ECCV 2022 Sign Spotting Challenge

Fact sheet

I. TEAM DETAILS

- Team leader name: Ryan Wong
- Username on Codalab: ryanwong
- Team leader affiliation: University of Surrey
- Team leader email: rwong@surrey.ac.uk
- Name of other team members (and affiliation):
  Necati Cihan Camgoz (University of Surrey)
  Richard Bowden (University of Surrey)
- Team website URL (if any):
- Competition track:
  - (X) Track 1: **MSSL** (multiple shot supervised learning).
  - ( ) Track 2: **OSLWL** (one shot learning and weak labels).

II. CONTRIBUTION DETAILS

A. Hierarchical I3D model for continuous sign spotting

The I3D model [1] has shown to be successful on sign language recognition datasets, such as WLASL [2] and MSASL [3], to identify a sign in a selected temporal region of a sign video. Therefore given a sign video sequence of 32 frames, it will only output a single sign prediction, which limits the network to coarse temporal predictions. We introduce an additional hierarchical component to the I3D model which learns coarse-to-fine predictions for accurate frame level sign predictions.

B. Method

For our approach we employ and modify a I3D model from [4] which was pretrained on the WLASL dataset. The ReLU activation functions are replaced by the Swish activations [5] as it has been shown to improve results for sign language recognition [6]. Instead of taking the output at the final layer with spatial temporal global average pooling and applying a fully connected layer for class predictions, we take the output before global average pooling and additional feature outputs before the 3D max pool layers in the I3D model. We therefore obtain 3 feature outputs each with a higher temporal resolution. For a given sequence length of 32 frames of dimensions 224×224, the base I3D model outputs the following features $1024 \times 4 \times 7 \times 7$, $832 \times 8 \times 14 \times 14$ and , $480 \times 16 \times 28 \times 28$, with a temporal resolution of 4, 8 and 16, respectively. As shown in fig. 1, a hierarchical network uses these inputs to output coarse-to-fine temporal predictions ranging from 4 (1 prediction every 8 frames), 8 (1 prediction every 4 frames), 16 (1 prediction every 2 frames) and 32 (1 prediction every 1 frames) temporally aligned predictions.

In fig. 1, we define the following:

- **POOL** - Global average spatial pooling
- **UP** - 3D transpose convolution which doubles the temporal dimension and halves the feature dimension.
- **CAT** - Concatenates the feature outputs
- **MERGE** - Consists of 3D convolution layer followed by Batch Normalisation and ReLu activation $(\times 2)$
- **conv** - convolution layer with kernel size of 1.
- **DOWN + POOL** - Consists of a ResNet Basic Block [7] which halves the feature dimension followed by spatial pooling.
- **interpolate + CAT** - Interpolates the temporal dimension to size 32 using the nearest approach and concatenates the input features.
- **FC** - Fully connected layer with output size of the number of classes.

Cross Entropy loss is used to predict the sign at each time segment for the course-to-fine predictions where the target is set to the sign label if it exists within the time segment and the additional unknown class (class 61) is set to be the target when the time segment does not have an known class label.

The final predictions are based on temporally interpolating the softmax of the logit features for each of the predictions to the original sequence length (32) and averaging the 5 output results to obtain the probabilities for each class prediction at frame level.

During training, the input to the model uses 32 consecutive frames of size 224×224 with random data augmentation such as random cropping, rotation, horizontal flipping, colour jitter and gray scaling. Mixup [8] is also applied with an α value of 1.0.

Models were trained on 5 fold cross validation (where folds were separated by signer number) for 200 epochs with a batch size of 8 and Adam optimizer with an initial learning rate of 3×10^{-4} and cosine annealing learning decay. The best checkpoint was chosen based on the best local validation F1-score.

This is repeated 3 times with different random sampling probabilities (rsp) were instead of selecting only frame regions around only known sign classes, we randomly select frame regions from other areas in the video based on the rsp. The probability used during our training was 0%, 10% and 50%. This is important as there is a trade off between precision and recall, choosing frame regions with known

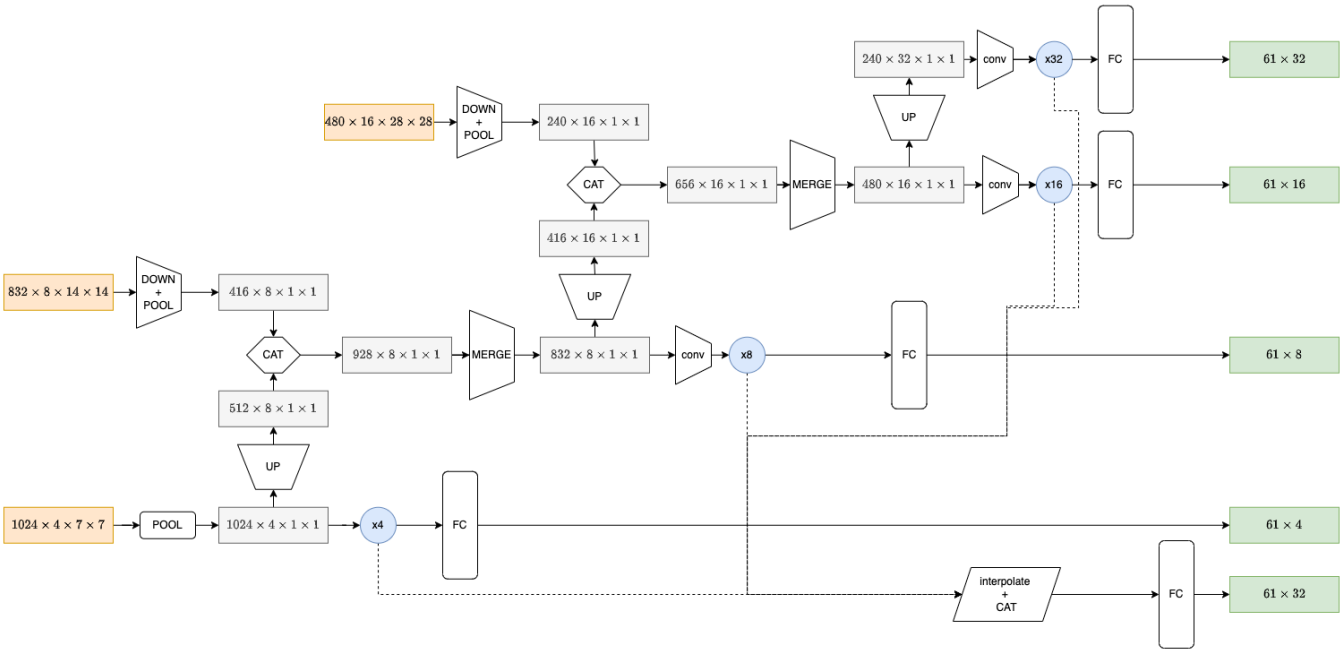

Fig. 1. Flow of the introduced hierarchical component of the network when given a sequence of 32 frames with dimensions 224×224. The three inputs are taken from outputs of various stages of the I3D model. The output consists of five temporal segment predictions of various segment lengths.

sign classes have the highest recall but significantly lower precision and models trained with 50% random frame region sampling had significantly higher precision but low recall.

For the final submission an additional 7 fold cross validation set is used consisting of a mixture of training and validation dataset and follows the same process as the original 5 fold approach.

For the final submission the models trained with the fold that consists of signer 5 in the local validation set is removed from the ensemble due to the degrade in performance. Therefore an ensemble based on the mean probability outputs of 30 models ($(4 \times 3) + (6 \times 3)$) are used for the final phase, were we apply a temporal stride stride of 1 over the test videos, taking the average between overlapping time segments.

C. Challenge results

Table I shows the obtained results, shown in the leaderboard of the challenge.

TABLE I
RESULTS FROM LEADERBOARD (TEST PHASE) OBTAINED BY THE PROPOSED APPROACH.

Rank position	avg F1
1	0.606554

D. Final remarks

One of the main benefits the hierarchical I3D model is that it can predict signs at frame level as opposed to using I3D model for single class prediction outputs for a given region in time. An important factor to consider is recall versus precision as models trained with an rsp of 0.0 has significantly higher recall but lower F1-score due to the low precision than models trained with a higher rsp.

III. ADDITIONAL METHOD DETAILS

- **Did you use pre-trained models?** (X) Yes, ( ) No
  The pretrained I3D model from [4] which was initially pretrained on the BSL-1K dataset and then the WLASL dataset was used as pretraining.
- **Did you use external data?** ( ) Yes, (X) No
- **Did you use any kind of depth information (e.g., 3D pose estimation trained on RGBD data)?** ( ) Yes, (X) No
- **At the final phase, did you use the provided validation set as part of your training set?** (X) Yes, ( ) No
  A second set of models were trained on training and validation set using the same process as models trained on only the training dataset. Instead of 5 fold cross validation used only with the training set, 7 folds were used with the training and validation set (splitting folds by signer).
- **Did you use other regularization strategies/terms?** ( ) Yes, (X) No
- **Did you use handcrafted features?** ( ) Yes, (X) No
- **Did you use any face / hand / body detection, alignment or segmentation strategy?** (X) Yes, ( ) No
  We used OpenPose [9] as a body detector to crop signer region.
- **Did you use any pose estimation method?** ( ) Yes,

(X) No

- **Did you use any spatio-temporal feature extraction strategy?** (X) Yes, ( ) No
  As described in the methodology features from various stages in the I3D model were used for input into the hierachical component of the network.
- **Did you explicitly classify any attribute (e.g., gender/handedness)?** ( ) Yes, (X) No

- **Did you use any bias mitigation technique (e.g. rebalancing training data)?**
  (X) Yes, ( ) No
  An imbalanced dataset sampler was used to sample lower frequency classes more frequently and higher frequency classes less frequently during training.

IV. Code repository

Code repository: `https://github.com/ryanwongsa/ECCV22_Chalearn-MSSL`

Data link: `https://drive.google.com/file/d/1FN0t3H5bAB6fL8lsjNPonSsr_fUanHR4/view?usp=sharing`

References

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[2] D. Li, C. Rodriguez, X. Yu, and H. Li, "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*, 2020, pp. 1459–1469.

[3] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv preprint arXiv:1812.01053*, 2018.

[4] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues," in *European conference on computer vision*. Springer, 2020, pp. 35–53.

[5] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[6] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413–3423.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.