UQÀM
Université du Québec à Montréal

Context matters: Self attention for Sign Language Recognition and Translation

Fares Ben Slimane

Sep, 2020

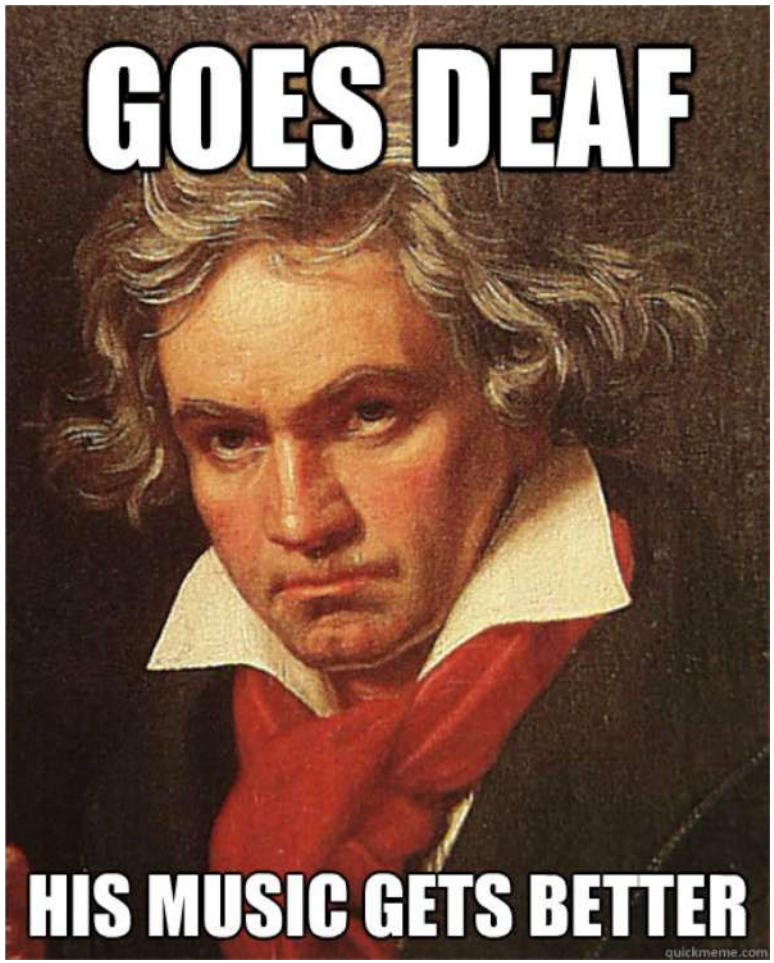
Directeur de recherche : Mohamed Bouguessa

Plan

1. What is sign language?
2. Sign Language Recognition
3. State of the art
4. Sign Language Recognition and Translation
5. Transformer Network
6. Proposed Solution
7. Conclusion

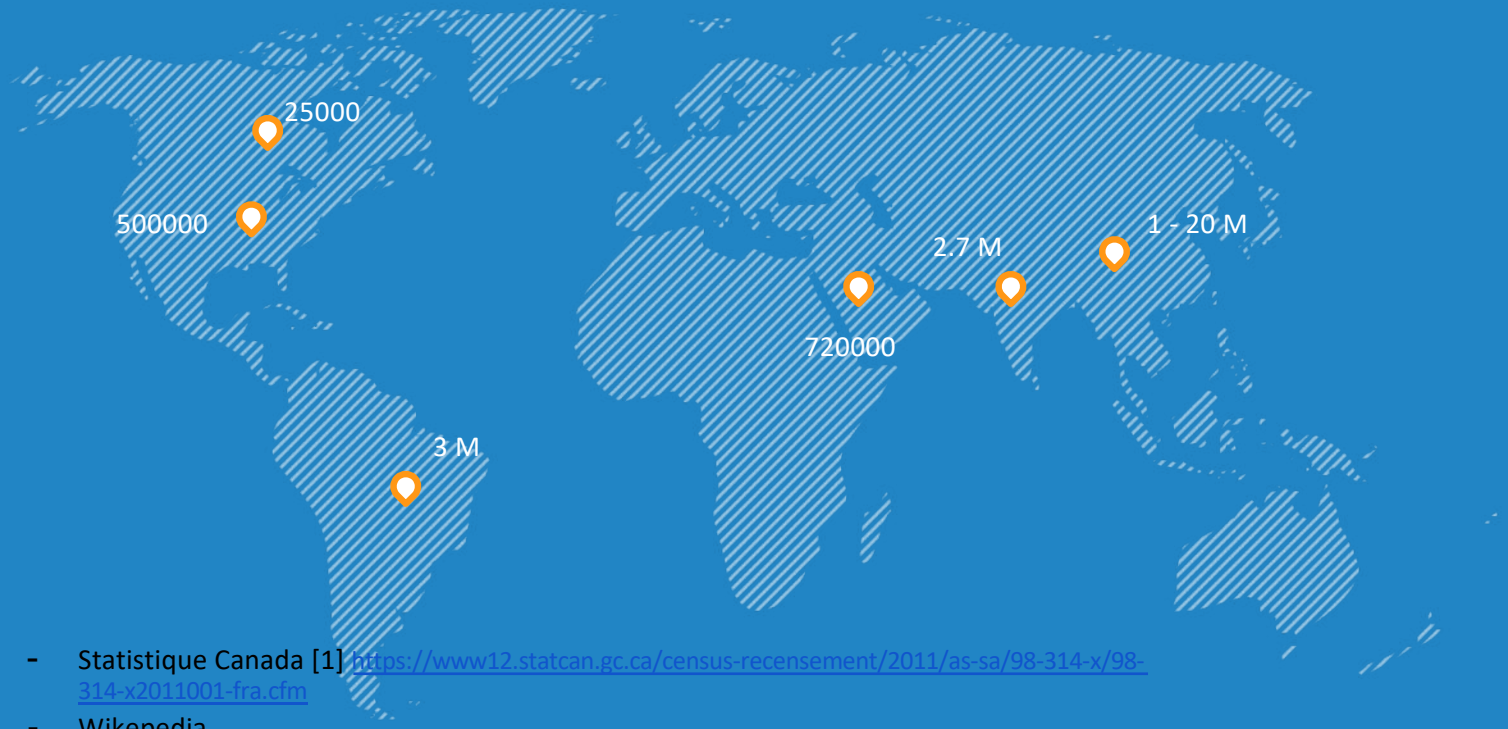
What's Sign Language?!?!

Sign language is a manual language used by deaf people and the hearing impaired to communicate.



130 Sign Languages

Estimate of Sign Language native speakers



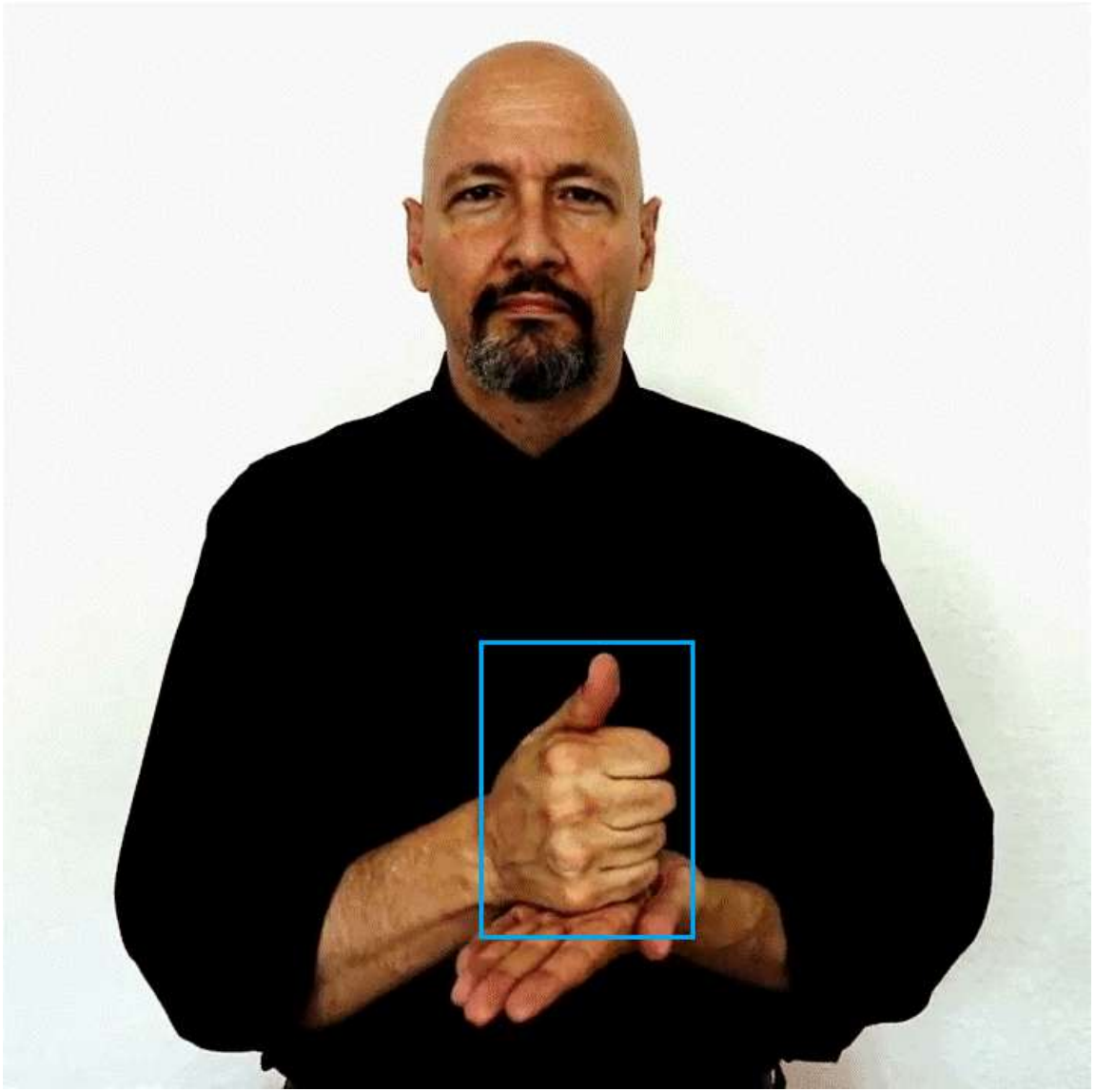
- Statistique Canada [1] https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-fra.cfm
- Wikepedia
- https://en.wikipedia.org/wiki/List_of_sign_languages_by_number_of_native_signers

Sign Language = Hand articulations + non-manual components

→ Multiple channels of information

Sign Language = Hand articulations + non-manual components

Dominant hand gets semantic context from non-manual components.



VS

Man

Woman

VS

Sit

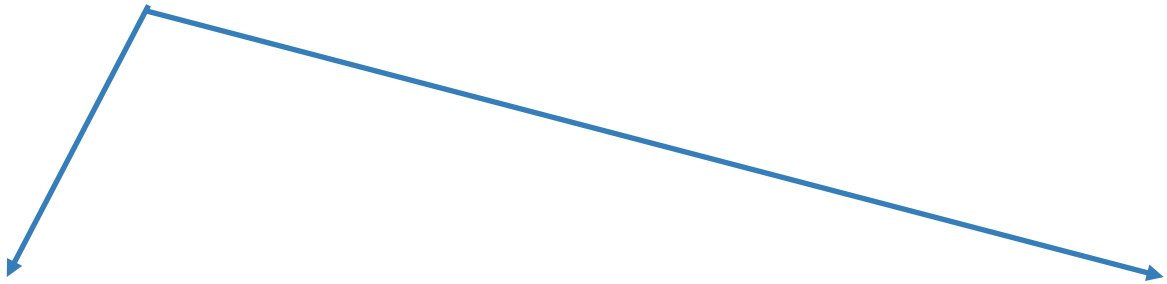
Chair

VS

- "Is it you?"
- "Are you ... ?"
- "Did you?"

- It's *you?!*?!
  (I am surprised that it is you.)

Sign Language = 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.

⟹ Because CONTEXT MATTER !!

Sign Language Recognition

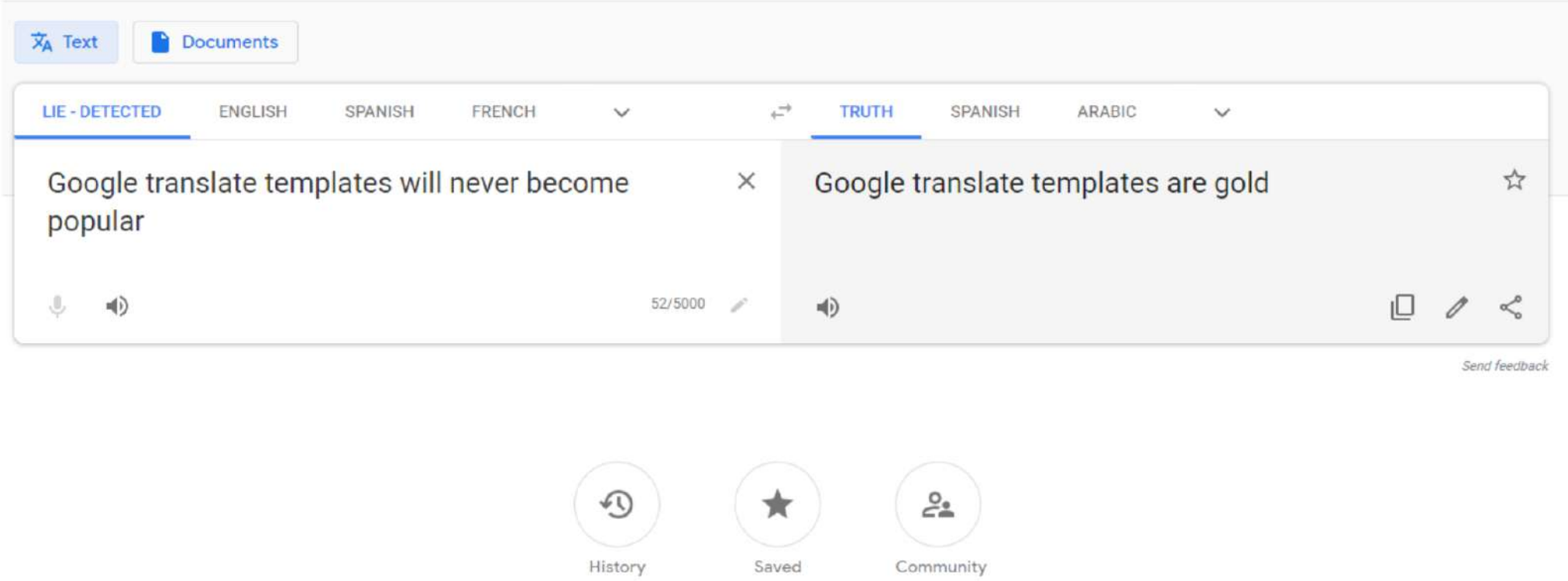
Sign Video

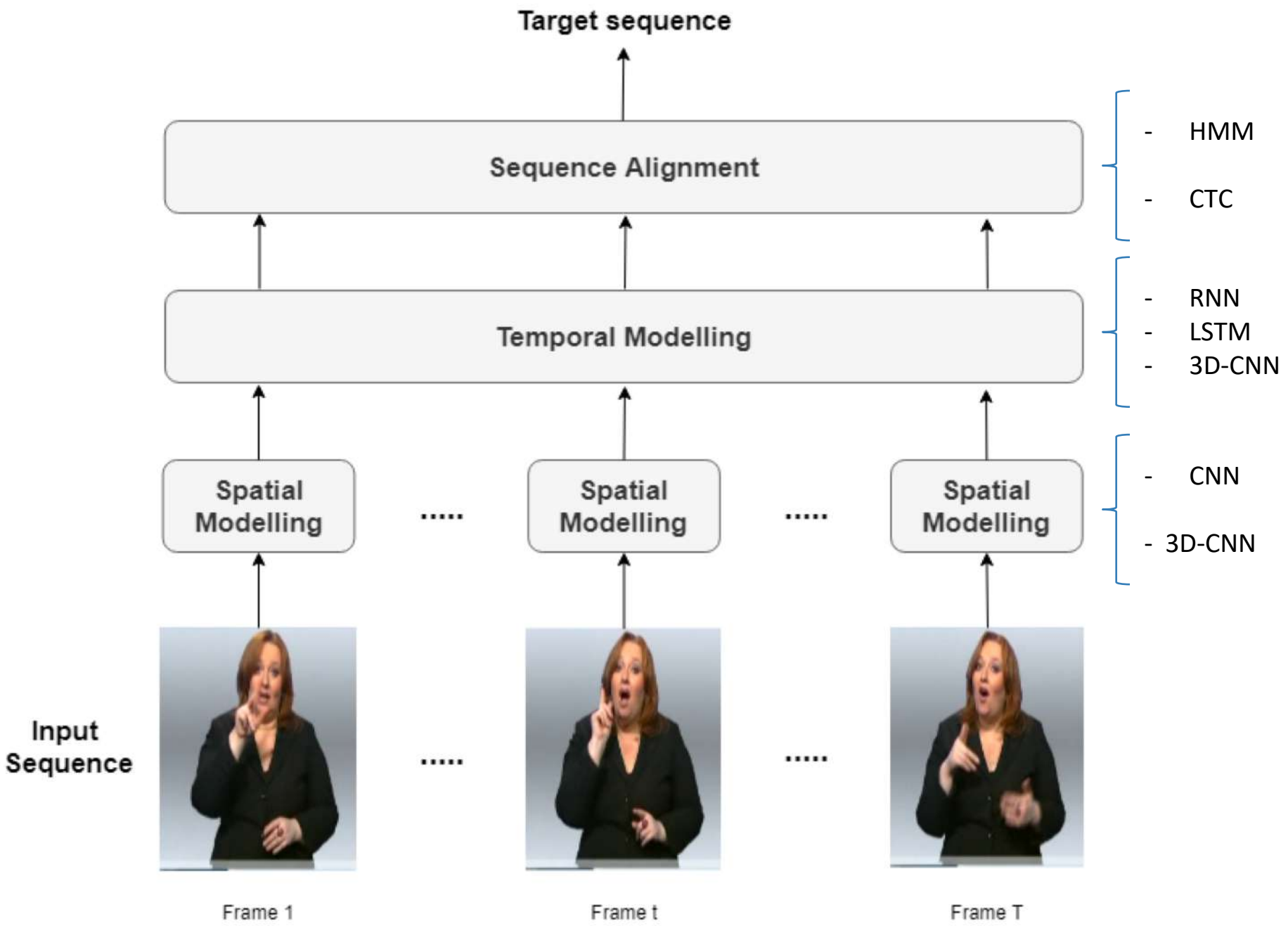
Word prediction

Recognition Model

NOW WEATHER
MORNING THURSDAY
TWELVE FEBRUARY

Sequence to Sequence problem


Text
Documents
LIE - DETECTED
ENGLISH
SPANISH
FRENCH
TRUTH
SPANISH
ARABIC
Google translate templates will never become popular
52/5000
Google translate templates are gold
Send feedback
History
Saved
Community



CTC : Connectionist Temporal Classification https://distill.pub/2017/ctc/

State of the art

Input

Output

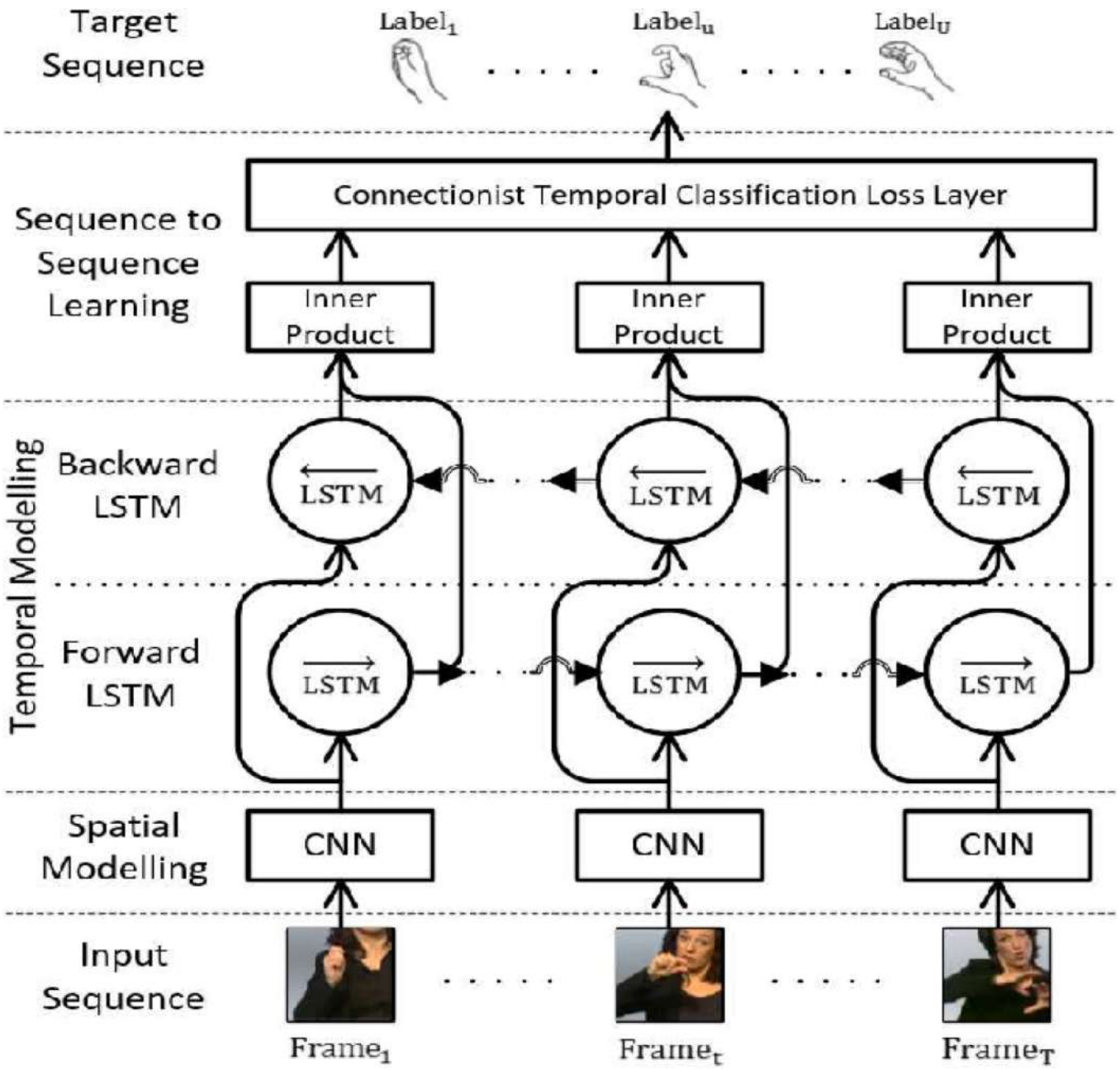
Recognition Model

Prediction

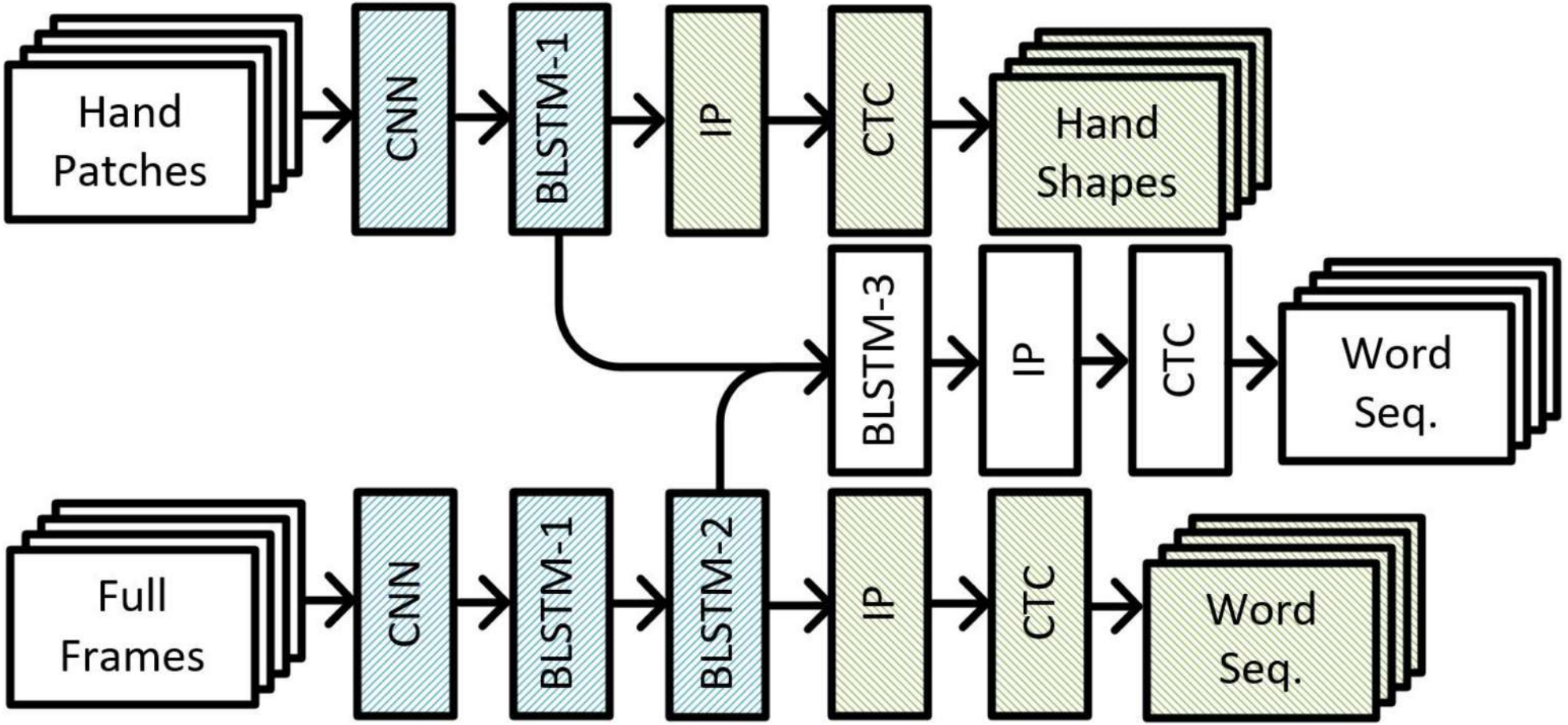
Recognition Model

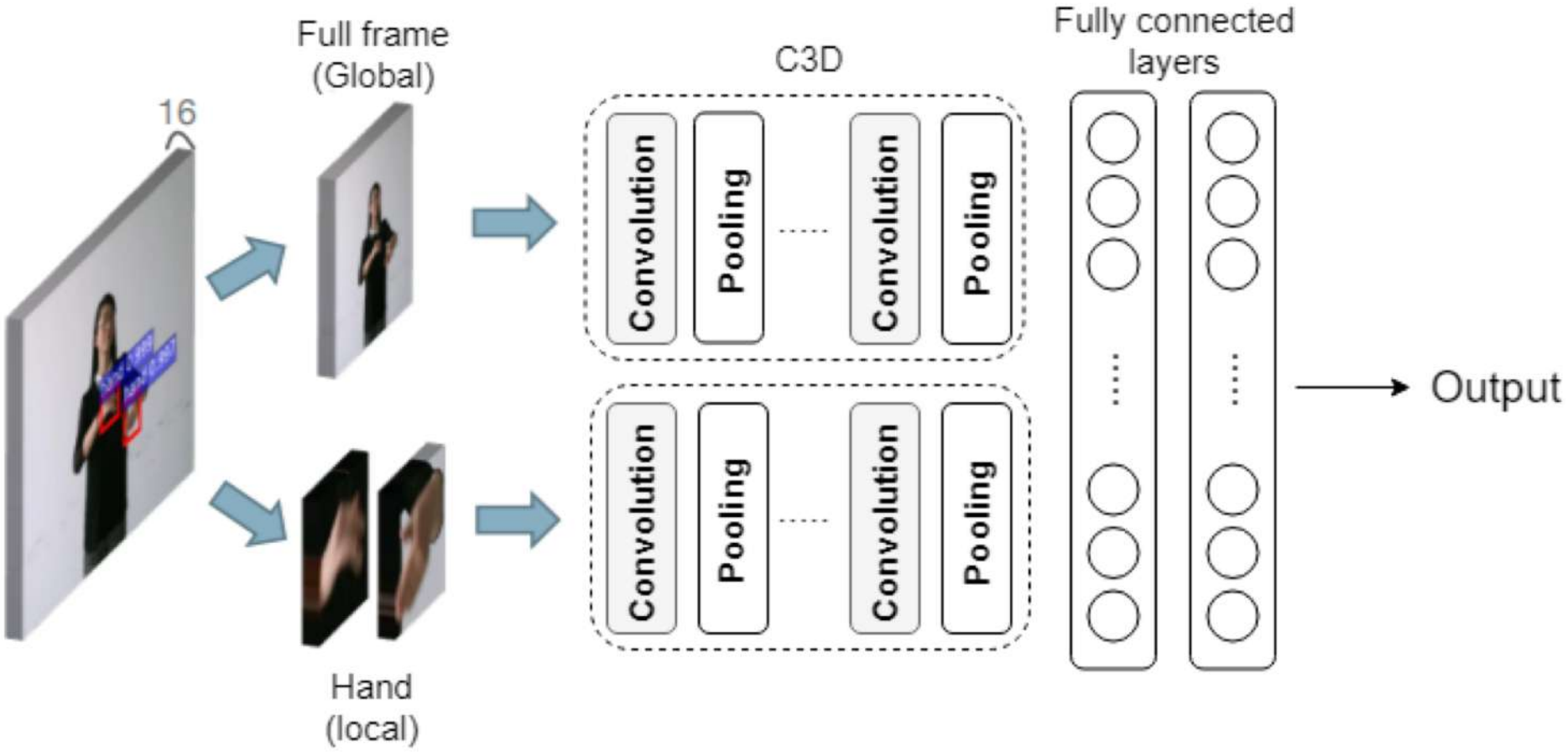
Prediction

Much of the current ignores the notion of incorporating contextual information to the handshape.



Source : SubUnets : End-to-end Hand Shape and Continous Sign Language Recognition (Camgaz et al 2017).

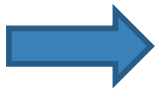

Hand Patches
CNN
BLSTM-1
IP
CTC
Hand Shapes
BLSTM-3
IP
CTC
Word Seq.
Full Frames
CNN
BLSTM-1
BLSTM-2
IP
CTC
Word Seq.



Source : Video-based Sign Language Recognition without Temporal Segmentation (Huang et al 2018).

From Sign Language Recognition To Translation

SLR is not simply a gesture recognition task.



learn the mapping of signs to their respective word predictions.

Sign Language has it is own linguistic and grammatical properties.

Word orders and grammar are important.

Source : Neural Sign Language Translation (Camgoz et al 2018)

From SLR (Sign Language Recognition) to SLT (Sign Language Translation)

- SLR seeks to recognize a sequence of continuous sign word gestures.

- SLT takes into account the different word orders and grammar.

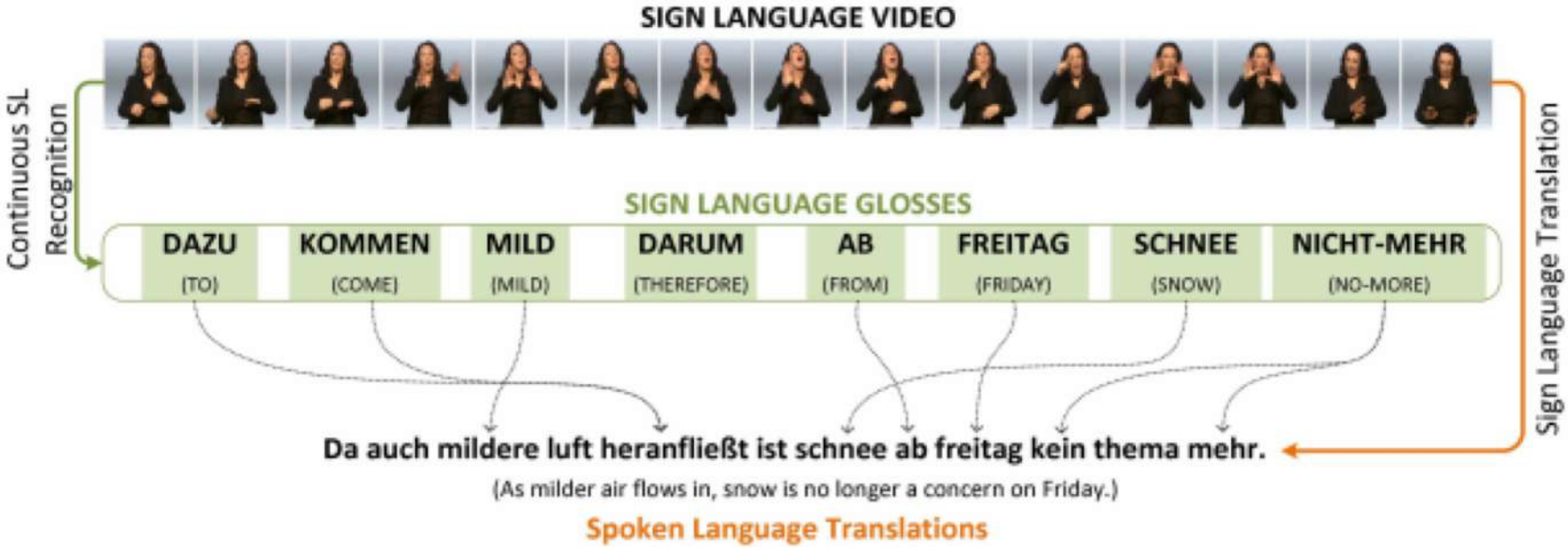


Figure 1. Difference between CSLR and SLT.

Sign2Gloss2Text (S2G2T)

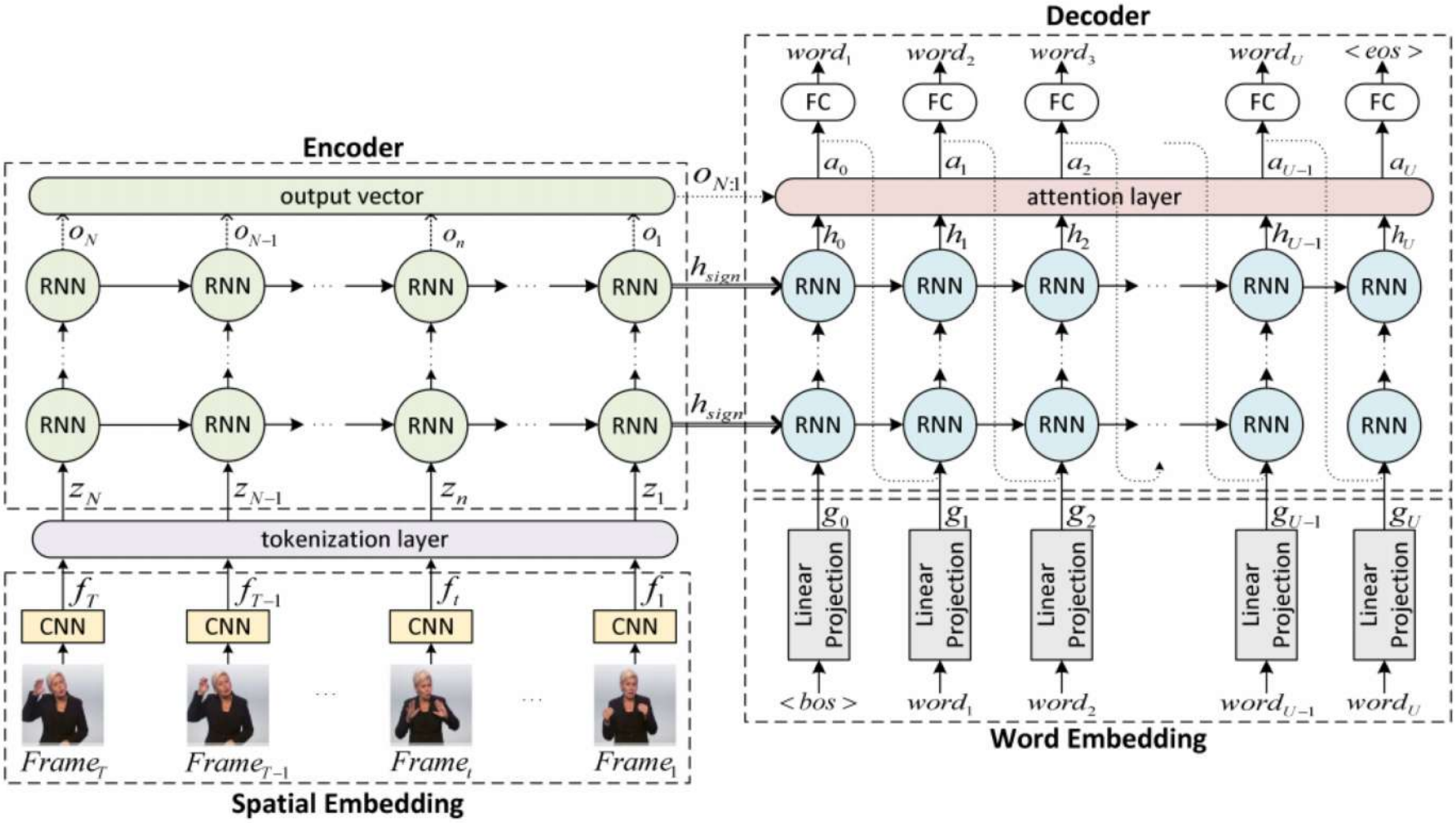
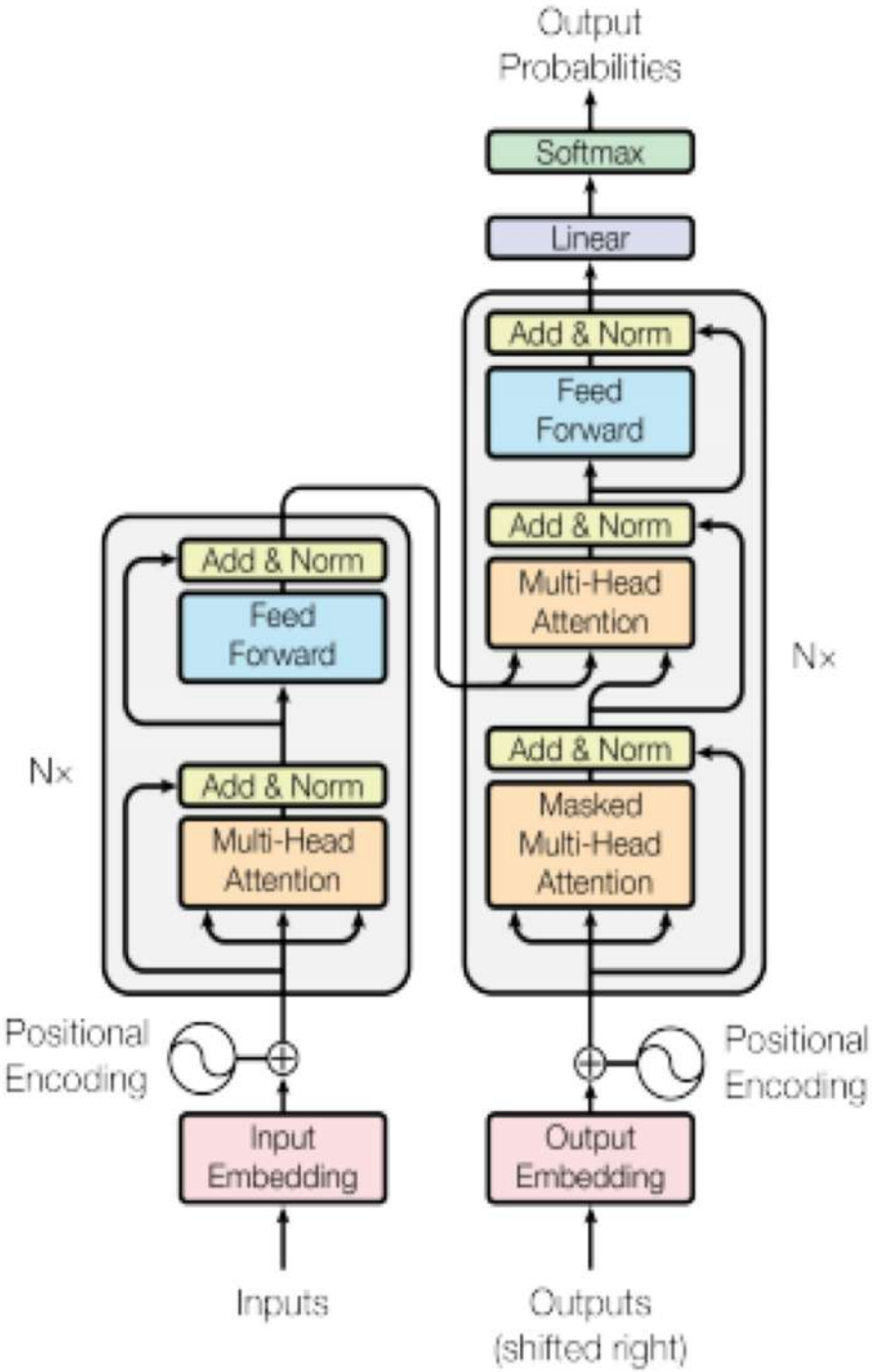


Figure 2. An overview of our SLT approach that generates spoken language translations of sign language videos.

Source : Neural Sign Language Translation (Camgoz et al 2018)

Transformer Network

Source: **Attention Is All You Need (Vaswani et al 2017)**


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

- **Word embeddings** :

Learning word representation using word2vec (linear layer), instead of doing one-hot encoding.

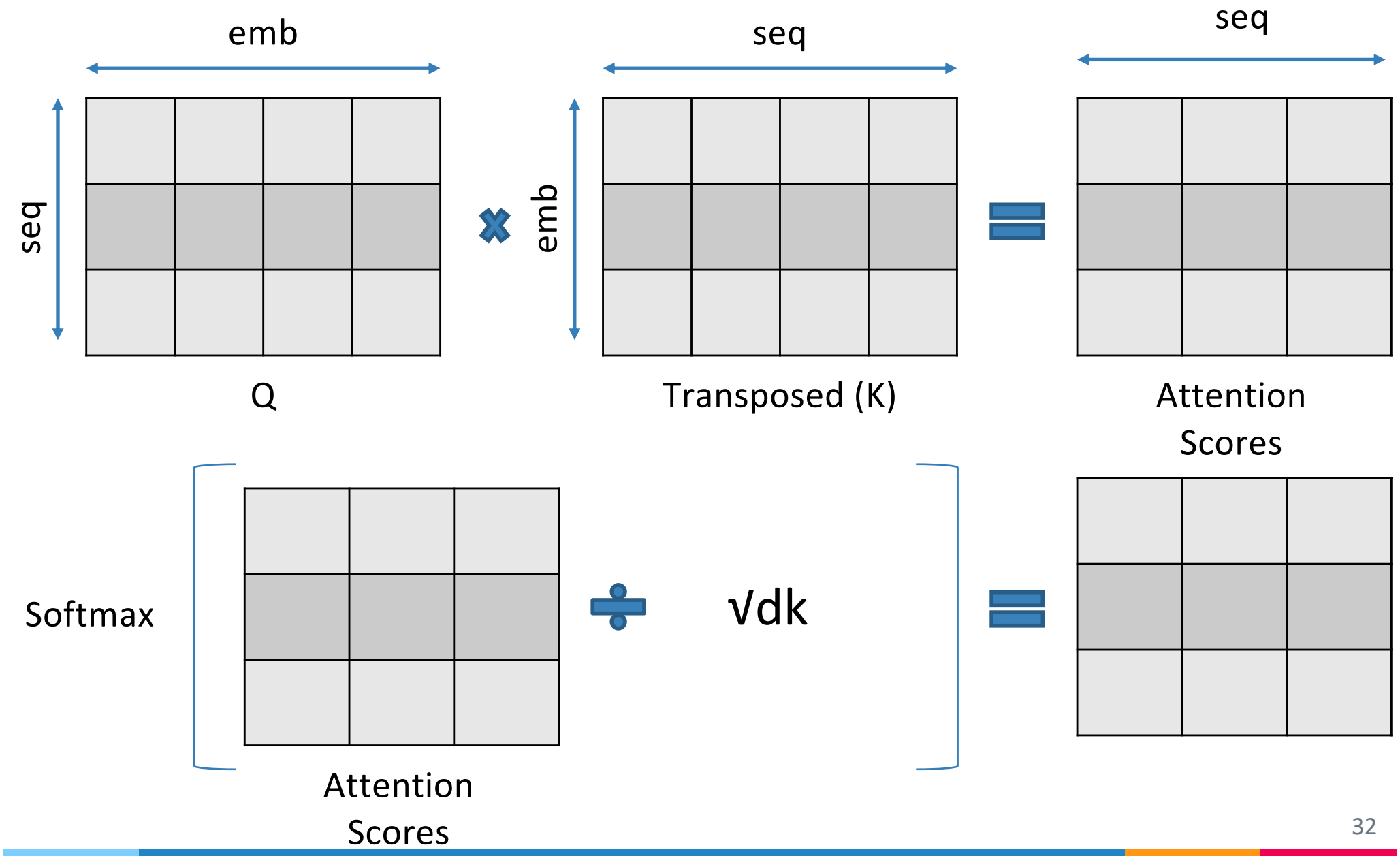
- **Self-Attention Mechanism** :

1- linearly project the input sequence to 3 identical representation (Query **Q**, Key **K** and Value **V**).

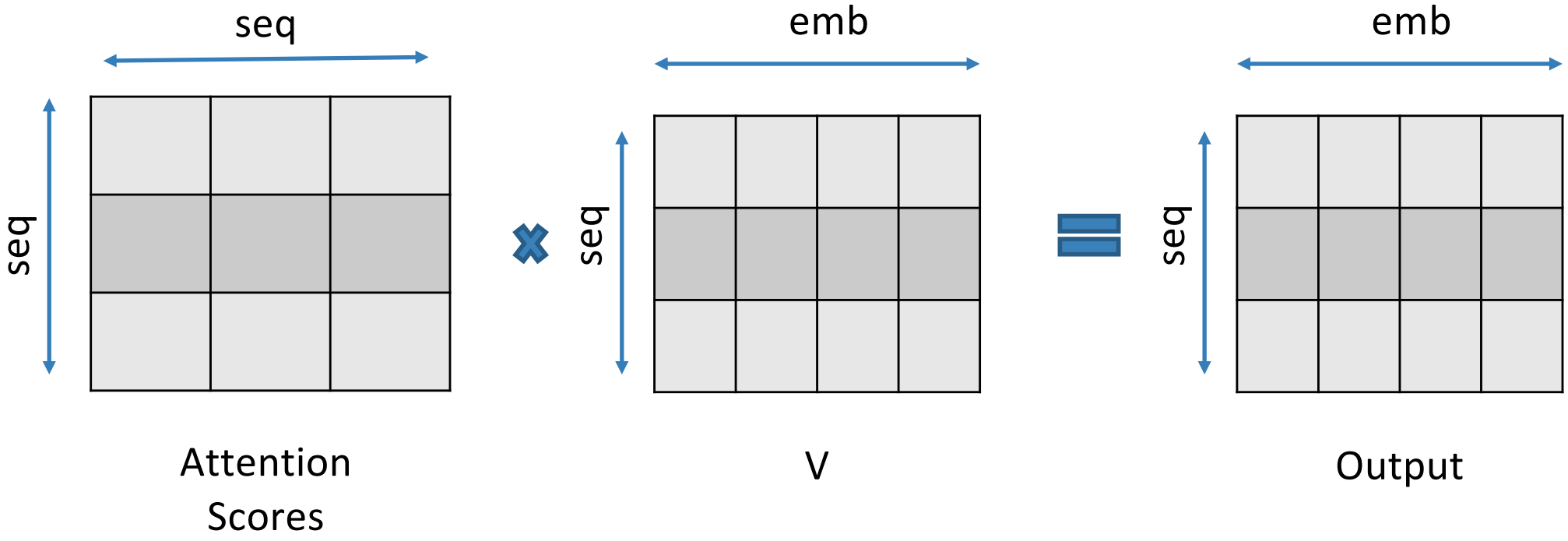
2- Scaled dot product to calculate the attention scores (similarity) of each feature with the rest of the features in the sequence by multiplying **Q** with **K**.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

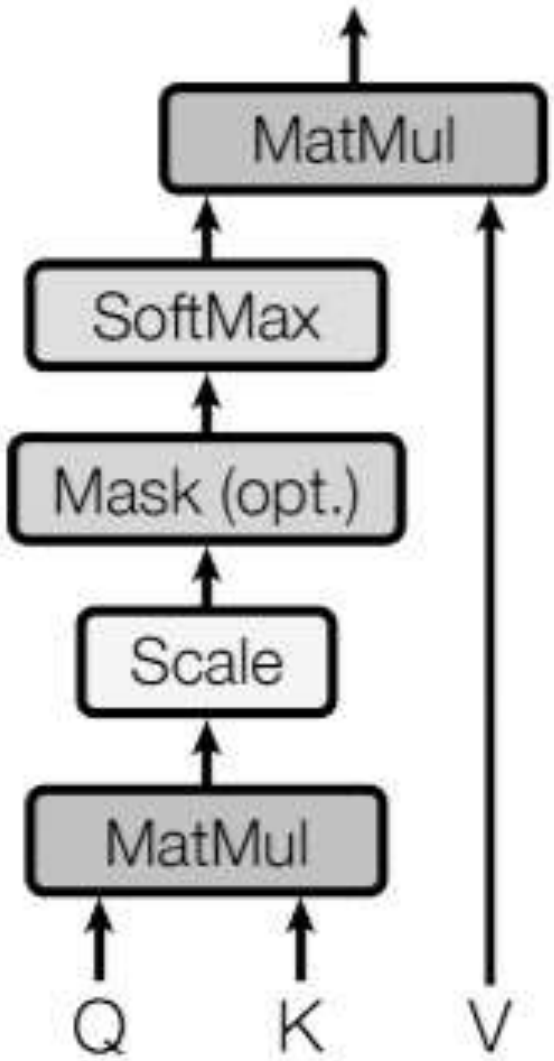
$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

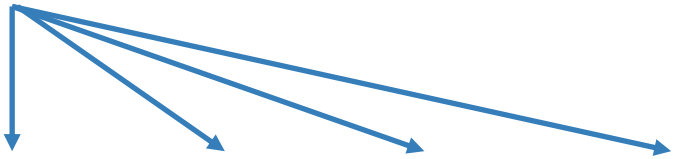


$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Query

My name is Fares



Key

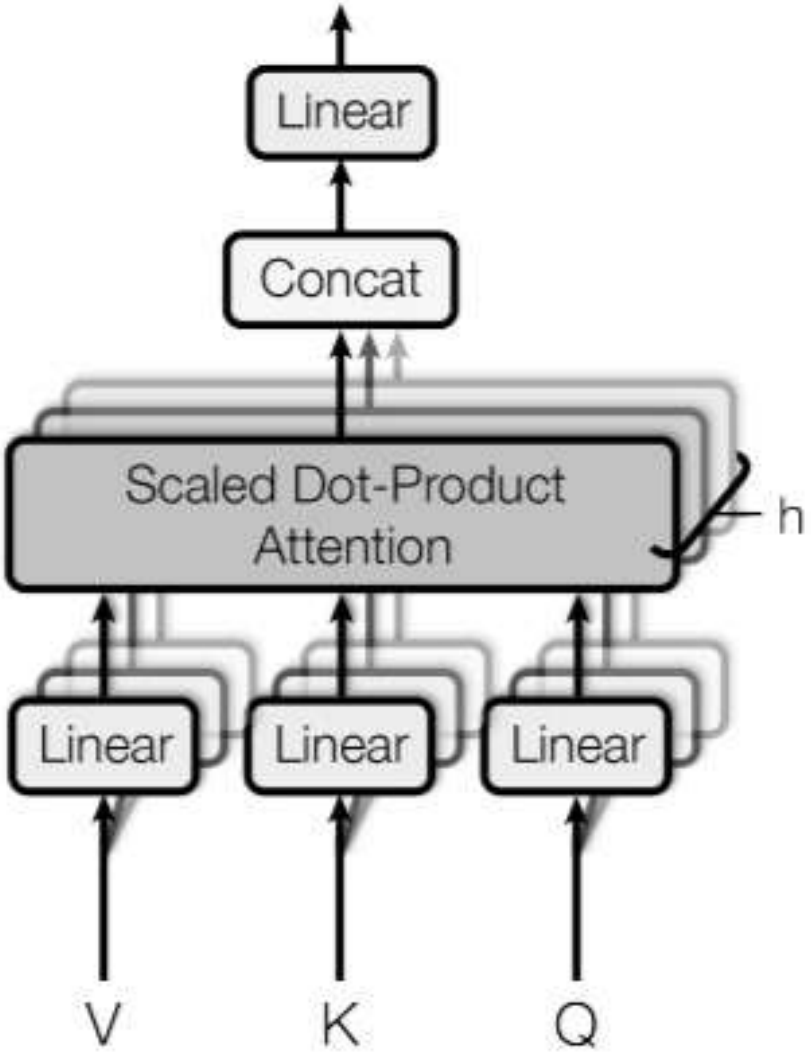
My name is Fares

Q / K	My	Name	Is	Fares
My	0.7	0.1	0.1	0.1
Name				0.6
Is				
Fares		0.6		0.9

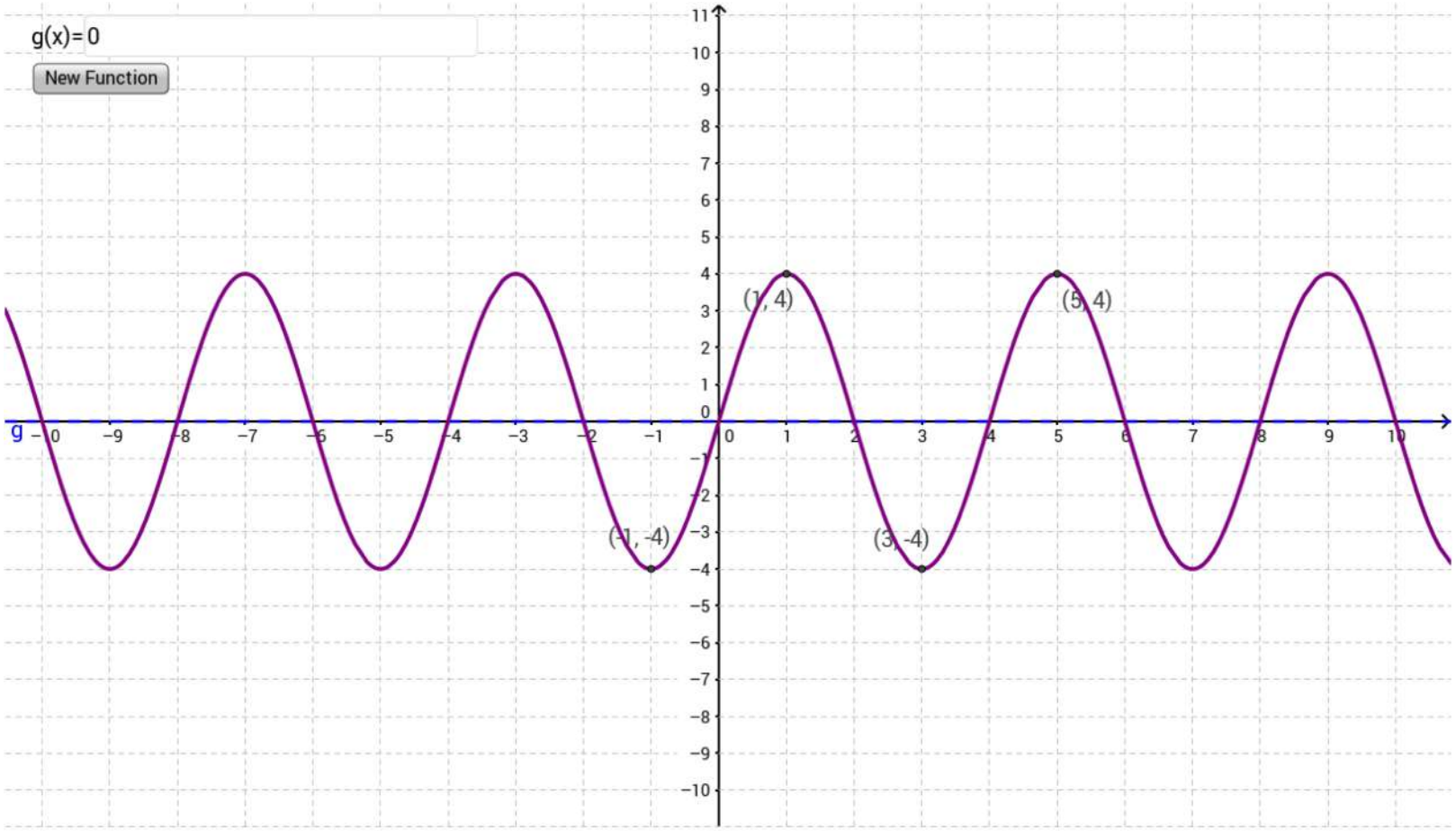
1

Attention Scores

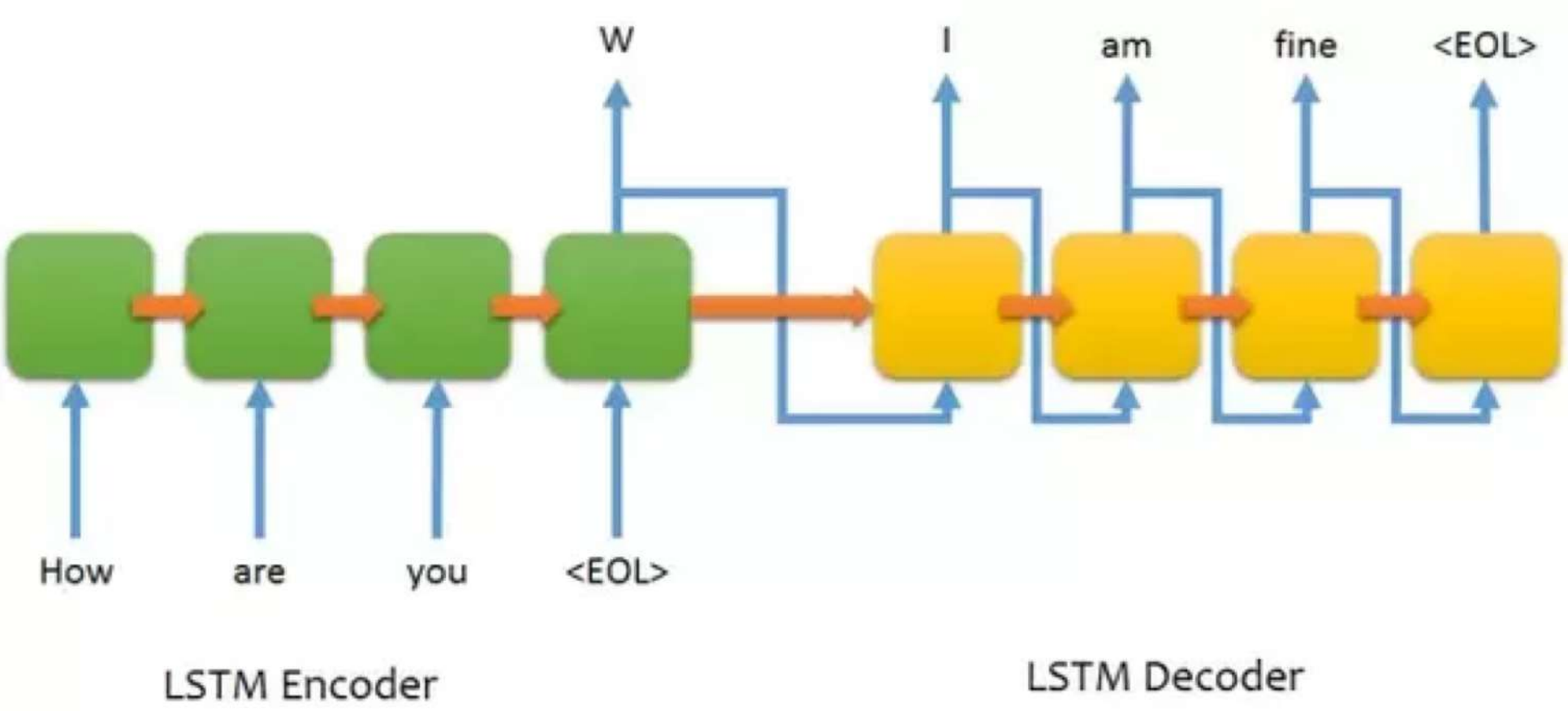
Multi-Head Attention

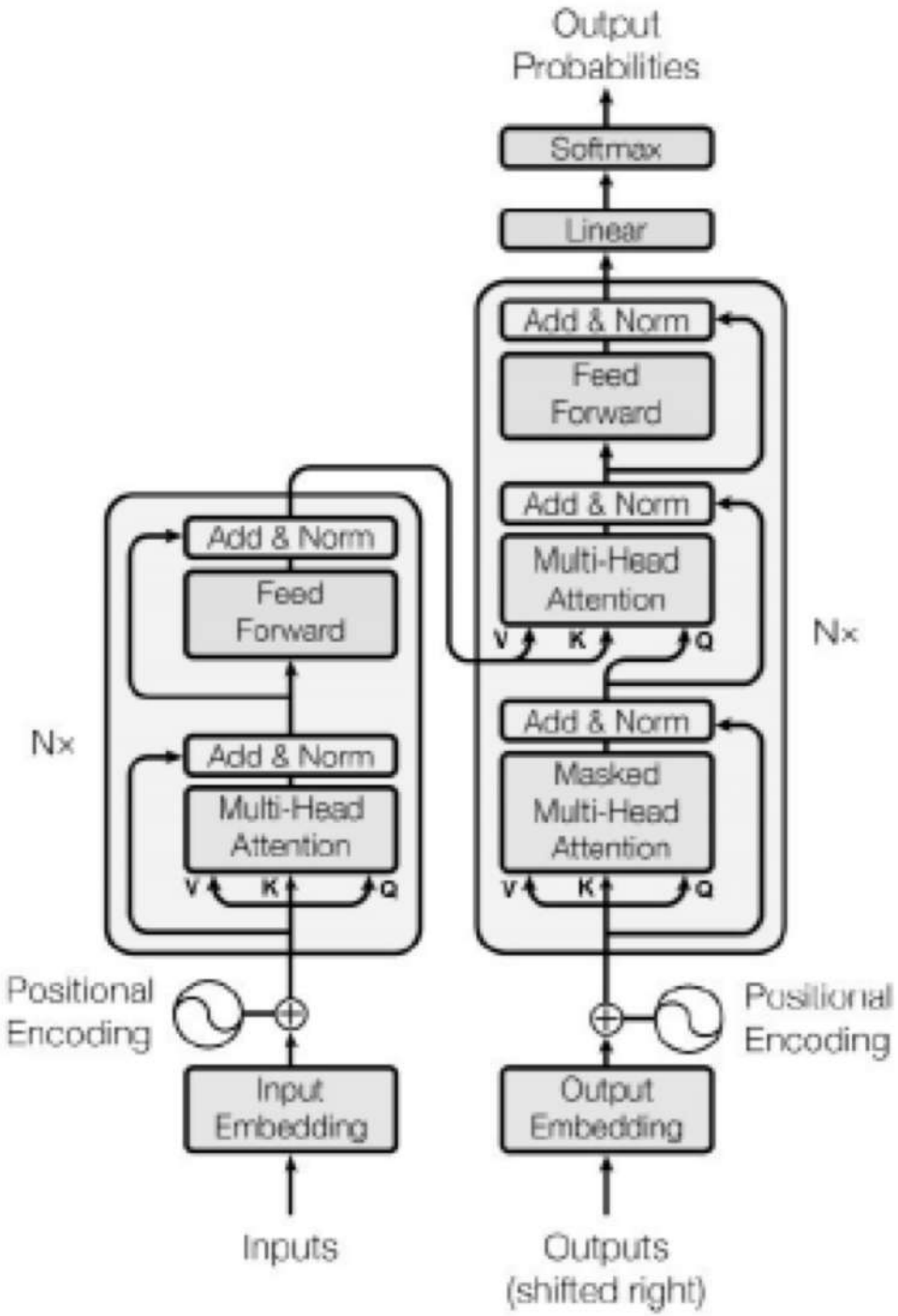


Positional encoding



Encoder-Decoder Network




Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
V
K
Q
Add & Norm
Masked Multi-Head Attention
V
K
Q
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
V
K
Q
Nx
Positional Encoding
Input Embedding
Inputs
Positional Encoding
Output Embedding
Outputs (shifted right)

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

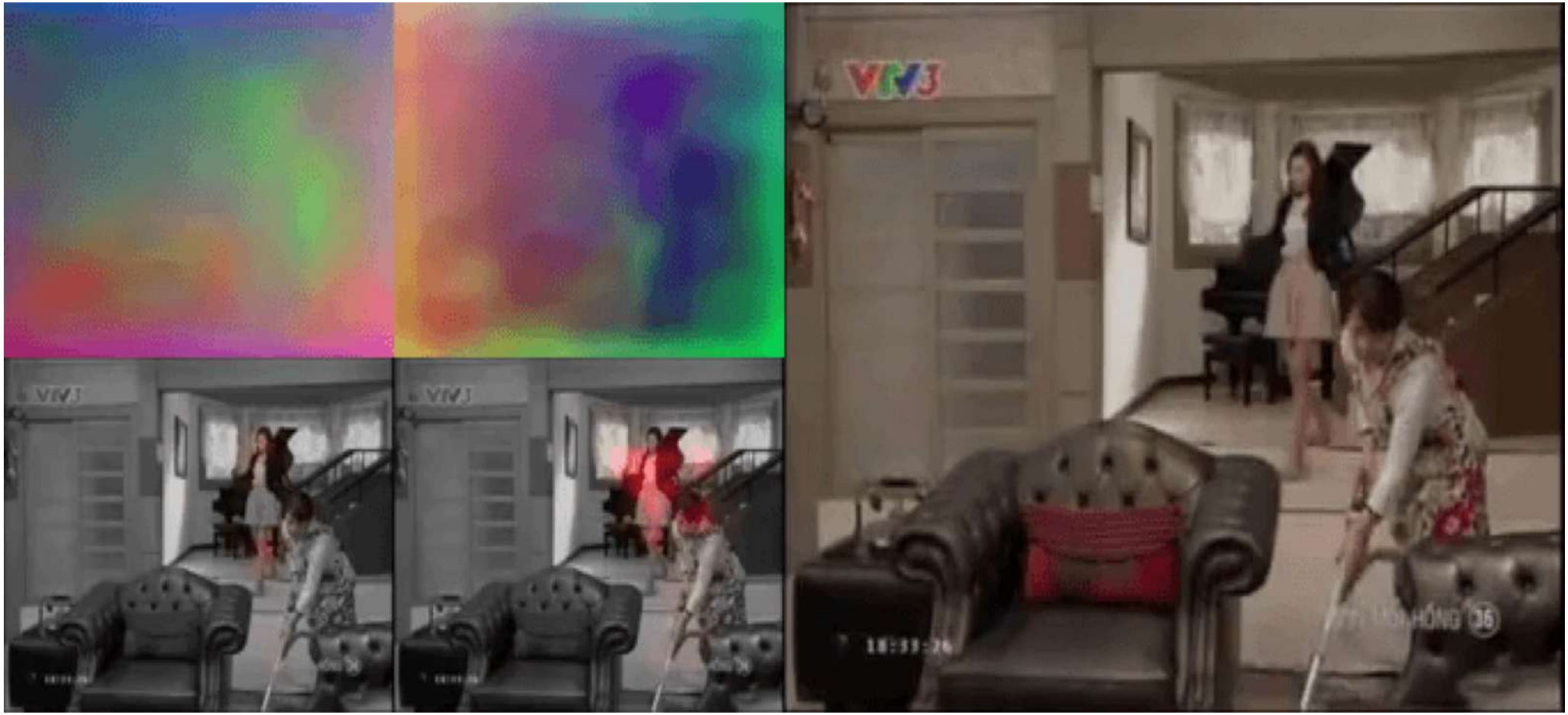
Video Action Transformer Network

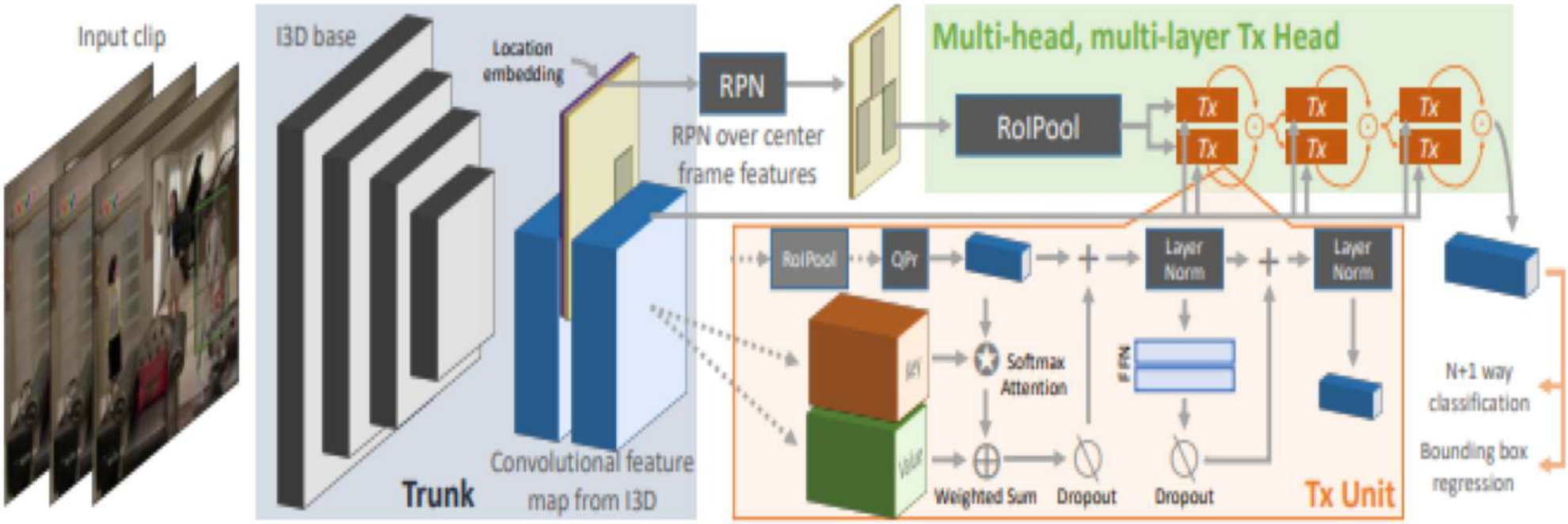
- Rohit Girdhar et al (2018)

- Use a Transformer-style architecture to aggregate features from the spatiotemporal context around the person to classify his actions.

- Human actions are recognizable from the state of the environment (scenes, objects), apart from their own (pose)

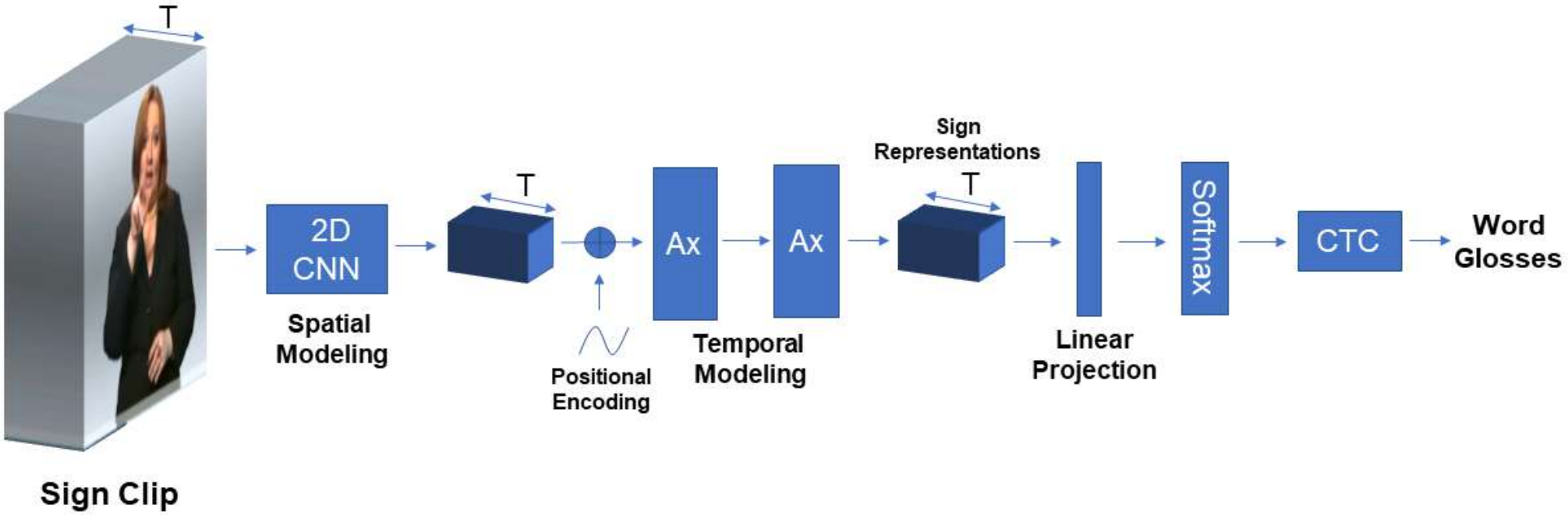
‘holding hands’ and ‘watching a person’.


VTV3

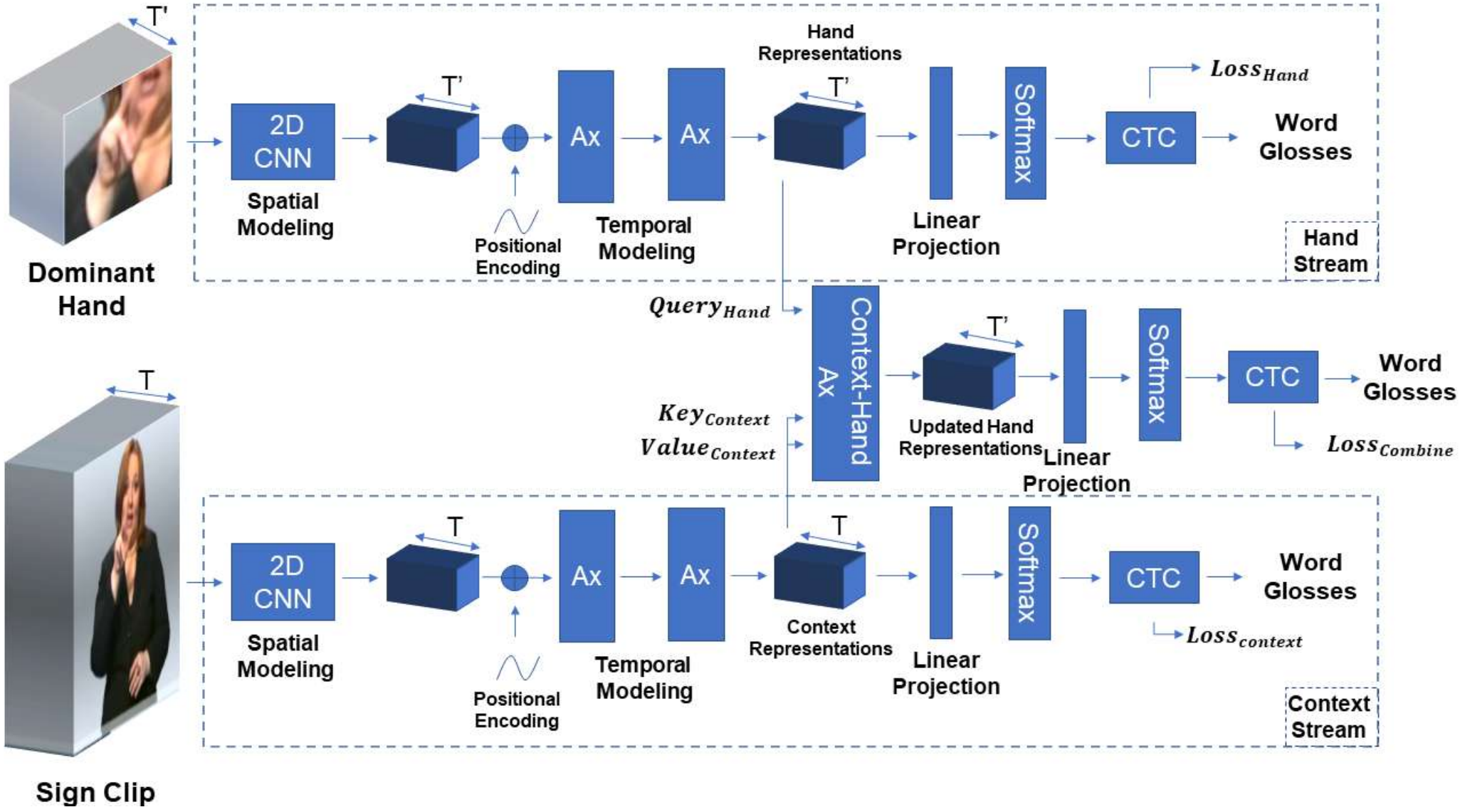

Input clip
I3D base
Location embedding
Trunk
Convolutional feature map from I3D
RPN
RPN over center frame features
Multi-head, multi-layer Tx Head
RoIPool
Tx
Tx
Tx
Tx
Tx
Tx
RoIPool
Q/K
Layer Norm
Layer Norm
Softmax Attention
FFN
Weighted Sum
Dropout
Dropout
Tx Unit
N+1 way classification
Bounding box regression

Proposed Solution

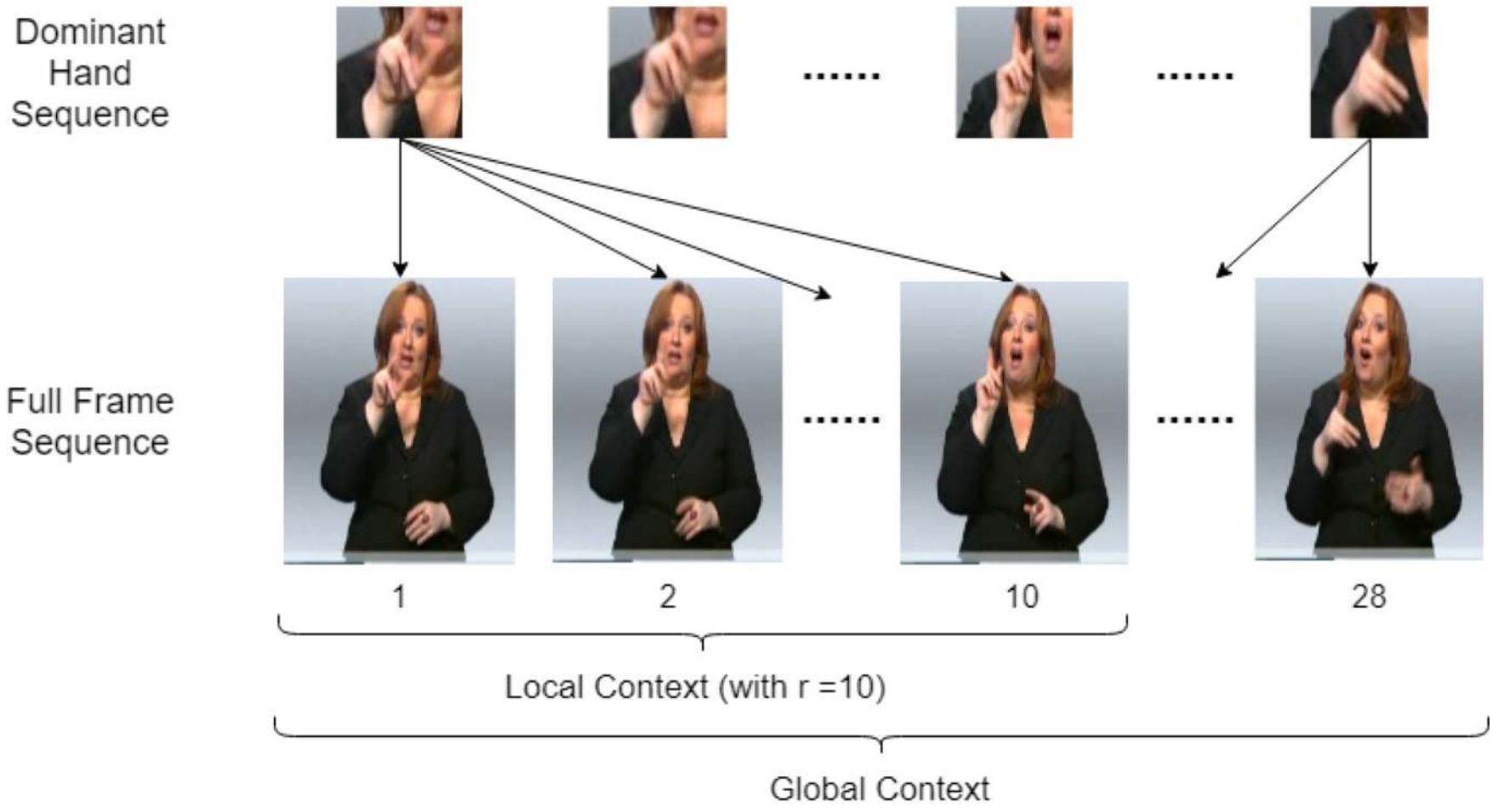
SIGN ATTENTION NETWORK



SIGN ATTENTION NETWORK
With Hand Stream



SIGN ATTENTION NETWORK
With Relative local Context masking



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.

Training Details

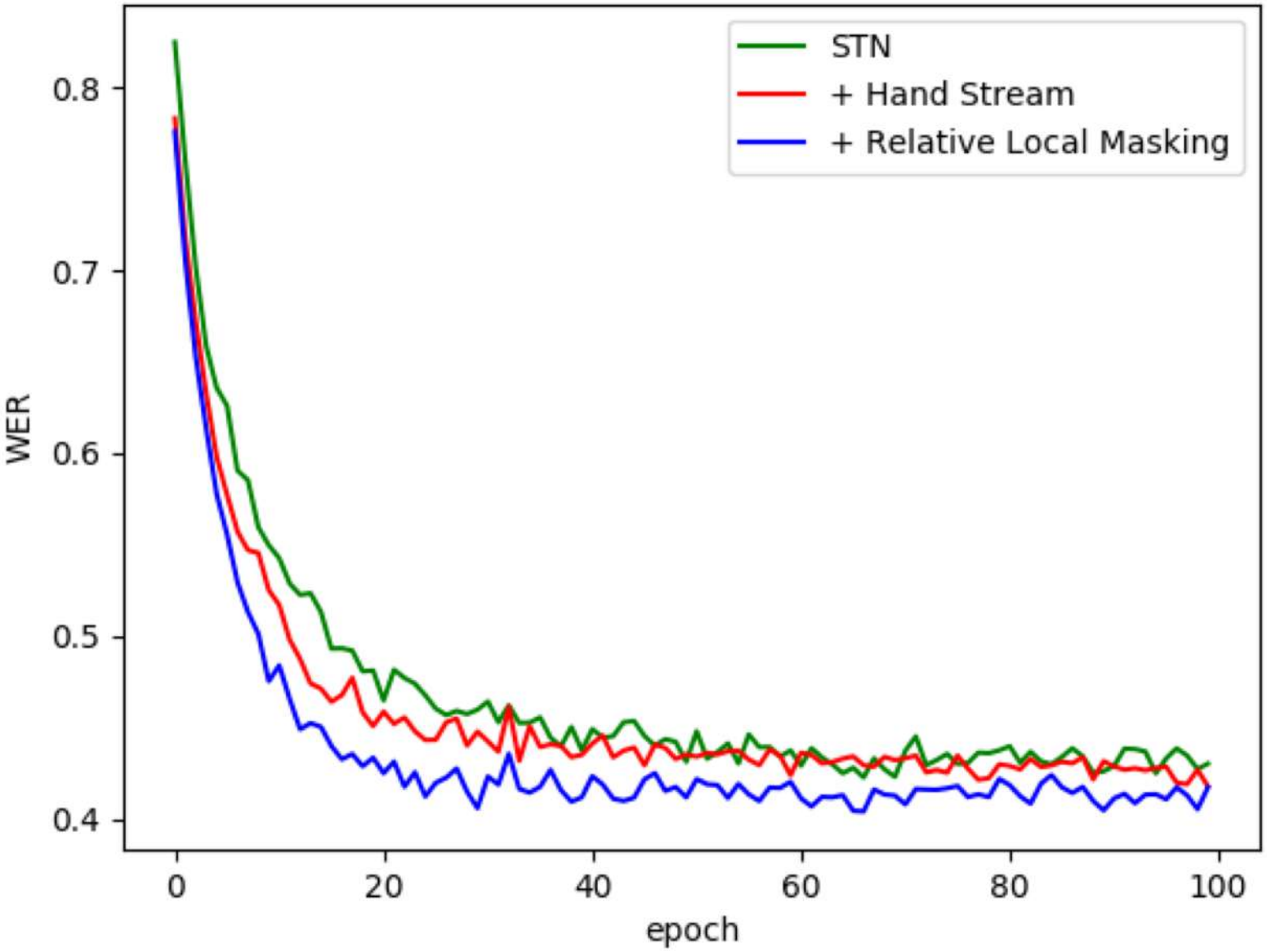
- Initialize the networks parameters using Xavier/glorot.
- Pre-train feature extraction on Imagenet.
- We use Adam optimization with default parameters.
- Data augmentation through random image x-y translation.
- Batch size of size 2.
- Padding mask to avoid looking at padded elements.

Word Error Rate (WER)

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

Comparison SAN approaches

	Dev	Test
SAN	35.33	35.45
+ Hand Stream	33.68	34.12
+ Relative Local Masking	**32.74**	**33.29**


STN
+ Hand Stream
+ Relative Local Masking
WER
0.8
0.7
0.6
0.5
0.4
0
20
40
60
80
100
epoch

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.

COMPARATIVE RESULTS ON THE RWTH-PHOENIX-WEATHER 2014 DATASET IN WER %

	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

Sign Language Translation

Neural Sign Language Translation

- Necati Cihan Camgoz, koller et al
- CVPR (2018)
- RWTH-PHOENIX-Weather 2014 T: Parallel Corpus of Sign Language Video, Gloss and Translation
- RWTH Aachen University, Germany

1. **Sign2Text (S2T)**: end-to-end pipeline translating from sign language video into spoken language.

2. **Sign2Gloss2Text (S2G2T)**: uses a SLR system as tokenization layer to add intermediate supervision.

Sign2Gloss2Text (S2G2T)

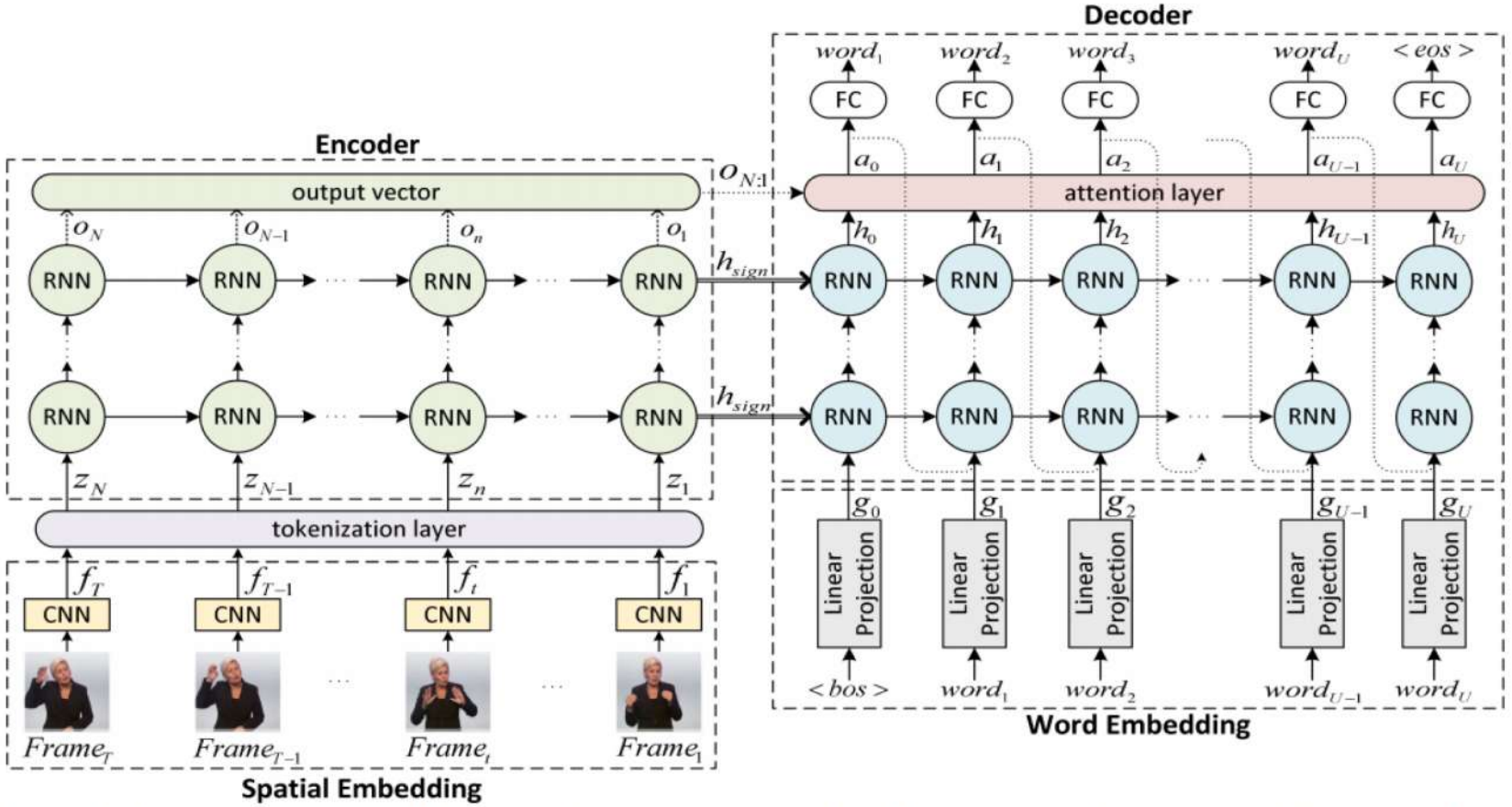


Figure 2. An overview of our SLT approach that generates spoken language translations of sign language videos.

- Prediction:

  Word glosses +
  German
  Translation

- JETZT WETTER MORGEN DONNERSTAG ZWOELF FEBRUAR
- MAINTENANT MÉTÉO DEMAIN JEUDI DOUZE FÉVRIER (Word glosses)
- et maintenant les prévisions météo pour demain jeudi le 12 août (Translation)

RWTH-PHOENIX-Weather 2014 T: Parallel Corpus of Sign Language Video, Gloss and Translation

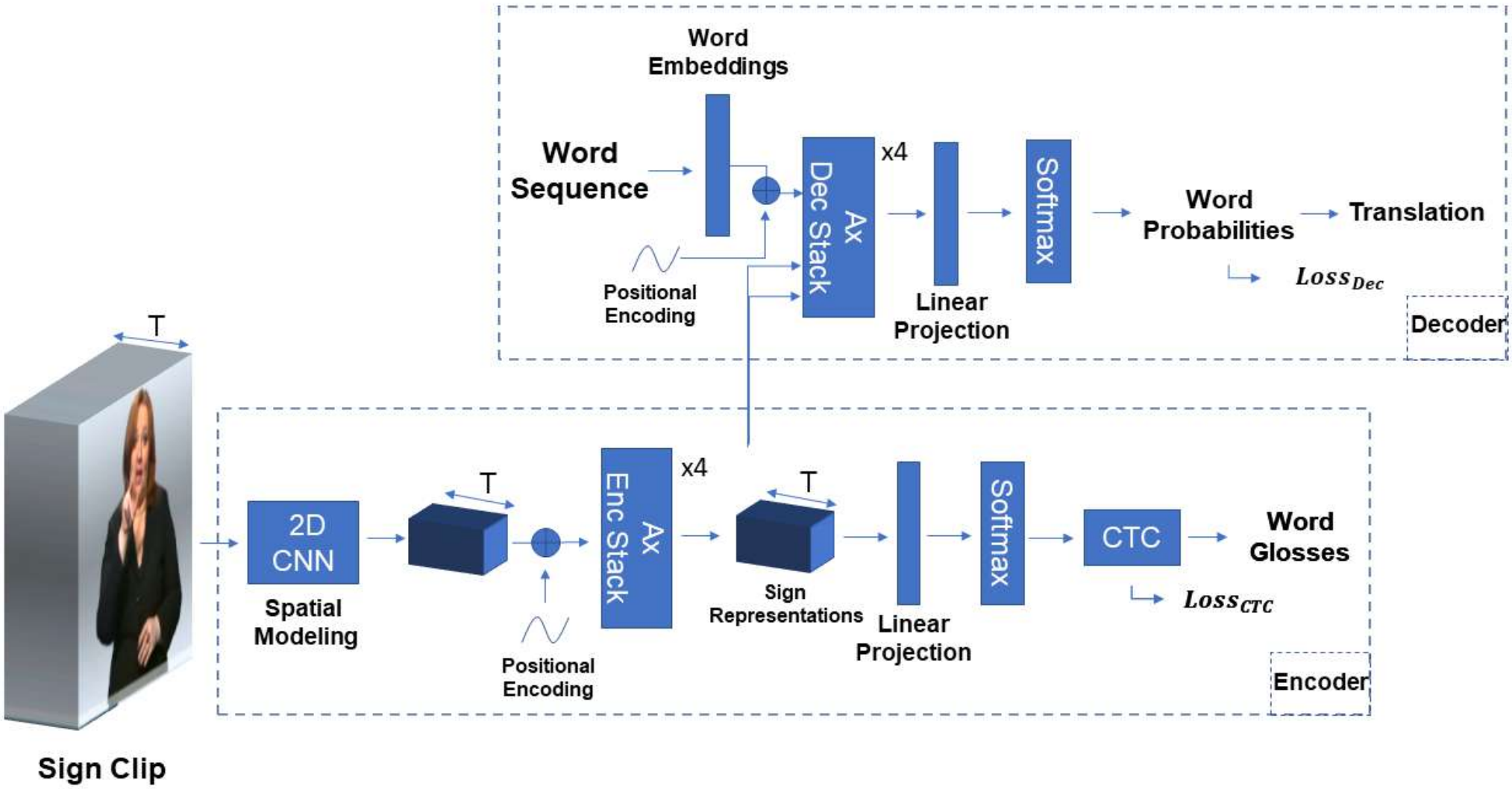
	DEV SET					TEST SET				
Approach:	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4
G2T	46.02	44.40	31.83	24.61	20.16	45.45	44.13	31.47	23.89	19.26
S2T	31.80	31.87	19.11	13.16	9.94	31.80	32.24	19.03	12.83	9.58
S2G→G2T	43.76	41.08	29.10	22.16	17.86	43.45	41.54	29.52	22.24	17.79
S2G2T	**44.14**	**42.88**	**30.30**	**23.02**	**18.40**	**43.80**	**43.29**	**30.39**	**22.82**	**18.13**

BLEU (bilingual evaluation understudy)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Metrics used for evaluating text quality.

SAN (encoder-decoder)



COMPARATIVE RESULTS ON THE RWTH-PHOENIX-WEATHER 2014 DATASET using Bleu and Rouge scores % (the higher the better).

	Dev					Test				
	Rouge	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge	Bleu-1	Bleu-2	Bleu-3	Bleu-4
STN (Encoder-Decoder)	**35.29**	**35.49**	**22.10**	**15.58**	**12.14**	**34.53**	**35.59**	**22.15**	**15.70**	**12.18**
S2T [5]	31.80	31.87	19.11	13.16	9.94	31.80	32.24	19.03	12.83	9.58
Hybrid STN	41.27	39.74	27.41	20.61	16.54	41.18	40.82	28.20	21.16	16.93
S2G2T [5]	**44.14**	**42.88**	**30.30**	**23.02**	**18.40**	**43.80**	**43.29**	**30.39**	**22.82**	**18.13**

NOTE:

1- S2G2T is pretrained on SLR.

2- Metrics like Bleu or Rouge can be sometimes misleading with regard to overall accuracy.

Approach	Translation
Ground Truth	die neue woche beginnt wechselhaft und kuhler.
	(the new week starts unpredictable and cooler.)
STN (ours)	die neue woche startet wechselhaft und wieder kühler .
	(the new week starts unpredictable and cooler again.)
Hybrid STN (ours)	die neue woche beginnt wechselhaft und wieder kühler .
	(the new week starts unpredictable and cooler again.)
S2T from (Camgoz et al., 2018)	am montag uberall wechselhaft und kuhler.
	(on monday everywhere unpredictable and cooler.)
S2G2T from (Camgoz et al., 2018)	die neue woche beginnt wechselhaft und wechselhaft.
	(the new week starts unpredictable and unpredictable.)
Ground Truth	im suden und sudwesten gebietsweise regen sonst recht freundlich.
	(in the south and southwest locally rain otherwise quite friendly.)
STN (ours)	am sonntag gebietsweise regen im bergland ist es sehr windig.
	(It is very windy on Sunday in the mountainous regions.)
Hybrid STN (ours)	im südwesten regnet es zum teil kräftig im osten sonst ist es freundlich.
	(in the southwest it is raining heavily in the east, otherwise it is friendly.)
S2T from (Camgoz et al., 2018)	von der sudhalfte beginnt es vielerorts.
	(from the southpart it starts in many places.)
S2G2T from (Camgoz et al., 2018)	am freundlichsten wird es im suden.
	(the friendliest it will be in the south.)
Ground Truth	und nun die wettervorhersage für morgen samstag den zweiten april .
	(and now the weatherforecast for tomorrow saturday the second april.)
STN (ours)	und nun die wettervorhersage für morgen samstag den fünften märz .
	(and now the weather forecast for tomorrow saturday the fifth of march.)
Hybrid STN (ours)	und nun die wettervorhersage für morgen samstag den zweiten märz.
	(and now the weather forecast for tomorrow saturday the second of march.)
S2T from (Camgoz et al., 2018)	und nun die wettervorhersage fur morgen freitag den sechsundzwanzigsten märz.
	(and now the weatherforecast for tomorrow friday the twentysixth march.)
S2G2T from (Camgoz et al., 2018)	und nun die wettervorhersage fur morgen samstag den siebzehnten april .
	(and now the weatherforecast for tomorrow saturday the seventeenth april.)

Thank you !!!

Réferences:

- http://lifeprint.com/asl101/gifs/t/thank-you.gif
- https://www.lifeprint.com/asl101/pages-layout/facialexpressions.htm
- https://i0.wp.com/www.reactiongifs.com/r/facepalm.gif
- https://rohitgirdhar.github.io/ActionTransformer/

Articles:

- Video Action Transformer Network
- Neural Sign Language Translation