Context matters: Self attention for Sign Language Recognition

Fares Ben Slimane  Mohamed Bouguessa

UQÀM
Université du Québec à Montréal
Goal: Sign Language recognition System

- Recognize continuous SL gestures.
- Capture SL gesture information from space and time.
Input: Only RGB data

Literature on Sign Language Recognition

- Glove-based methods.
- Motion tracking gadgets and sensors.

"jsc2012e237333" by NASA Johnson is licensed under CC BY-NC 2.0
Sign Language Gestures = Hand articulations + non-manual components

Multiple channels of information

- Non-manual components provide semantic context to the dominant hand.
Man VS Woman
- "Is it you?"
- "Are you ... ?"
- "Did you?"

VS

- It's you?!?!
  (I am surprised that it is you.)
Sign Language Gestures =
Dominant Handshape
+
Contextual Information

- Spatial Information
  (Around the handshape)

- Temporal Information
  (Hand and Body movements)
Key insight

“Signs require recognizing the handshape accompanied by its contextual information.”

CONTEXT MATTERS !!
Typical architecture for Sign Language Recognition

Input Clip

Spatial Modeling

CNN
3D-CNN
...

T

Temporal Modeling

RNN
LSTM
3D-CNN
...

T

Sequence Alignment

CTC
HMM
...

Transformer Network

CTC : Connectionist Temporal Classification
HMM : Hidden Markov Model

Attention Is All You Need by Ashish et al.
SIGN ATTENTION NETWORK (SAN)

2D CNN → Spatial Modeling → 2D CNN → Temporal Modeling → Sign Representations → Linear Projection → Softmax → CTC → Target Sequence

Sign Clip
Key insight:
- Signs are recognizable from the state of the body, apart from the handshape

Solution:
- Exploit spatiotemporal context around the handshape to recognize the sign
SIGN ATTENTION NETWORK
With Hand Stream

Dominant Hand

2D CNN
Spatial Modeling

Ax
Temporal Modeling

Hand Representations

Ax
Temporal Modeling

Linear Projection

Softmax

CTC
Target Sequences

LossHand

Hand Stream

SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition
Sign language components have complex relations.

**Attention**

Efficiently capture dependencies between manual and non-manual components.
SIGN ATTENTION NETWORK
With Relative local Context masking

Dominant Hand Sequence

Full Frame Sequence

Local Context (with r =10)

Global Context
Implementation Details

- Extract T keyframes (mostly 64) from original video clip.

- Resize full-frame and hand frames to 224 x 224 and 112 x 112.

- Normalize input images by subtracting the dataset’s image pixels mean.

- We use the MobilenetV2 CNN architecture for feature extraction.
Word Error Rate (WER)

$$WER = \frac{\text{#deletions} + \text{#insertions} + \text{#substitutions}}{\text{#number of reference observations}}$$

Comparison SAN variations

|               | Dev  | Test |
|---------------|------|------|
| SAN           | 35.33| 35.45|
| + Hand Stream | 33.68| 34.12|
| + Relative Local Masking | 32.74| 33.29|
Pretraining methods

| Pre-Training                      | Dev  | Test  |
|----------------------------------|------|-------|
| ImageNet                         | 32.74| 33.29 |
| RWTH-PHOENIX-Weather 2014        | 29.02| 29.78 |

Model is too complex to generalize using our dataset.

Better initialization scheme for our model by firstly training the spatial feature extractor (CNN) on the same dataset.
Quantitative Evaluation on the RWTH-PHOENIX-WEATHER 2014 DATASET in WER %

| Method                          | Dev  | Test  |
|---------------------------------|------|-------|
| SAN                             | 29   | 29.7  |
| Koller et al. (CNN-2BLSTM) [7]   | 32.7 | 32.9  |
| Koller et al. (CNN) [7]          | 33.7 | 33.3  |
| Huang et al. [9]                 | -    | 38.3  |
| Cui et al. [17]                  | 39.4 | 38.7  |
| Koller et al. [6]                | 38.3 | 38.8  |
| Camgoz et al. (HMM-LM) [10]      | 40.8 | 40.7  |
| Camgoz et al. (CTC) [10]         | 43.1 | 42.1  |
| Koller et al. [35]               | 47.1 | 45.1  |
| Koller et al. [27]               | 57.3 | 55.6  |

O. Koller, J. Forster, and H. Ney. [Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers.](https://doi.org/10.1016/j.cviu.2015.08.016) *Computer Vision and Image Understanding*, volume 141, pages 108-125, December 2015.
Qualitative Analysis

SAN

+ Hand Stream

+ Relative Local Masking
Thank you !!!
References:

- Memes created by : Meme Generator - Imgflip
- SIGN Recognition Dataset: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/

Papers:

- Video Action Transformer Network
- Neural Sign Language Translation
- SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition

Code (GitHub):

The official PyTorch implementation of "Context Matters: Self-Attention for sign Language Recognition"

- https://github.com/faresbs/slrt