rijksuniversiteit
groningen

HandTalk: American sign language recognition by 3D-CNNs

Julia Walczyńska

American Sign Language

› predominant sign language of deaf communities in the United States

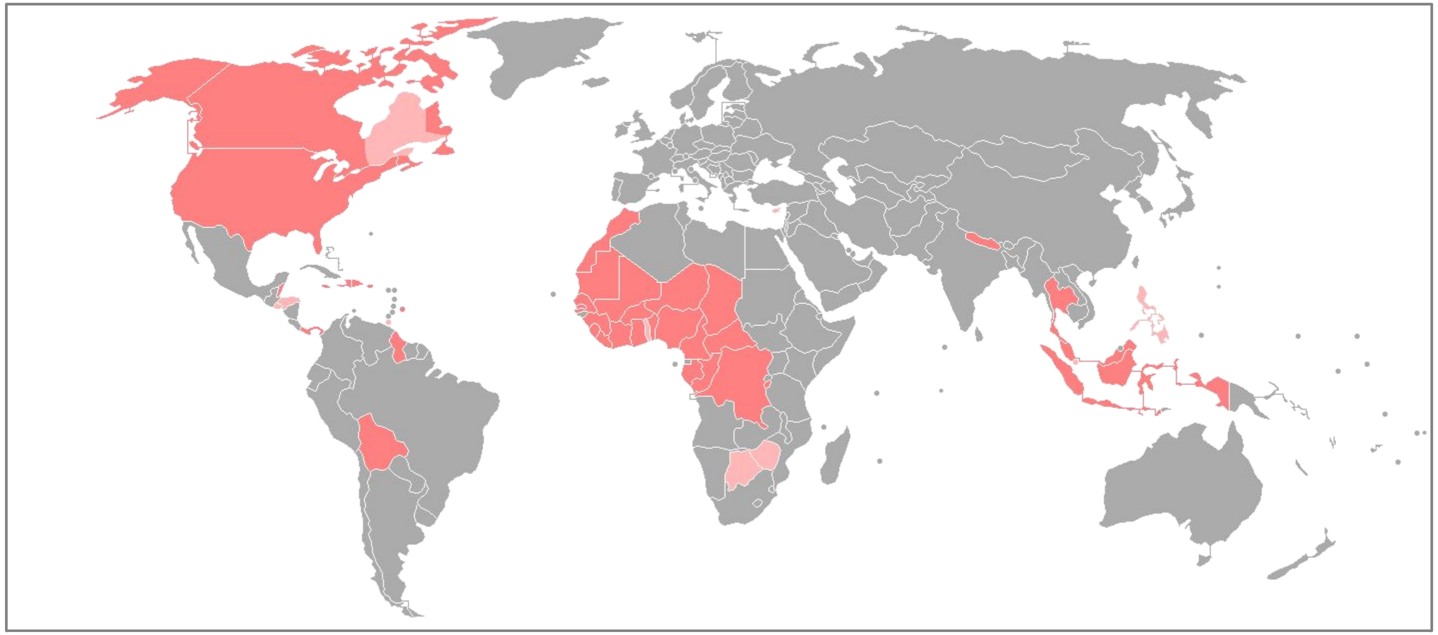

Figure 1: Areas where ASL is a national sign language (dark pink) or in significant use alongside with other languages (light pink). Source: Wikipedia contributors. (2022, June 10). American Sign Language. In Wikipedia, The Free Encyclopedia.

ASL phonology

- three types of signs:
  - one-handed
  - symmetric two-handed
  - non-symmetric two-handed
- each sign has five parameters:
  - handshape
  - movement
  - palm orientation
  - location
  - non-manual markers

Figure 2: Comparison of signs differing by only one of the parameters: (1) palm orientation, (2) handshape

rijksuniversiteit groningen

Goal

creating an american sign language classification system

- suitable for mobile devices
- camera-based (RGB images as input)
- to be used in real-time

rijksuniversiteit
groningen

Dataset

WLASL - A large-scale dataset for Word-Level American Sign Language

- over 20 000 videos
- 2 000 classes
- over 100 different signers
- methods of comparison: I3D and Pose-TGCN

WLASL100 - version of WLASL with 100 classes, over 2 000 videos by 97 signers

- I3D accuracy: 65.89%
- Pose-TGCN accuracy: 25.97%

3D-MobileNetV2

- mobile tailored computer vision model
- low number of operations and memory needed
- achieved 94.59% accuracy on Jester dataset
  - Jester: largest available hand gesture dataset

3D-MobileNetV2 architecture

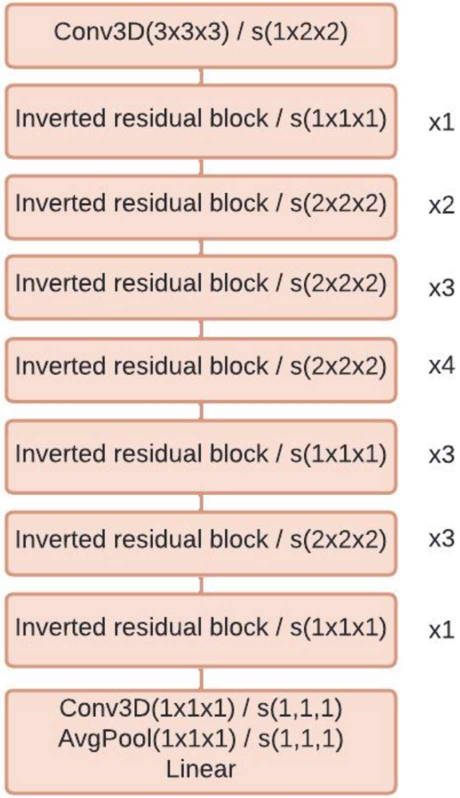

Figure 3: 3D-MobileNetV2 architecture. Inverted residual blocks are either with stride 1 or with stride 2 (meaning spatio-temporal 2x downsampling). On the right side of each block, it is noted how many times it should be repeated.

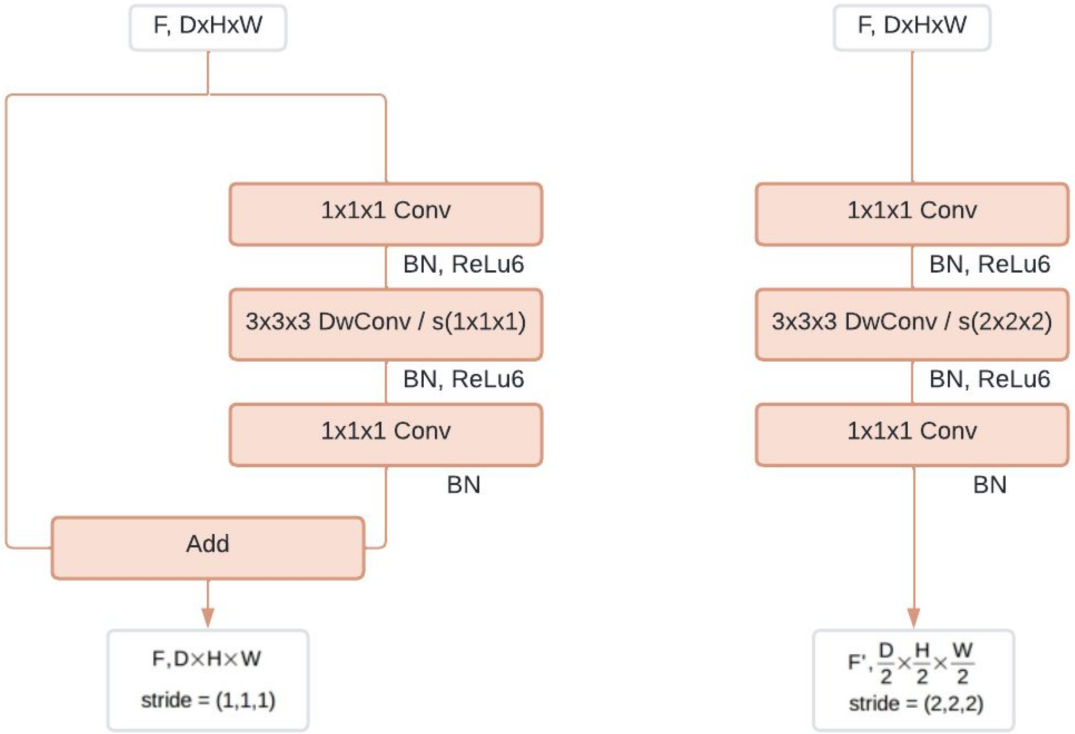

Figure 4: 3D-MobileNetv2 block (left) and 3D-MobileNetv2 block with spatiotemporal x2 downsampling (right). Based on source:: Kopuklu, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0).

rijksuniversiteit groningen

Linear Bottlenecks

- three layers: 1x1x1, 3x3x3 and 1x1x1 convolution
  - 1x1x1 layers are computationally cheap and are responsible for reducing and later restoring dimensions
- in a classic bottleneck, Rectified Linear Unit (ReLu) function is used at the end…
  - …but it was found to hurt the performance as it destroys too much information
- no ReLu in the last layer of linear bottleneck

Inverted residuals

- inverted bottleneck
  - first 1x1x1 conv expands features rather than reduces them
- allows memory efficient implementations

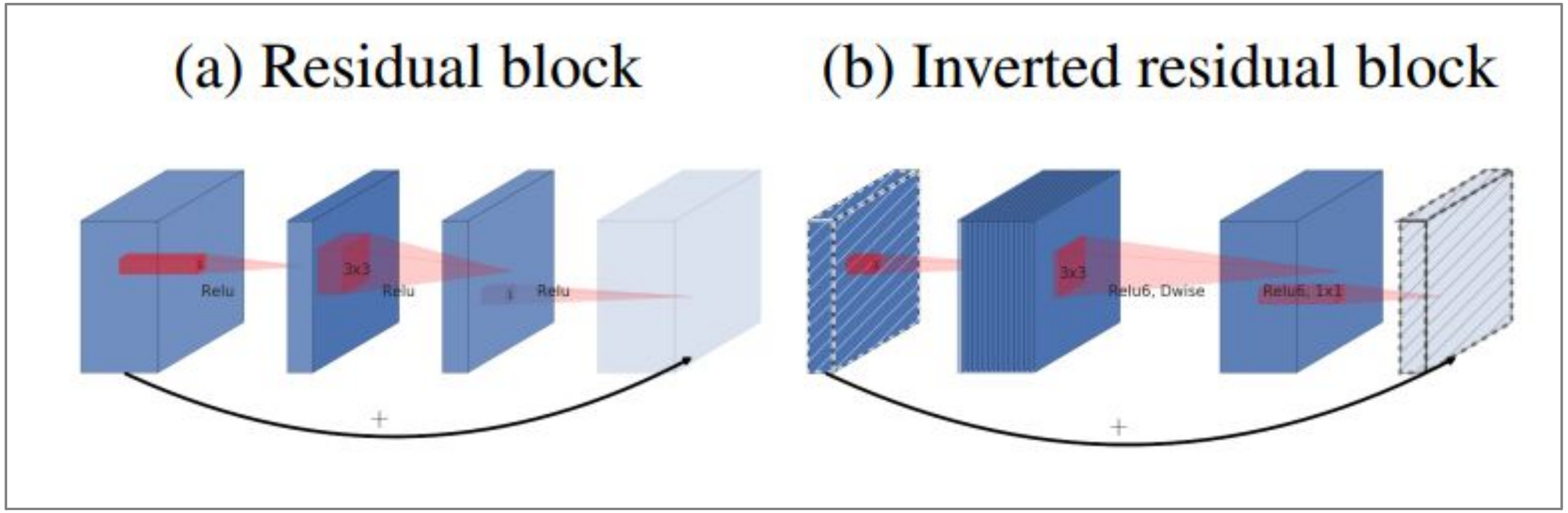

Figure 5: Comparison of residual and inverted residual block. Source: Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

Depthwise Separable Convolutions

- comparing to standard convolution layers, the computation cost is 8-9 times lower with almost no impact on accuracy
- key idea: replace full convolutional operator with a factorized version splitting it into two layers
  - first layer: depthwise convolution
    - lightweight filtering
    - applies a single convolutional filter per input channel
  - second layer: pointwise convolution
    - building new features through computing linear combinations of the input channels
    - 1x1 convolution

Data preprocessing

- Extracting frames from video
- Resize
- Random timewise crop
- Padding
- Random horizontal flip
- Random brightness and contrast
- Random crop

Training

- 3D-MobileNetV2 with input 112x112x32
- AdamW optimizer with initial learning rate of 0.0001 and weight decay 0.001
- cross entropy loss
- ReduceLROnPlateau scheduler
- 80% videos used for training and 20% for validation
- batch size: 32
- about 400 epochs
- trained on Peregrine cluster
  - 6 cores @ 2.7 GHz (12 cores with hyperthreading)
  - 128 GB memory
  - 1 Nvidia V100 GPU accelerator card

Results

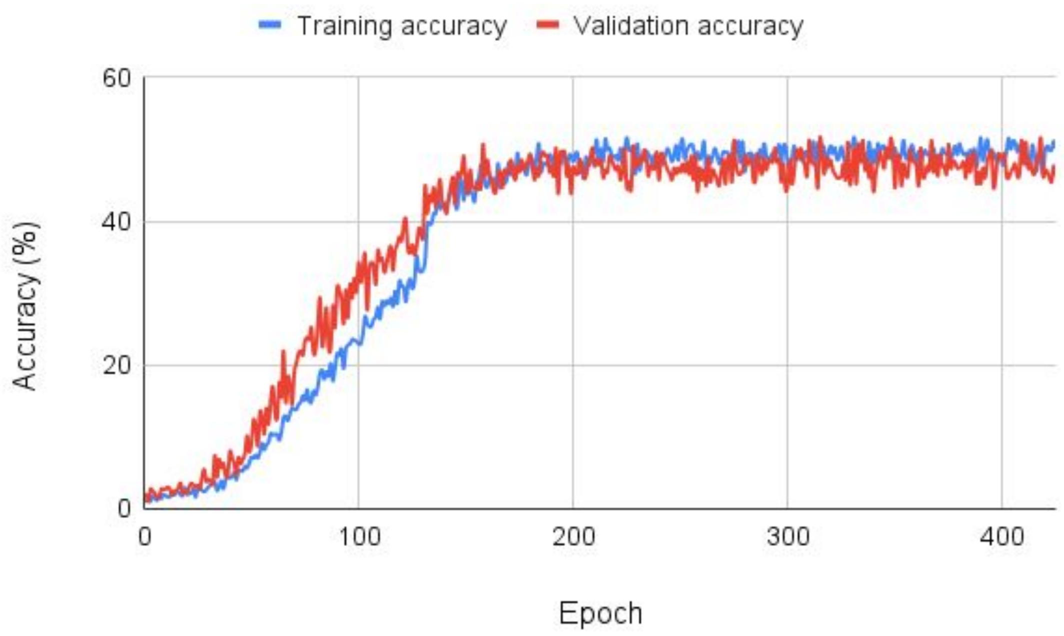

Figure 6: Training accuracy and validation accuracy of MobileNetV2 trained on WLASL100 dataset.

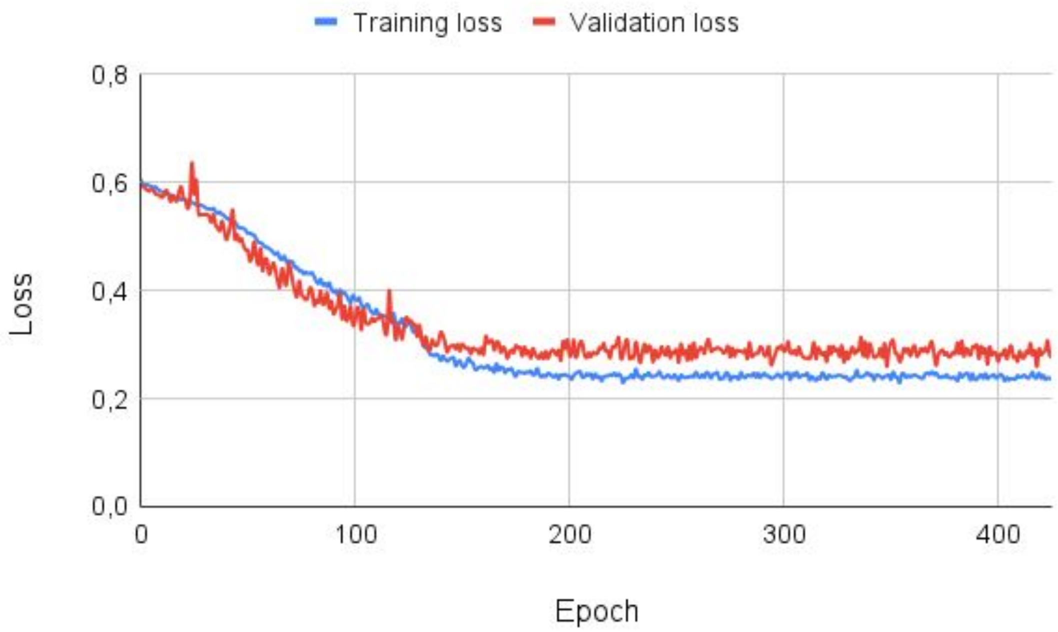

Figure 7: Training loss and validation loss of MobileNetV2 trained on WLASL100 dataset.

Conclusion and discussion

- low number of training samples per class
- sometimes there are few different signs with the same meaning
- other methods included extracting hand key points, using optical flow

rijksuniversiteit
groningen

Bibliography

Kopuklu, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0).

Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1459-1469).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

Wikipedia contributors. (2022, June 10). American Sign Language. In *Wikipedia, The Free Encyclopedia*. URL: https://en.wikipedia.org/w/index.php?title=American_Sign_Language&oldid=1092395775

Rules of dominant, passive, and symmetrical hands. In *Handspeak.* URL: https://www.handspeak.com/learn/index.php?id=98