IMPLEMENTING THE SIGNBERT FRAMEWORK FOR ISOLATED SIGN LANGUAGE RECOGNITION (ISLR)

Heysem Ismail
Department of Computer Science
Hacettepe University
Ankara, Beytepe, Turkey
heysem.ismail@cs.hacettepe.edu.tr

ABSTRACT

Sign language recognition task does not have enough data sources to be able to overcome over-fitting problems. So, frameworks like signBERT Hu et al. (2021) achieve model pre-training by leveraging self-supervised techniques like mask and reconstruction of visual tokens. The proposed framework, jointly uses different types of embeddings and masking strategies, and incorporates pre-trained model of hand reconstruction. This project try to show the effectiveness of pre-training models using self-supervised techniques without the need of extra data. So that, we try to rebuild the most ideas in this framework and re-implement some experiments that illustrate how this method can improve the performance of Sign Language Recognition (SLR). Instead of 2 Chinese and 2 American sign language datasets used in the original paper Hu et al. (2021), one large-scale Turkish AUSTL dataset and an other one American dataset called WLASL are treated in this project. The code and models are available at: https://www.github.com/Heysem82/CMP719Project

1 INTRODUCTION

The Sign Language Recognition (SLR) research field concerns with translating the visual gestures that used by deaf community to communicate, into words and sentences understandable by hearing people. Our paper Hu et al. (2021) focuses on its word-level branch called Isolated Sign Language Recognition (ISLR) which is a challenging classification problem mainly depends on the movements of hands or the upper part of body generally. This problem can be treated using RGB-based or Pose-based methods. RGB-based methods consider each pixel in video frames to understand the visual signs, which gives high score results, but computationally it is expensive method. On the other hand, pose-based representation selects to determine and follow the movement of certain and limited key-points of body, like joints of hands and body upper parts. Although, pose-base methods are compact and computationally more efficient, its performance is significantly lower than RGB-based ones. That because its depending on the pose estimation algorithms which give approximated and not always accurate values. In addition, the limitation of sign data sources prevents getting high classification accuracy when using pose-base methods alone. So that, self-supervised methods expected to be very helpful in compensate the lack of data, and consequently raising the performance without losing the efficiency. The studied paper Hu et al. (2021) try to leverage the success of self-supervised pre-training methods previously used in Bert models which use masking then prediction and reconstruction techniques. Concepts like masked language modeling (MLM) and next sentence prediction (NSP) are main factors of the great success obtained by Bert models. SignBERT Framework is developed to reduce the performance gap between pose-based and RGB-based methods. It aims at incorporating pre-training using self-supervised learning techniques like mask modeling strategies to capture the hierarchical contextual information contained in the sign data sources. To implement this strategies, SignBERT framework considers hand and body key-points as a visual tokens, masks some of them, then, is enforced to reconstruct the masked ones. For accurate key-points reconstruction, SignBERT benefits from prior hand-aware pre-trained model called MANO model. This model is learn from large number of hand scans and images to be one of best tools that gives accurate results of hand modeling from law level hand pose and shape parameters. Evaluating Sign-

BERT Framework is performed through two main tasks: self-supervised pre-training on large-scale dataset, then fine-tuning on (ISLR) downstream classification task. In the next sections, detailed description of the framework, evaluation methods, and metrics used in the project are explained. Then, designs and settings of the implemented experiments, and comparisons of results with those of the original paper are detailed. Finally, list of challenges encountered during the phases of this project and solutions followed to deal with problems are presented.

2 Approach

2.1 Model description

Pre-training phase is implemented by applying self-supervised technique that mask and reconstruct visual tokens, similar to mask modeling strategy that applied in the original BERT paper Devlin et al. (2018). For that purpose, three masking strategies are applied: masked joint ,masked frame, and identity strategies. Masked joint model, randomly and in certain ration, selects number of key-points in each frame to be masked. Additionally, masked frame strategy considers each frame as a token in sequence of tokens (frames), and randomly masks number of frames completely.Finally, in the process of identity modeling, tokens are kept unchanged and fed into the framework. The backbone of the pre-trained model is an encoder-decoder architecture where the encoder part is attention-based Transformer, while the decoder is a hand-model-aware part incorporates pretrained model called MANO model to better model hierarchical context over the hand sequence. Based on parameters describe the hand status (θ and β for pose and shape respectively), and weak-perspective camera parameters($\mathbf{c}_r$, $\mathbf{c}_o$, and c_s for rotation, translation and scale respectively), MANO model can perform accurate hands reconstruction from the masked input sequence. These parameters are extracted as latent semantic embedding which is the output of fully connected layer takes the representation generated by Transformer encoder as input. MANO model treats one hand at the same time, so that left or right hands information are fed separately. Three types of embeddings: gesture state, temporal, and hand chirality are used sequentially before the pre-extracted and masked 2D hand pose sequences are fed into the framework. In gesture state embedding, the joints and the physical connections among them are represented in undirected spatial graph as nodes and edges, respectively. Then spectral-based GCN model Cai et al. (2019) is applied for getting more semantic representations like what is obtained when using word2Vec embeddings in NLP tasks. As the self-attention layer of encoder is invariant to order information, positional encoding strategy like in Vaswani et al. (2017) is required, where at the same time t frames of both hands take the same positional encoding. For embedding chirality of hands at each frame, two special tokens 'L' and 'R' implemented by the WordPiece embeddings Wu et al. (2016) are adopted. The output of hand reconstruction using MANO model is 3D pose and joint coordinates information. For getting 2D joints information, orthographic projection is applied. For fine-tuning the model previously pre-trained, the decoder part (MANO model and latent semantic extraction layer) is removed, and a prediction head is injected at the end. pre-extracted and embedded pose information is fed to the new pretrained architecture, and the representation generated by encoder transformer is used for classification task, where traditional training methodologies are applied. Figure 1 from the paper Hu et al. (2021) give a detailed preview of the framework described above:

2.2 Evaluation Methods

pre-training process in the original paper is performed using 4 datasets, two of them are American sign language datasets named MSASL and WLASL, while the others are Chinese sign languages known as CLR500 and NMFs-CSL. However, because I could not have access to the Chinese ones I use large-scale free Turkish sing language dataset called: AUTSL instead. In addition, due to limited computation resources, and one American sign language WLASL dataset is treated. As a result, for pre-training AUTSL and WLASL are used, while the fine-tuning task is evaluated just using the Turkish AUTSL dataset. For implementing pose estimation and extracting full 2D key-points from videos, MMPose tool is used in the paper. Instead, I use framework from Google called mediaPipe, which has a dedicated library forpose estimation of hand key-points called hand landmarker. It is, compared with MMPose, easy-of-use, and generates pose information with a high confidence. During pre-training, sum of two losses are used as objective function: hand reconstruction loss $\mathcal{L}_{rec}$ and regularization $\mathcal{L}_{rec}$ loss, as in next equation: $\mathcal{L}_{rec} = \mathcal{L}_{rec} + \lambda\mathcal{L}_{rec}$

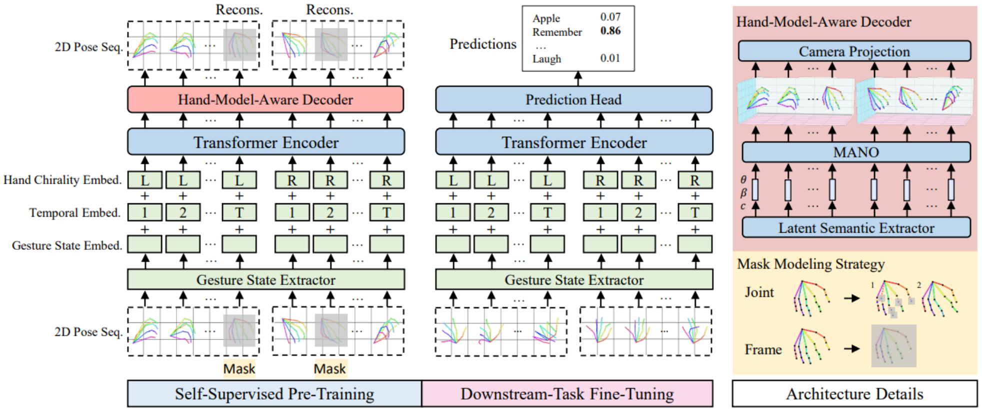


Figure 2. Illustration of our SignBERT framework, which contains self-supervised pre-training and fine-tuning for the downstream sign language recognition. The pre-extracted 2D hand pose sequence of both hands is fed into the framework. Each hand pose is viewed as a visual token, embedded with gesture state, temporal and hand chirality information. In self-supervised pre-training, we design several mask modeling strategies and incorporate model-aware hand prior to better exploit hierarchical contextual representation. For the downstream SLR task, the pre-trained Transformer encoder is fine-tuned with the prediction head to perform recognition.

Figure 1: Summary of SignBERT framework

where λ is a weighting factor. In the paper, reconstruction loss $\mathcal{L}_{rec}$ is calculated by adding a factor that considers just joints with prediction confidences more than certain threshold $\epsilon \geq 0.5$. However, mediapipe tool by-default applies that threshold. So that, reconstruction loss is calculated in a traditional way; the difference between the original coordinates and predicated ones, like in this equation: $\mathcal{L}_{rec} = \sum_{t,j} \left\| \widetilde{J}_{2D}(t,j) - J_{2D}(t,j) \right\|_1$. On the other hand, regularization loss constrains inputs of MANO model: shape and pose parameters, for ensuring generating a plausible mesh and keeping the signer identity unchanged. It is calculated like in the paper, as follows: $\mathcal{L}_{reg} = \sum_t \left(\|\theta_t\|_2^2 + w_\beta \|\beta_t\|_2^2 + w_\delta \|\beta_t - \beta_{t-1}\|_2^2 \right)$, where w_δ and w_β are hyper-parameters work as weighting factors.

The effectiveness of the self-supervised pre-training methodology applied in the framework, are measured on downstream isolated SLR task.

2.3 Evaluation Metrics

For comparing the results of experiments, the paper uses both per-instance and per-class Top-1 and Top-5 accuracy as evaluation metric, because the number of samples not equals in each class of Chinese datasets. However, in our experiments, just per-instance Top-1 and Top-5 accuracy are measured and reported, as the numbers of samples are balanced in each class of the Turkish sign dataset.

3 Experimental Details

3.1 Models Settings

The baseline model is obtained by removing decoder part, and training from scratch; without applying self-supervised pertaining, encoder Transformer model that has the same architecture as in its original paper Vaswani et al. (2017), provided with prediction head. Both of input embeddings to and outputs from this model have the same dimension. The number of heads in multi-head self-attention layer is 8 in Vaswani et al. (2017). But, as getting a state of the art performance is not a goal of this project, and for reducing the computational complexity, number of heads is set as (4) in our experiments. However, as setting number of blocks in encoder to 3 gives the best result in the paper, it is also applied in our experiments. Number of classes to be classified is set as 226 like in AUTSL dataset. For pre-training experiments, the above model is modified to include fully connected layer that takes the input from the top of encoder and extracts the latent semantic embedding vector of MANO model parameters. The size of this parameters are set to comply with input size required by MANO model. So that, the vector is of 42 dimensions divided as 25, 10, 3, 3, and 1 for pose embedding θ ,shape embedding β, global rotation $\mathbf{c}_r$, translation $\mathbf{c}_o$, and scale c_s respectively. For fine-tuning experiments, latent semantic extraction layer and MANO model are removed from the previous pretrained model, and with the same architecture and settings of the baseline, the pre-trained model are trined. In general these settings are similar to those in the paper with some modifications as mentioned above. For using spectral-based GCN model for gesture state embeddings, certain configurations have to be done before that. Connections between joints, symmetric joints, and how joints are grouped fro pooling, are settings defined according to what mentioned in Cai et al. (2019). Moreover, distinguishing right hand from left one is done by embeddings of special tokens 'R' and 'L'. pre-trained WordPiece embedding model from Bert models is used for this purpose. It accepts the dimension size to be a multiple of 768. So that, all other embeddings and outputs are set to comply with this number.

3.2 Training Details

The implementing the above concepts and designs is performed across two phases: extracting pose information from videos, then feeding it to the framework for pre-training and fine-tuning experiments. For the first stage the paper uses MMPose tool for extracting full 133 2D key-points of the body, because their experiments study two cases. In first one, just hand joints are considered, while the impact of using full body joints are studied in the second case. However, because my experiments focus on implementing just the first case, hand landmarker pre-trained model from MediaPipe framework are used in the first stage. For that purpose, every frame in video is resized to 224×224 size, then it is fed to the hand landmarker model for extracting pose information of joints in both left and right hands. Finally, these coordinates information of 21 key-points of every hand in each frame are stored in Parquet file, so that one Parquet file for each video is the output of this stage. In second phase, content of these files are the input to the propose framework. Pre-training and fine-tuning experiments are implemented on 1 GPU device P100 from Kaggle has 16 GB Memory, while training from scratch is done on locally NVIDIA GeForce RTX 3050 Ti Laptop GPU of 4 GB memory. For all experiments, Adam optimizer is adopted as optimization method, with weight decay of L2 regularization is set as 0.0001, and momentum equals 0.9. Due to the ambiguity in target of learning rate decay strategy mentioned in the paper, whether it is for training or for fine-tuning experiment, I consider it for fine-tuning, and settings from next version of paper are adopted for pre-training. Therefore, learning rate starts with 0.001 and and reduced by 0.1 factor every 20 epochs in finetuning, and learning rate of 0.0001, with a warm-up of 6 epochs, and linear learning rate decay when model is pretrained. Total number of epochs is not determined in the paper. So, considering the computing cost ,it is set as 60 for training, pre-training, and fine-tuning experiments. Hyper-parameters of losses functions λ, w_δ and w_β are set like the paper as 0.01, 10.0 and 100.0, respectively. Mask threshold of key-points masking is set to 0.15.

3.3 Datasets Settings

AUSTL dataset contains about 28000 video clips covering 226 classes of Turkish sign language, while WLASL dataset has about 12000 ones belongs to 2000 sign words, and both datasets are used for pre-trianing experiments. For fine-tuning and training from scratch, just AUSTL dataset is targeted, as it is more balanced in terms of instance in each class. While 50% and 25% of AUSTL dataset are dedicated for training and validation respectively, the rest 25% of data are used as test set. Otherwise, because number of frames in each video is different, max number of frames is set as 64 during all pre-training and training experiments.

3.4 Results Discussion and Comparison

The results of paper Hu et al. (2021) experiments find that pre-training encoder transformer model using self-supervised learning with different masking strategies, improves the prediction accuracy when the pre-trained model is fine-tuned and tested on downstream classification task. Table 1 shows results from studied paper Hu et al. (2021).

For evaluating these findings, some of experiments are implemented using the settings detailed above with some modifications for getting results faster.

Table 1: Effectiveness of the masking strategy on MSASL dataset in paper. The first row denotes the baseline,

Mask		100		200		1000	
Joint	Frame	P-I	P-C	P-I	P-C	P-I	P-C
		63.01	62.72	57.69	57.56	41.85	38.30
✓		72.66	72.75	68.51	69.72	48.87	45.39
	✓	74.77	75.48	68.65	69.20	49.02	46.02
✓	✓	76.09	76.65	70.64	70.92	49.54	46.39

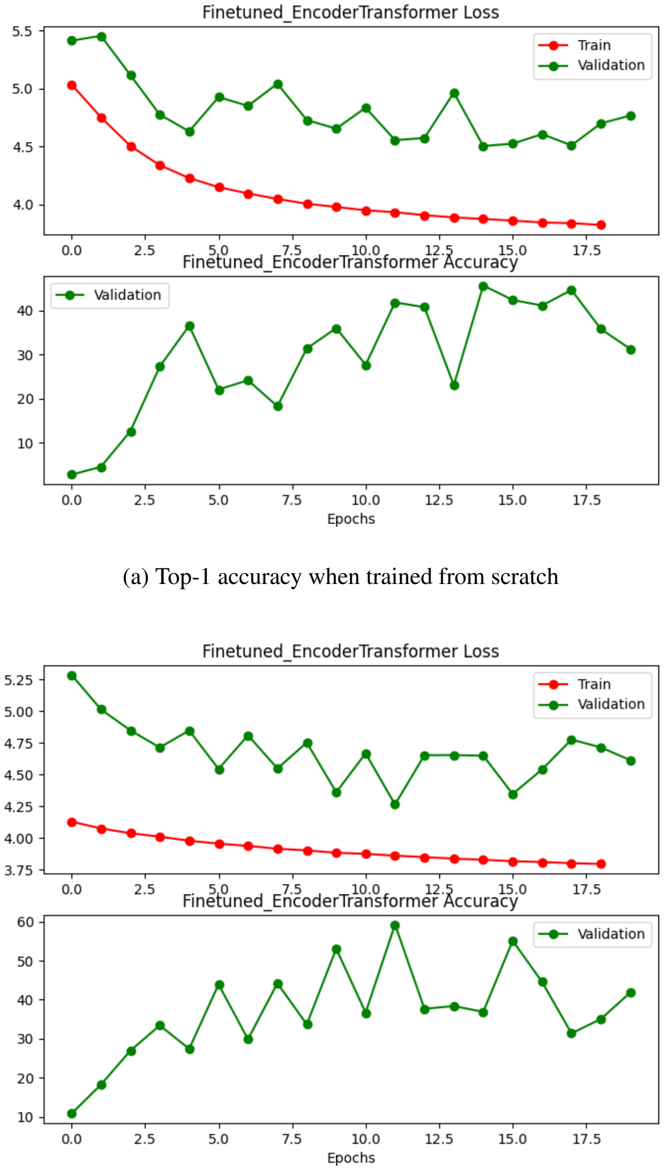
- Training Encoder Transformer Model from scratch without pre-training. For this experiment, I Follow the training recipe mentioned above with some modification to accelerate the training process. So that, Hand chirality embedding is disabled, embedding dimension is set to 256 instead of 768, total epochs is reduced to 20, number of heads is set to 4 instead of 8, and learning rate is set to fixed 0.001 with disabled scheduling. The first experiment results are shown in part (a) from Figure 2. It can be noticed that the best top-1 validation accuracy of model is about 45.6% and is 74.2% for top-5 accuracy, while evaluating the best checkpoint on test set gives about 47%.
- Pre-training Encoder-Decoder Model with Self-supervised Methods. In this experiment, just joint masking strategy is applied, and acceleration recipe is followed too. So that, just one dataset (WLASL) is used for pre-training, 4 heads for self-attention, total epochs is 20, fixed learning rate equals 0.001, and disabled Hand chirality embedding. The resulting pretrained model, without considering checkpoint with the best loss, will be used the next experiment.
- Fine-tuning the pre-trained model. The decoder part is removed and prediction head is added, then training is done using the same acceleration configuration of the first experiment. Part(b) of Figure 2 clearly shows that how the best top-1 and top-5 values of validation accuracy significantly improved to 59.2% and 83% in order. Moreover, top-1, top-5 accuracy when the model is evaluated on test test, confirms this findings when increases to 73% top-1 and 90.7% top-5 comparing to 53% top-1 and 70% top-5 before pre-training process.

In both exp1 and exp3, the evaluation results on test set are higher than those with validation ones, which clearly show that our model architecture and the recipe followed for training it, produces trained model has high ability of generalization to unseen data.

4 Challenges and Solutions

In this project we try to build SignBERT framework proposed in the paper, and to implement some experiments that evaluate the contribution of self-supervised learning methods to the accuracy of the downstream task like isolated sign language recognition. During my work, a lot of unexpected challenges are faced. The most important ones of them are summarized in the next points:

- No open source or free implementation code is available online. So, I forced to consider every part of the framework without any guide, which takes a long time for me.
  As a solution I read a lot of references that explain the briefed ideas in the paper. For example, gesture state embedding requires deep understanding of spectral-based GCN model where joints of hands and their connections are represented in graph, and certain pooling strategics are used. An other example is how special tokens 'R' and 'L' is embedded in pretrained Bert model.
- I could not reach the most of datasets used in the paper, especially the Chinese ones. So, a large-scale Turkish sign language is used instead. Although it is not as large as those in paper, but it is still sufficient to demonstrate the impact of self-supervised pre-training on the model performance.



(a) Top-1 accuracy when trained from scratch

(b) Top-1 accuracy after pre-training

Figure 2: Classification performance of Encoder-Transformer Model before and after self-supervised pre-training

- The documentation of MMPose library of pose estimation used in paper for extracting pose information, is not easy to be followed. Also, it extracts full body joints information. As my experiments focus on hand information, I found that hand landmarker from MediaPipe framework, which is specialized for hands landmarks, more convenient.

- One of most challenging problems, in terms of implementation, is that the number of frames in each video may be different from others. To deal with this issue, max frame number is determined, and the videos that have frames less than max number are padded. Then, this padding mask has to be considered when calculating the self attention values.

- During training from scratch, validation accuracy suffered from fluctuated values. Therefore, step decay learning rate scheduling is applied, and complexity of model is reduced by decreasing number of heads and number of encoder blocks.
- Because of high computational cost, some complexities are disabled, like hand chirality embedding with pretrained Bert model, which requires vector of size 768 dimensions at least. Also, scheduling are disabled and learning is raised and fixed at 0.001.
- In terms of possible improvements of this project, many types of experiments can be designed and applied.For example, additional datasets can be added to the pretraining process, frame masking strategy can be activated, disabled hand chirality embeddings and scheduling can be enabled for better performance results. An other proposed improvement is using discrete Variational Autoencoder (d-VAE) instead of MANO model, which is suggested and implemented in Zhao et al. (2023)

REFERENCES

Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2272–2281, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 18: 1527–1554, 2018.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11087–11096, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. *arXiv preprint arXiv:2302.05075*, 2023.