

Sign Language Translation

Mahdi Ranjbar

Deep Learning

January 16, 2024

Introduction

Sign Language Translation (SLT) involves the complex process of converting continuous sign language videos into corresponding spoken language sentences. This multimodal challenge demands not only an accurate interpretation of the signer's movements and expressions but also the creation of an appropriate textual translation. Despite significant advancements, the journey towards a fully effective automatic SLT system remains ongoing, highlighting the complexity of this task.

The impact of mastering SLT cannot be overstated. It holds the potential to significantly enhance communication between individuals who use sign language and those who do not, bridging a vital gap in interpersonal interactions.

Recent developments in SLT echo trends seen in broader areas of computer vision and natural language processing. The approach predominantly involves training deep neural networks on extensive datasets. However, a major hurdle in this field is the scarcity of publicly available sign language datasets, particularly those that offer a parallel corpus—videos paired with their textual translations. Such datasets are crucial for benchmarking and advancing the current state-of-the-art in SLT.

SLT Pipeline

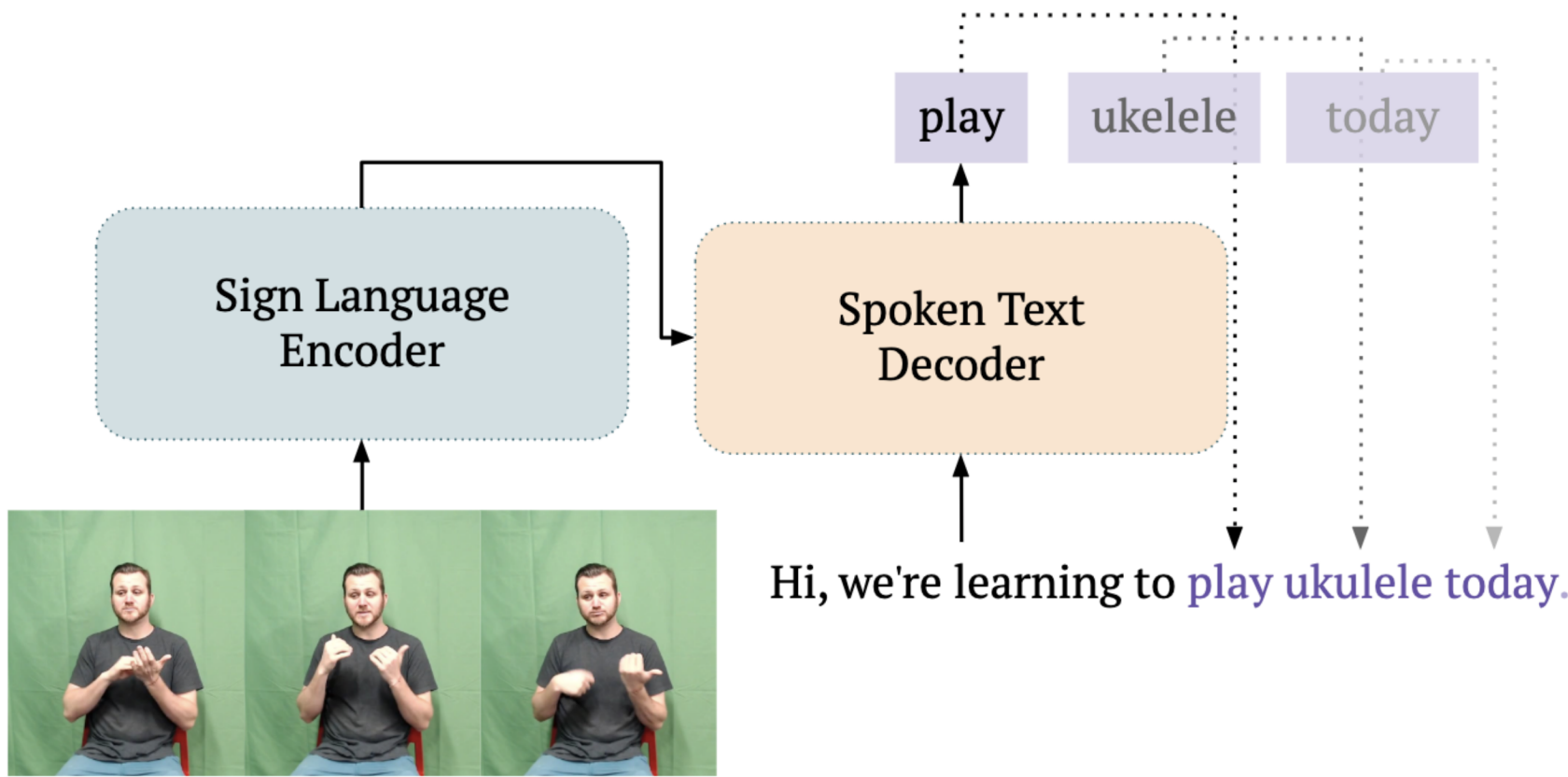


Figure 1: A basic pipeline for sign language translation.

Dataset

How2Sign [2], a multimodal and multi-view continuous American Sign Language (ASL) dataset, consisting of a parallel corpus of more than 80 hours of sign language videos and a set of corresponding modalities including speech, English transcripts, and depth.



Figure 2: A basic pipeline for sign language translation.

Methodology

The input video stream is tokenized with a pre-trained I3D feature extractor [1]. These tokens are fed into the encoding layers of the Transformer. The decoder of the Transformer operates with lowercase and tokenized textual representations.

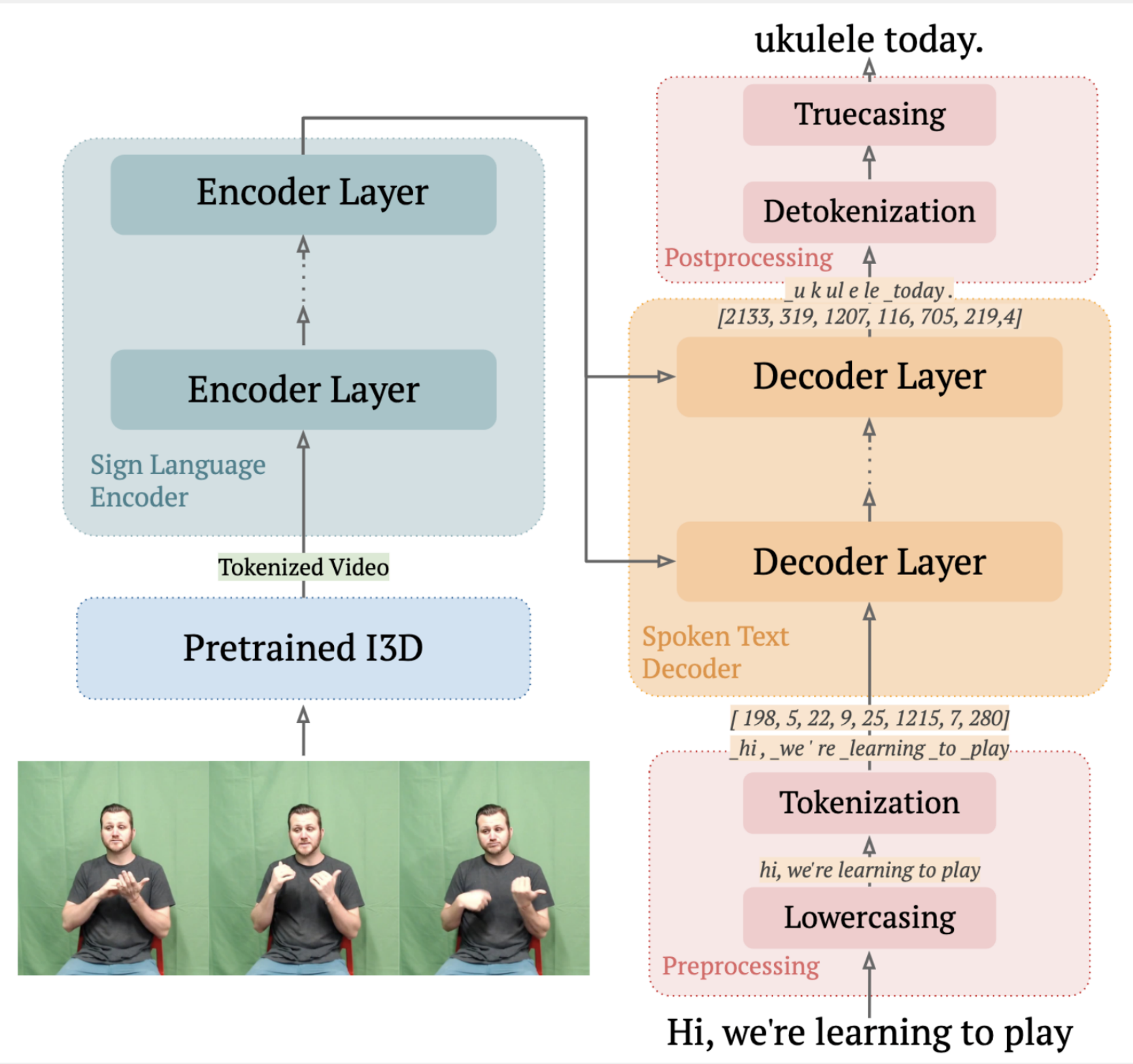


Figure 3: The input video sequence is fed into a Transformer to generate the output text sequence.

Metric

The BLEU (Bilingual Evaluation Understudy) metric is widely used in the evaluation of machine translation quality. It measures the correspondence between a machine's output and that of a human. The key idea of BLEU is to compare the n-gram of the translated text with the n-gram of the reference text, typically a human translation, to determine quality.

The BLEU score is calculated as follows:

$$\text{BLEU} = \min\left(1, \frac{\text{output length}}{\text{reference length}}\right) \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (1)$$

where p_n is the modified n-gram precision, w_n is the weight for each n-gram (usually uniform), and N is the size of the n-gram.

Additionally, an alternative metric to BLEU is reducedBLEU(rBLEU). This metric consists of removing certain words from the reference and the prediction before computing the BLEU score. There is a blacklist of words that are frequently used in the training data but do not contribute much to the meaning of the sentences, such as articles, prepositions, and pronouns.

Proposed Work


ID			rBLEU	BLEU
(1)	Input Reference Prediction	 And that's a great vital point technique for women's self defense . It's really a great point for women's self defense .	30.29	38.25
(2)	Input Reference Prediction	 In this clip I'm going to show you how to tape your cables down. In this clip I'm going to show you how to improve push ups .	24.88	64.53
(3)	Input Reference Ours	 In this segment we're going to talk about how to load your still for distillation of lavender essential oil . — Ok , in this clip , we're going to talk about how to fold the ink for the lid of the oil .	6.77	29.82
(4)	Input Reference Ours.	 You are dancing , and now you are going to need the veil and you are going to just grab the veil as far as possible . — So, once you're belly dancing , once you've got to have the strap , you're going to need to grab the thumb , and try to avoid it.	4.93	8.04
(5)	Input Reference Ours	 But if you have to setup a new campfire , there's two ways to do it in a very low impact ; one is with a mound fire , which we should in the campfire segment earlier and the other way to setup a low impact campfire is to have a fire pan , which is just a steel pan like the top of a trash can . — And other thing I'm going to talk to you is a little bit more space , a space that's what it's going to do, it's kind of a quick , and then I don't want to take a spray skirt off, and then I don't want it to take it to the top of it.	0.85	3.79
(6)	Input Reference Ours	 So, this is a very important part of the process . It's a very important part of the process .	0	61.86

Figure 4: Qualitative examples from the proposed best-performing model. In bold the words remaining to compute rBLEU. Together with selected frames from the input video.

Reflection

- Investigate the impact of video features using SWIN features instead of I3D features
- Analyse both quantitatively and qualitatively other translation metrics such as ROUGE, BERT-Score, etc. and state their advantages as well as their limits for the SLT task
- No text augmentation is used during training in the proposed work. One way to perform text augmentation is to add a rephrasing module at training time.

References

- [1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299--6308, 2017.
- [2] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735--2744, 2021.