Advanced Machine Learning

Final Project Presentation — MobileViTs for Sign Language Recognition

Dario Loi (1940849)
Alessio Olivieri (1973323)
Alessandro Monteleone (1883922)

M.Sc. in Computer Science,
Sapienza, University of Rome.

A. Y. 2024 – 2025

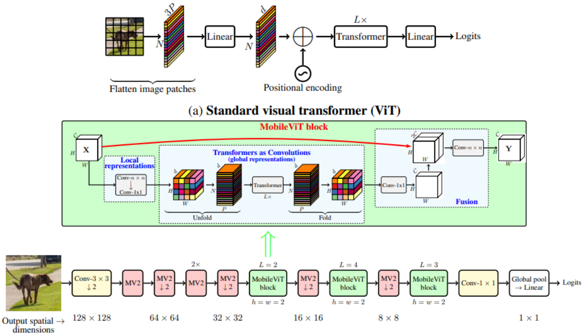
Task and Motivation

Task: Continuous Sign Language Recognition (CSLR). A sequence-to-sequence problem, where the input is a video and the output is a phrase in natural language.

Dataset: RWTH-PHOENIX-Weather 2014T. Around $\approx$ 4 thousand videos, 25 frames per second, 210×260 pixels.

MobileViTs

MobileViTs[2] are a lightweight version of Vision Transformers (ViTs) that are designed to run on mobile devices.



(b) **MobileViT**. Here, Conv-$n \times n$ in the MobileViT block represents a standard $n \times n$ convolution and MV2 refers to MobileNetv2 block. Blocks that perform down-sampling are marked with $\downarrow 2$.

Figure: MobileViT architecture.

Our novel MobileViT architecture

We adapt the MobileViT architecture to CSLR by performing spatio-temporal aggregation across multiple frames. This comes at no additional cost in terms of parameters, while allowing full exploitation of the temporal dimension.

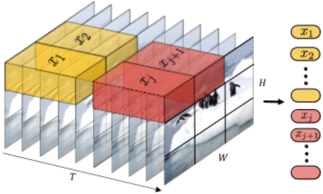


Figure: When the temporal dimension is processed in patches, the volume is called a tubelet.

First approach: Encoder-only MobileViT

To ensure a minimal amount of parameters, we first tried to train an encoder-only MobileViT. We use CTC loss to perform per-frame prediction of tokens from the encoder output, fusing repeated tokens.

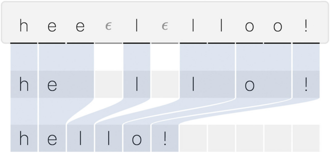


Figure: Output of a CTC loss trained model. repeated tokens separated by a blank token are fused.

Second approach: Encoder-decoder MobileViT

To compare ourselves with a more classical approach, we also trained an encoder-decoder MobileViT. We attach a decoder head to the encoder output, and train the model with classic CE loss and teacher forcing.
This massively increases the number of parameters w.r.t. the encoder-only model, but it simplifies the training process.

Encoder-only MobileViT: Results

The encoder-only model gets trapped in a local minimum (predicting 0-length sequences) and does not recover. Repeated experiments with different configurations did not yield better results.

Figure: Validation loss for the encoder-only MobileViT.

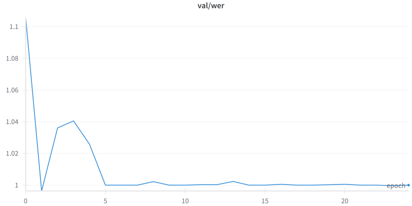


Figure: Validation Word Error Rate for the encoder-only MobileViT, stuck at 1.0.

Encoder Decoder MobileViT: Results

The encoder-decoder approach shows a smoother convergence and a lower WER w.r.t. the encoder-only model.

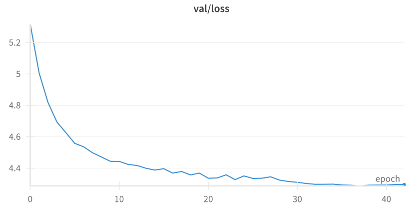


Figure: Validation loss for the encoder-decoder MobileViT.

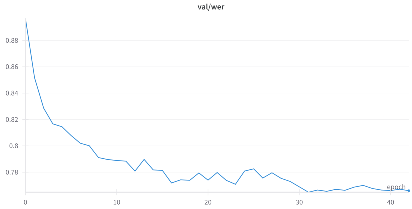


Figure: Validation Word Error Rate for the encoder-decoder MobileViT.

Quantitative results

Model	Parameters (M)	WER
Encoder-only XXS MobileViT (ours)	6.09	1.00
Encoder-decoder XXS MobileViT (ours)	7.80	0.79
SlowFastSign[1]	52.90	0.18
LCSA[3]	16.17	0.21

Table: Parameters (in millions) and validation Word-Error-Rate (WER) for the models.

Conclusions

CSLR is a challenging task that requires massive computational resources. Even while employing models that are specifically designed for mobile devices, we still struggle to achieve sufficient capacity to match state-of-the-art models.

For both the encoder-only and encoder-decoder models, limited experiments with increasing parameter size showed improvements in terms of WER. Showing that scaling the model up might be beneficial.

Thank you for your attention!

Questions?

J. Ahn, Y. Jang, and J. S. Chung.
SlowFast Network for Continuous Sign Language Recognition, Sept. 2023.
arXiv:2309.12304 [cs].

S. Mehta and M. Rastegari.
MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer, Mar. 2022.
arXiv:2110.02178 [cs].

R. Zuo and B. Mak.
Local Context-aware Self-attention for Continuous Sign Language Recognition.
In *Interspeech 2022*, pages 4810–4814. ISCA, Sept. 2022.