL interpreting services: USD 6.4 billion
- Non-interpreting language services: USD 49.4 billion

[Source: https://www.nimdzi.com/asi-interpreting/]
Target Users

• Target Users:
  Research-Focused Community
  Automated ASL Translation Developers
  Data Contributors

• Model shared on Hugging Face
## Data Sources

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Vocabulary</th>
<th># Signers</th>
<th># Hours</th>
<th># Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>WLASL</td>
<td>Words</td>
<td>2000</td>
<td>119</td>
<td>14</td>
<td>25513</td>
</tr>
<tr>
<td>MS–ASL</td>
<td>Words</td>
<td>1000</td>
<td>22</td>
<td>25</td>
<td>21083</td>
</tr>
<tr>
<td>YouTube–ASL</td>
<td>Sentences</td>
<td>60000</td>
<td>2519</td>
<td>984</td>
<td>11093</td>
</tr>
</tbody>
</table>
Modeling Approach

ASL Videos → Text translation
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
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**

- **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

<table>
<thead>
<tr>
<th>Stage</th>
<th>Operation</th>
<th>Output Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>stride 5, RGB</td>
<td>frames × 224²</td>
</tr>
<tr>
<td>conv₁</td>
<td>1 × 3², 16</td>
<td>frames × 112²</td>
</tr>
<tr>
<td>block₂</td>
<td>1 × 5², 16, 40, 3 × 3², 16, 40</td>
<td>frames × 56²</td>
</tr>
<tr>
<td>block₃</td>
<td>3 × 3², 16, 40, 3 × 3², 16, 64, 3 × 3², 40, 96, 3 × 3², 40, 120</td>
<td>frames × 28²</td>
</tr>
<tr>
<td>block₄</td>
<td>3 × 3², 72, 240, 3 × 3², 72, 160, 3 × 3², 72, 240, 3 × 3², 72, 192</td>
<td>frames × 14²</td>
</tr>
<tr>
<td>block₅</td>
<td>3 × 3², 72, 240, 3 × 3², 72, 240</td>
<td>frames × 14²</td>
</tr>
<tr>
<td>block₆</td>
<td>5 × 3², 144, 480, 1 × 5², 72, 144</td>
<td>frames × 7²</td>
</tr>
<tr>
<td>conv₇</td>
<td>2 × 1², 96</td>
<td>frames × 6² × 96</td>
</tr>
<tr>
<td>conv₈</td>
<td>3 × 1², 4</td>
<td>frames × 4² × 48</td>
</tr>
<tr>
<td>flatten₉</td>
<td>1 × 1², 768</td>
<td>frames × 768</td>
</tr>
<tr>
<td>dense₁₀</td>
<td>1 × 1², 107</td>
<td>frames × 107</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output Shape</th>
<th>Param #</th>
</tr>
</thead>
<tbody>
<tr>
<td>shared (Embedding)</td>
<td>multiple x 768</td>
<td>24674304</td>
</tr>
<tr>
<td>encoder (TFT5MainLayer)</td>
<td>multiple</td>
<td>109628544</td>
</tr>
<tr>
<td>decoder (TFT5MainLayer)</td>
<td>multiple</td>
<td>137949312</td>
</tr>
</tbody>
</table>

Total params: 222903552 (850.31 MB)
Trainable params: 222903552 (850.31 MB)
Non-trainable params: 0 (0.00 Byte)
Word-Level Generation

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

<table>
<thead>
<tr>
<th>CAPTION</th>
<th>PREDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>praise the lord</td>
<td>praise the lord</td>
</tr>
<tr>
<td>fox</td>
<td>fox</td>
</tr>
<tr>
<td>delicious</td>
<td>delicious</td>
</tr>
<tr>
<td>rainbows rainbows high up in the sky</td>
<td>rainbows rainbow high up in the sky.</td>
</tr>
<tr>
<td>a 0d grade</td>
<td>a 0c grade</td>
</tr>
<tr>
<td>scrub your hands for at least 20 seconds.</td>
<td>dry your hands using a clean towel.</td>
</tr>
<tr>
<td>school performances for deaf children</td>
<td>encouraging deaf performers to participate.</td>
</tr>
<tr>
<td>raindrops, raindrops falling to the ground.</td>
<td>raindrops.</td>
</tr>
</tbody>
</table>

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 email 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Data Size</th>
<th>SacreBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>How2Sign</td>
<td>45k captions</td>
<td>2.21 / 8.03</td>
</tr>
<tr>
<td>Google</td>
<td>45k captions</td>
<td>1.22</td>
</tr>
<tr>
<td>Google</td>
<td>610k captions</td>
<td>3.95 / 12.39</td>
</tr>
<tr>
<td>SignSense (Ours)</td>
<td>20k captions</td>
<td>1.98</td>
</tr>
</tbody>
</table>
Impact of increasing SignSense sample size

- SacreBLEU
- Avg. Cosine Similarity

Training dataset size

SacreBLEU

Avg Cosine Similarity
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


