Is context all you need? Scaling Neural Sign Language Translation to Large Domains of Discourse

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, Richard Bowden

Sign Language Translation (SLT) is a challenging task that aims to generate spoken language sentences from sign language videos, both of which have different grammar and word/gloss order. From a Neural Machine Translation (NMT) perspective, the straightforward way of training translation models is to use sign language phrase-spoken language sentence pairs. However, human interpreters heavily rely on the context to understand the conveyed information, especially for sign language interpretation, where the vocabulary size may be significantly smaller than their spoken language equivalent.

Taking direct inspiration from how humans translate, we propose a novel multi-modal transformer architecture that tackles the translation task in a context-aware manner, as a human would. We use the context from previous sequences and confident predictions to disambiguate weaker visual cues. To achieve this we use complementary transformer encoders, namely: (1) A Video Encoder, that captures the low-level video features at the frame-level, (2) A Spotting Encoder, that models the recognized sign glosses in the video, and (3) A Context Encoder, which captures the context of the preceding sign sequences. We combine the information coming from these encoders in a final transformer decoder to generate spoken language translations.

We evaluate our approach on the recently published large-scale BOBSL dataset, which contains ~1.2M sequences, and on the SRF dataset, which was part of the WMT-SLT 2022 challenge. We report significant improvements on state-of-the-art translation performance using contextual information, nearly doubling the reported BLEU-4 scores of baseline approaches.