Scaling up ConvAtt for Sign Language Recognition

Gastón Rios[1,3]0000-0003-0252-7036, Pedro Dal Bianco[1,3]0000-0001-7197-8602, Franco Ronchetti[1,2]0000-0003-3173-1327, Facundo Quiroga[1,2]0000-0003-4495-4327, Santiago Ponte Ahon[1,3]0009-0002-8540-6546, Oscar Stanchi[1,4]0000-0003-0294-2053, and Waldo Hasperué[1,2]0000-0002-9950-1563

[1] Instituto de Investigación en Informática LIDI - Universidad Nacional de La Plata, 50 & 120, La Plata, 1900, Buenos Aires, Argentina
{grios,pdalbianco,fronchetti,fquiroga,sponte,ostanchi,whasperue}@lidi.info.unlp.edu.ar
[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA), La Plata, 1900, Buenos Aires, Argentina
[3] Becario Doctoral - Universidad Nacional de La Plata, 50 & 120, La Plata, 1900, Buenos Aires, Argentina
[4] Becario Doctoral - CONICET, Godoy Cruz 2290, Ciudad Autonoma de Buenos Aires, 1425, Buenos Aires, Argentina

Abstract. Sign language is crucial for communication within the deaf community, making Sign Language Recognition (SLR) essential for bridging the gap between signers and non-signers. However, SLR models often face challenges due to limited data availability and quality. This paper investigates various data augmentation and regularization techniques to enhance the performance of a lightweight SLR model. We focus on recognizing signs from the French Belgian Sign Language using a novel model architecture that integrates convolutional, channel attention, and self-attention layers. Our experiments demonstrate the effectiveness of these techniques, achieving a top-1 accuracy of 49.99% and a top-10 accuracy of 83.19% across 600 distinct signs.

Keywords: Handshape Recognition, Unbalanced Data, Limited Data, Sign Language, Human Motion Prediction

1 Introduction

Sign language uses hand movements, facial expressions, and body cues. Globally, there are over 300 distinct sign languages, which are generally not mutually intelligible. There is a critical need for Sign Language Recognition (SLR) systems to facilitate communication and improve technological accessibility for sign language users. Recent years have seen promising advancements in SLR systems [5]. These approaches primarily leverage deep learning techniques in computer vision or multi-modal processing. However, the efficacy of these deep

learning models is often constrained by the limited availability and diversity of sign language data. Unlike voice recognition, which benefits from vast datasets, sign language resources are comparatively scarce, making data collection a challenging, time-consuming, and costly endeavor. While speech processing models achieve human-like recognition [12] using hundreds of thousands of hours of voice recordings, sign language datasets rarely exceed a hundred hours [1].

In this paper, we address these challenges by exploring various regularization and data augmentation techniques to enhance the accuracy of deep learning SLR models[5]. A key innovation in our approach is the use of synthetic data augmentation [10], where we generate synthetic samples to pretrain our model. This technique allows us to artificially expand our dataset, potentially improving the model's generalization capabilities and robustness. Our generator model architecture was mainly inspired by siMLPe [6], and is therefore named CSiMLPe.

For our experiments, we utilized pose data from the French Belgian Sign Language Isolated (LSFB-ISOL) dataset to train a 1D convolutional neural network equipped with transformer layers, which we call ConvAtt. This model architecture combines the strengths of convolutional networks in capturing local patterns with the ability of transformer layers to model long-range dependencies, making it well-suited for the sequential nature of sign language data.

The remainder of this paper is structured as follows: Section 2 provides an overview of related works in Sign Language Recognition . Section 3 details the materials and methods employed in our research, including a comprehensive description of the French Belgian Sign Language Isolated (LSFB) dataset used and the architectures of our proposed models, ConvAtt and CSiMLPe. Section 4 presents our experiments and results, encompassing the setup configuration, various regularization techniques, data augmentation methods, and a thorough analysis of our experimental outcomes using the LSFB dataset. Finally, Section 5 concludes the paper by summarizing our key findings and suggesting potential directions for future research.

1.1 Contributions

Our research demonstrates that a combination of multiple techniques, including regularization, Discrete Cosine Transform (DCT), data augmentation, and synthetic data augmentation, significantly enhances the performance of our Sign Language Recognition (SLR) model to state-of-the-art levels. Specifically, our key contributions are:

- Development of a state-of-the-art SLR model (ConvAtt). This model achieves a top-1 accuracy of 49.99% and a top-10 accuracy of 83.19% on the LSFB dataset.
- Introduction of a novel sign language generation model (CSiMLPe) that can be conditioned by sign labels to produce specific hand gesture sequences. This model enables synthetic data generation, addressing the scarcity of sign language datasets.

[5] Code available at `https://github.com/okason97/HandCraft`

– Successful pretraining of our SLR model using synthetically generated data, resulting in improved performance without relying on additional external data sources. This approach offers a promising solution to the data scarcity problem in SLR research.
– Comprehensive evaluation of various techniques' impact on SLR model performance, providing insights that can guide future research in this field.

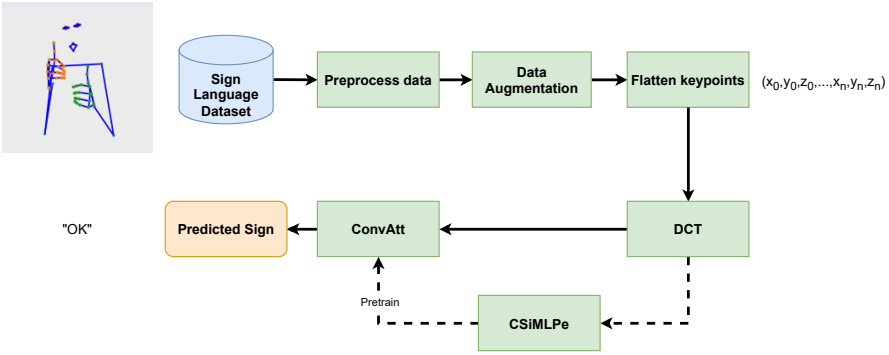

Fig. 1: Pipeline of the SLR process utilized. We first preprocess the data by removing samples and keypoints that do not meet our criteria, adjusting the data to a fixed frame length through cropping or padding, and normalizing it. Next, we apply a series of data augmentation techniques to each sample. The data is then flattened to a single dimension and a DCT transformation is applied to it before using is as input. A CSiMLPe generator model is trained and used to pretrain ConvAtt. Finally, this one-dimensional data is fed into the ConvAtt model, which makes the final prediction.

2 Related Work

SLR involves the classification of individual sign language gestures into written words or glosses. SLR models are trained using various types of data, including videos, images, depth maps , and poses of the signer's hands, body, and face [4], typically in a multi-modal approach . SLR can be classified as continuous, where sign language sentences are translated directly into text, or isolated, where a single sign is classified [4].

Currently, state-of-the-art SLR and gesture recognition commonly employ models based on convolutional neural networks [5] and transformer architectures [16], although combinations of the two architectures have been effectively implemented [13]. However, due to the data limitations of sign language datasets, innovative data representation methods and training pipelines have been developed

to enhance these models. Pose information extracted by pose recognition models has shown great success in improving performance [13]. This can be attributed to a better representation of the input data, retaining sufficient discriminative information to classify the signs while removing task-irrelevant information. Discrete Cosine Transform (DCT) [6] had been used in these cases to improve the representation of the data, encoding temporal information into it. In addition to these methods, data augmentation has proven to be an essential tool to increase the robustness of the model and reduce overfitting . Furthermore, data augmentation can diminish the representation distance between video and text data, easing data scarcity . In particular, synthetic data augmentation [10] had proven to be an effective method to generate new data samples from scratch.

3 Materials and methods

3.1 LSFB Dataset

The French Belgian Sign Language Isolated (LSFB-ISOL) dataset [3] is built upon the LSFB Corpus. It spans 25 hours of videos and poses of continuous isolated signs performed by 85 different signers. In this paper, we focus solely on the poses, as they reduce domain complexity and enable faster processing times for models. The final dataset comprised 60 landmarks: 12 facial (eyes and mouth), 6 body, and 21 per hand. After filtering, it contained 52,350 sign poses across 610 classes, with 10% allocated for testing. To mitigate class imbalance, all classes were oversampled to match the most populous class. We standardize input size by sampling 30 contiguous frames per clip, applying circular padding for shorter sequences. Temporal information is encoded using Discrete Cosine Transform (DCT) [6]. For pose prediction, we employ Inverse DCT to revert to the original representation.

3.2 Model Architectures

This section details our Sign Language Generation (SLG) model (CSiMLPe) and Sign Language Recognition (SLR) model (ConvAtt). We employ CSiMLPe to generate a synthetic dataset used to pretrain ConvAtt.

CSiMLPe CSiMLPe comprises four key components: label embedder, fully connected layers, transpose operations, and adaptive layer normalization.

Given the complex and extensive label space of signs in each language, label embedding is crucial for efficient input processing. This embedding layer also enables our model to learn semantic relationships between different signs. The process begins with applying a fully connected layer to the spatial dimension of our input motion sequence. Subsequently, a series of m blocks operate over the temporal dimension of the data. Each block consists of an adaptive layer normalization (AdaLN) conditioning module [11] and a fully connected layer operating on the temporal dimension. The AdaLN module calculates its shift,

scale, and gate parameters using the embedded sign label as input. The final step involves applying a fully connected layer in the spatial dimension to produce the output.

ConvAtt Our SLR model combines convolutions and self-attention mechanisms to extract local and global input information. We employ 1D convolutional layers to leverage the local information of adjacent keypoints through its sliding window operation. Conversely, a self-attention module allows our model to discern position-wise local features and engage in content-based global interactions. Additionally, Efficient Channel Attention (ECA) [15] captures cross-channel interactions without performance degradation.

The model architecture begins with a linear encoding layer for initial processing. This is followed by a sequence of ConvAtt blocks, each comprising three convolutional modules and a self-attention module. Within each convolutional module, a depthwise convolutional layer is applied, along with an Efficient Channel Attention (ECA) [15] module and Batch Normalization . The inputs and outputs of each convolutional module are managed by fully-connected layers with Gaussian Error Linear Unit (GELU) activation functions and incorporate residual connections.

4 Experiments and results

This section presents a comprehensive evaluation of our ConvAtt model for Sign Language Recognition (SLR) on the LSFB-ISOL dataset. We conducted extensive experiments to assess the model's performance and analyze the impact of various regularization and data augmentation methods. Throughout these experiments, we use Top-1 and Top-10 accuracy as our primary evaluation metrics.

4.1 Setup configuration

For the optimization of our model, we employ RAdam with 1×10^{-3} weight decay [8] complemented by the lookahead strategy [17] which facilitates faster convergence and reduced variance by using a second set of parameters that are periodically updated using k future steps. The model undergoes training over 50 epochs with a batch size of 2048. Regarding hyperparameters, we opt for an embedding size of 64, 128 channels for the convolutional layers, and a depth of 8 blocks. We initialize the weights using orthogonal initialization. Throughout the training phase, we apply categorical cross-entropy as our loss function.

4.2 Regularization

The inherent low variance within the domain predisposes models to overfitting. To mitigate this, we have implemented a variety of regularization techniques. Dropout and DropPath [7] are employed within each convolutional module to

inhibit the co-adaptation of activations, thereby diminishing the likelihood of the network's reliance on a limited subset of weights for its predictions. Additionally, we utilize an Exponential Moving Average (EMA) [9] as a teacher model, which serves to enhance generalization and robustness while attenuating parameter noise.

There is evidence that large learning rates regularize the training improving generalization [14] . To achieve "super-convergence" [14] we employed a one-cycle learning rate scheduler. We initiate training with a high learning rate γ, escalating from 3×10^{-3} to 1×10^{-2} within a brief span of epochs. Subsequently, γ is diminished post-peak to a nadir of 4×10^{-4} in a singular cycle.

4.3 Data augmentation

To enhance our training data, we implemented a series of data augmentation techniques. Affine transformations were utilized to efficiently flip, scale, and rotate the landmarks. Specifically, we applied horizontal flips with a probability of 50%, scaling by factors ranging from 0.95 to 1.05, and rotations within ± 5 degrees.

To further prevent overfitting and bolster the model's robustness—thereby reducing its dependence on specific keypoints or frames—we employed random masking to frames and random cutout [2] to blocks of 9 adjacent keypoints. Each of these transformations was applied with a probability of 20%.

4.4 Synthetic pretrain

We applied synthetic data to pretrain our model. Given an input space $P := (x_1, y_1), ..., (x_n, y_n))$ where x is the list of poses that compose each sign. To train our model we process it to the form $X := (v_1, t_1, y_1), ..., (v_n, t_n, y_n))$ so that given a target timestep and a total number of frames f, $v_i = x_{i,0..T}$ is the input pose, $t_i = x_{i,T..f} - x_{i,T}$ the target residuals and y the sign class. We train our CSiMLPe generator $\mathcal{G}$ to predict a new target $\mathcal{G}(x, y) = t'$. Similarly to SiMPLe, our objective function $\mathcal{L}$ includes a $\mathcal{L}_{\in} - norm$ minimization term between the ground-truth t and t' and a $\mathcal{L}_{\in} - norm$ minimization of the predicted motion m' and the ground-truth one m, where $m_t = x_{t+1} - x_t$ represents the velocity between frames x_t and x_{t+1}. We use $\mathcal{G}$ to generate z new samples by randomly sampling v' and y' from X in a stratified way to obtain $X' := (v'_1, t'_1, y'_1), ..., (v'_z, t'_z, y'_z))$. We then concatenate v' and t' into x' to obtain a new input space $P' := (x'_1, y'_1), ..., (x'_n, y'_n))$ that we use to pretrain ConvAtt for a reduced number of epochs. Afterwards, we fine-tune the model with P following the normal procedure.

5 Results

We conducted a comprehensive series of experiments to evaluate the performance of our ConvAtt model and assess the impact of various data augmentation and regularization techniques.

Table 1 presents a comparison of our best-performing ConvAtt model against current state-of-the-art models on the LSFB-ISOL dataset. For all experiments, we maintained consistent data preprocessing steps and utilized the same held-out test set for evaluation. We employed Top-1 and Top-10 accuracy metrics as our primary performance indicators, allowing for direct comparison with existing benchmarks. Our model achieves competitive performance, with a Top-1 accuracy of 49.99% and a Top-10 accuracy of 83.19%. While these results are slightly lower than the previous state-of-the-art in Top-1 accuracy, they are comparable in Top-10 accuracy.

In the following subsections, we present detailed analyses of our regularization techniques and data augmentation methods, providing insights into their individual and combined effects on model performance.

Model	**Top-1**	**Top-10**
ConvAtt-S [Ours]	42.7	81.9
ConvAtt [Ours]	49.99	83.19
LSFB classifier [4]	54.4	83.4

Table 1: Comparison of SLR results on the LSFB-ISOL dataset. The table depicts each model's Top-1 and Top-10 scores. Our best model was trained using OneCycle, Drop Path, Look Ahead and Frame Masking. ConvAtt-S is a smaller version of ConvAtt with 4 blocks.

5.1 Regularization results

To assess the impact of various regularization techniques on our ConvAtt model's performance, we conducted an ablation study. The results of this study are presented in Table 2.

We first evaluated Dropout, a widely used regularization technique. This method achieved a Top-1 accuracy of 43.47% and a Top-10 accuracy of 77.84%. We used this technique as our baseline model since it consistently improved the performance across our experiments. Next, we tested Drop Path, which demonstrated better performance with a Top-1 accuracy of 45.52% and a Top-10 accuracy of 80.78%. This improvement indicates that Drop Path may be more effective than using only standard Dropout for our specific architecture and dataset, possibly due to its ability to prevent overfitting in deeper layers. The Exponential Moving Average (EMA) technique proved to be particularly effective, yielding a Top-1 accuracy of 49.55% and a Top-10 accuracy of 80.82%. This significant improvement over both Dropout and Drop Path underscores the value of EMA in stabilizing model predictions and enhancing generalization. Finally, we implemented the OneCycle learning rate scheduler, which produced

the best results among all tested regularization techniques. With a Top-1 accuracy of 51.137% and a Top-10 accuracy of 81.66%, OneCycle demonstrated its effectiveness in improving both model performance and training efficiency. We also experimented with combinations of these techniques. However, we found that a careful balance of regularization is crucial to prevent hindering the model performance.

Regularization	**Top-1**	**Top-10**
Baseline (Dropout)	43.47	77.84
Drop Path	45.52	80.78
EMA	46.17	80.68
OneCycle	51.137	81.66
OneCycle + Drop Path	48.53	82.15
OneCycle + EMA	49.55	80.82

Table 2: Comparison of SLR results using different regularization methods on the LSFB-ISOL dataset. The table depicts each model's Top-1 and Top-10 scores. Dropout is used as the baseline model, with every other method being applied in addition to it.

5.2 Data augmentation results

We conducted experiments using each data augmentation method individually to analyze the impact of each one in the training of our model. The results are presented in Table 3.

Our baseline model achieved a Top-1 accuracy of 43.47% and a Top-10 accuracy of 77.84%. This serves as our reference point for assessing the impact of each augmentation technique. Horizontal flipping of the input data showed a modest improvement, increasing the Top-1 accuracy to 44.03% and the Top-10 accuracy to 78.37%. Scaling augmentation further improved performance, with a Top-1 accuracy of 45.06% and a Top-10 accuracy of 78.45%. This technique likely helps the model become more robust to variations in the size of the signing space or distance from the camera. Rotation augmentation yielded the best results among the geometric transformations, achieving a Top-1 accuracy of 45.38% and a Top-10 accuracy of 79.10%. This improvement indicates that slight rotations help the model handle variations in signer orientation or camera angle. Frame masking proved to be particularly effective for Top-10 accuracy, reaching 80.03% while maintaining a competitive Top-1 accuracy of 45.21%. This suggests that randomly masking frames encourages the model to learn more robust temporal features. Interestingly, random cutout resulted in a significant performance drop, with Top-1 and Top-10 accuracies of 37.72% and 72.83% respectively. This unexpected result suggests that removing spatial information

from the input may be too disruptive for sign language recognition, where the spatial relationships between keypoints are crucial. The most striking improvement came from synthetic pretraining, which achieved the highest performance with a Top-1 accuracy of 50.54% and a Top-10 accuracy of 81.12%. This significant boost in performance demonstrates the value of leveraging synthetic data to pretrain the model, likely helping it to learn more general and robust features before fine-tuning on the real dataset.

Model	**Top-1**	**Top-10**
Baseline	43.47	77.84
Flip	44.03	78.37
Scale	45.06	78.45
Rotate	45.38	79.10
Frame Mask	45.21	80.03
Random Cutout	37.72	72.83
Synthetic Pretrain	50.54	81.12

Table 3: Comparison of SLR results using different data augmentation methods on the LSFB-ISOL dataset. The table depicts each model's Top-1 and Top-10 scores.

6 Conclusions & Future Work

Our research yielded a state-of-the-art SLR model (ConvAtt), achieving 49.99% top-1 and 83.19% top-10 accuracy on the LSFB dataset. Ablation studies revealed that all tested regularization methods improved performance, underscoring the importance of stabilized training and learning rate management. Data augmentation analysis demonstrated the efficacy of most geometric transformations and frame masking, while highlighting the need for careful technique selection. Notably, our novel conditional pose sequence generation model (CSiMLPe) enabled the creation of a synthetic dataset, which, when used for pretraining, proved to be the most effective data augmentation method.

7 Future Work

In our future work, we plan to incorporate new regularization techniques and refine the data augmentation methods presented in this study to further boost the model's accuracy. We aim to explore on the possible combination of regularization and data augmentation techniques that improve the model performance without incurring in an over regularization.

References

1. Bianco, P.D., Ríos, G., Ronchetti, F., Quiroga, F., Stanchi, O., Hasperué, W., Rosete, A.: Lsa-t: The first continuous argentinian sign language dataset for sign language translation (2022)
2. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout (2017)
3. Fink, J., Frénay, B., Meurant, L., Cleve, A.: Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2021)
4. Fink, J., Poitier, P., André, M., Meurice, L., Frénay, B., Cleve, A., Dumas, B., Meurant, L.: Sign language-to-text dictionary with lightweight transformer models. In: Elkind, E. (ed.) Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. pp. 5968–5976. International Joint Conferences on Artificial Intelligence Organization (8 2023). https://doi.org/10.24963/ijcai.2023/662
5. Gan, S., Yin, Y., Jiang, Z., Wen, H., Xia, K., Xie, L., Lu, S.: Signgraph: A sign sequence is worth graphs of nodes. In: CVPR2024
6. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: A simple baseline for human motion prediction (2022)
7. Larsson, G., Maire, M., Shakhnarovich, G.: Fractalnet: Ultra-deep neural networks without residuals (2017)
8. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond (2021)
9. Morales-Brotons, D., Vogels, T., Hendrikx, H.: Exponential moving average of weights in deep learning: Dynamics and benefits. Submitted to Transactions on Machine Learning Research (2023)
10. Mumuni, A., Mumuni, F., Gerrar, N.K.: A survey of synthetic data augmentation methods in machine vision. Machine Intelligence Research (Mar 2024). https://doi.org/10.1007/s11633-022-1411-7
11. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023)
12. Sinha, R., Azadpour, M.: Employing deep learning model to evaluate speech information in acoustic simulations of auditory implants. Research square (06 2023). https://doi.org/10.21203/rs.3.rs-3085032/v1
13. Skobov, V., Bono, M.: Making body movement in sign language corpus accessible for linguists and machines with three-dimensional normalization of MediaPipe. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 1844–1855. Association for Computational Linguistics (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.124
14. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates (2018)
15. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks (2020)
16. Wong, R., Camgoz, N.C., Bowden, R.: Sign2GPT: Leveraging large language models for gloss-free sign language translation. In: The Twelfth International Conference on Learning Representations (2024)
17. Zhang, M.R., Lucas, J., Hinton, G., Ba, J.: Lookahead optimizer: k steps forward, 1 step back (2019)