- ORIGINAL ARTICLE -

ConvAtt Network: A Low Parameter Approach For Sign Language Recognition

Red ConvAtt: Un Acercamiento Con Bajos Parámetros Para El Reconocimiento De Lengua De Señas

Gastón Rios[1,3], Pedro Dal Bianco[1,3], Franco Ronchetti[1,2], Facundo Quiroga[1,2], Santiago Ponte Ahon[1,3], Oscar Stanchi[1,4], and Waldo Hasperué[1,2]

[1]*Instituto de Investigación en Informática LIDI, Universidad Nacional de La Plata, Argentina*
{grios, pdalbianc, fronchetti, fquiroga, sponte, ostanchi, whasperue}@lidi.info.unlp.edu.ar
[2]*Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA), Argentina*
[3]*Becario Doctoral, Universidad Nacional de La Plata, Argentina*
[4]*Becario Doctoral, CONICET, Argentina*

Abstract

Despite recent advances in Large Language Models in text processing, Sign Language Recognition (SLR) remains an unresolved task. This is, in part, due to limitations in the available data. In this paper, we investigate combining 1D convolutions with transformer layers to capture local features and global interactions in a low-parameter SLR model. We experimented using multiple data augmentation and regularization techniques to categorize signs of the French Belgian Sign Language. We achieved a top-1 accuracy of 42.7% and a top-10 accuracy of 81.9% in 600 different signs. This model is competitive with the current state of the art while using a significantly lower number of parameters.

Keywords: Deep Learning, Sequence Classification, Sign Language Recognition, Unbalanced Data

Resumen

A pesar de los avances recientes en grandes modelos de lenguaje para el procesamiento de texto, el Reconocimiento de Lenguas de Señas (SLR por sus siglas en inglés) aún es una tarea sin resolver. Esto es, en parte, debido a las limitaciones en los datos disponibles. En este artículo, investigamos cómo combinar convoluciones 1d con capas transformer para capturar las características locales y las interacciones globales utilizando un modelo de SLR de pocos parámetros. Experimentamos usando múltiples técnicas de aumento de datos y regularización para categorizar señas de la lengua de señas belga-francesa. Como resultado, obtuvimos una exactitud top-1 de 42.7% y top-10 de 81.9% en 600 señas diferentes. Este modelo es competitivo con el estado del arte actual, utilizando una cantidad significativamente menor de parámetros.

Palabras claves: Aprendizaje Profundo, Clasificación De Sequencias, Reconocimiento De Lenguas De Señas, Datos Desbalanceados.

1 Introduction

Sign language is a visual language expressed through hand movements, facial expressions, and body cues. Deaf individuals primarily use it, but it is also utilized by non-deaf people in certain contexts, such as medical sign language. More than 300 different sign languages exist worldwide [1]. These languages are usually not mutually intelligible. For instance, a person fluent in Argentinian Sign Language cannot communicate with someone fluent in French Belgian Sign Language. Furthermore, Sign Language is not commonly used by individuals who are not part of the Deaf community or do not regularly interact with it. This underscores the importance of systems for SLR in facilitating communication and interaction with technology for Sign Language users.

There have been multiple approaches to SLR systems with promising results [2, 3, 4, 5, 6]. These works mainly utilize deep learning through computer vision or multi-modal processing with images, videos and poses of the signers. However, the limited availability and diversity of sign language data can constrain the accuracy of these deep learning models, which require a large quantity of high-quality samples. Compared to voice recognition data, sign language data sources are fewer, making data collection a challenging, time-consuming, and expensive task. In addition, a unique dataset must be created for each sign language due to their lack of mutual intelligibility, which limits the possibility of combining multiple data sources. While speech processing models achieve human-like recognition [7, 8] and generation [9, 10] using hundreds of thousands of hours of voice recordings, sign language datasets rarely surpass a hundred hours [11].

In this paper, we explore various regularization and data augmentation techniques to enhance the accuracy of deep learning SLR models[1]. We utilized pose data from the French Belgian Sign Language Isolated (LSFB-ISOL) dataset to train a compact 1D convolutional neural network equipped with transformer layers, we call this model ConvAtt. Figure 1 shows a graphical representation of our pipeline.

A description of the LSFB-ISOL dataset and the preprocessing used is contained in Section 3. A detailed definition of the model architecture is provided in Section 4.1, while the hyperparameters used during training are outlined in Section 4.2. We implemented spatial and temporal data augmentations in conjunction with multiple regularization techniques. Sections 4.3 and 4.4 describe these data augmentation and regularization techniques respectively. Section 5 elaborates on the conclusions derived from the results obtained.

1.1 Contribution

Our research demonstrates that by using regularization and data augmentation it is possible to develop a compact Sign Language Recognition (SLR) model with results similar to state-of-the-art. Specifically, our key contributions are:

- Development of ConvAtt, a compact Sign Language Recognition (SLR) model that combines 1D convolutions and self-attention mechanisms. This model achieves competitive results (42.7% top-1 accuracy and 81.9% top-10 accuracy on the LSFB dataset) while using significantly fewer parameters than other leading models.
- Implementation and evaluation of various regularization techniques, including Dropout, DropPath and OneCycle, to mitigate overfitting.
- Exploration of data augmentation techniques for SLR, including affine transformations (flipping, scaling, rotation) and masking methods (random frame masking, random cutout). The paper demonstrates that poorly tuned data augmentation can actually harm model performance.
- Proposal of the ConvAtt model and associated pipeline as a baseline for future SLR research, with the code made publicly available.

2 Related Works

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 [12], depth maps [13], and poses of the signer's hands, body, and face [4, 14], typically in a multi-modal approach [15]. 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 [2, 3] and transformer architectures [4, 5], although combinations of the two architectures have been effectively implemented [6, 16]. These models are widely used in vision and natural language processing [17, 18, 19]. 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 like Mediapipe [20] and Openpose [21] has shown great success in improving performance [6]. 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 [22]. Graph representation of sign sequences has also been proposed, obtaining state-of-the-art results [2]. In addition to these methods, data augmentation has proven to be an essential tool to increase the robustness of the model and reduce overfitting [23]. Furthermore, data augmentation can diminish the representation distance between video and text data, easing data scarcity [24].

3 Dataset

The French Belgian Sign Language Isolated (LSFB-ISOL) dataset [25] is built upon the LSFB Corpus. It spans 25 hours of videos and poses of continuous isolated signs performed by 85 different signers. An example of a frame extracted from the dataset can be seen in Figure 2.

In this paper, we focus solely on the poses, as they reduce domain complexity and enable faster processing times for models. Pose data was extracted using the MediaPipe tool and subsequently normalized. This resulted in 478 3D landmarks for the face, 33 for the body, and 21 2D landmarks for the hands. We selected a subset of 45 face landmarks of the eyes and mouth to reduce data redundancy. This totals 99 landmarks for our input.

After filtering out signs with less than 20 samples or more than 60 frames, the dataset contained 52,350 sign poses across 610 classes. We isolated 10% of the samples for testing. Since the dataset is highly unbalanced, all classes were oversampled to contain the same number of samples as the most numerous class.

For processing, keypoints were formatted in a single dimension as $\{x_i, y_i, z_i | i = 0, 1, 2, ..., n\}$ where n represents the number of keypoints in each sample. Additionally, each sample incorporates a temporal dimension, maintaining an equivalent count of keypoints. To facilitate uniform input size, we randomly sam-

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

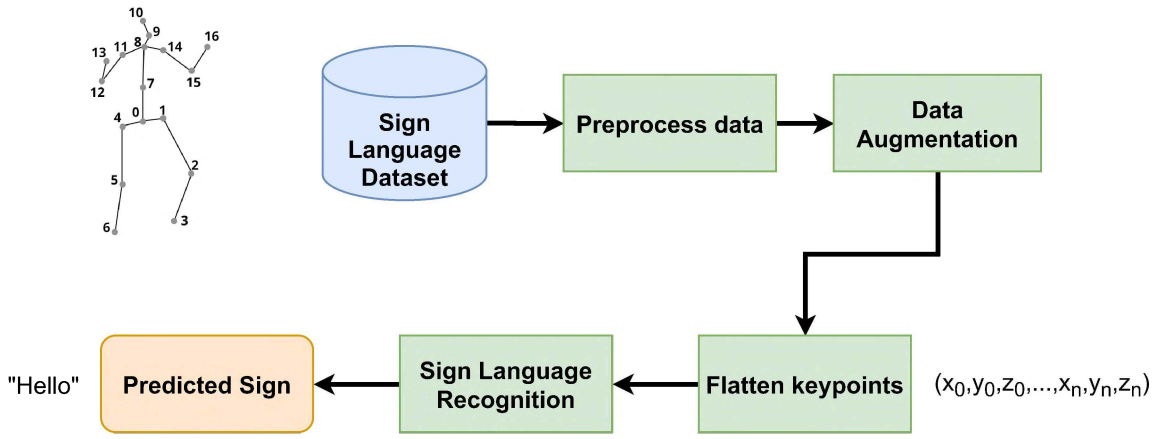

Figure 1: Pipeline of the SLR process utilized. We first preprocess the data by removing samples and keypoints that don't meet our criteria, adjusting the data to a fixed frame length through cropping or padding, and normalizing it. Next, we apply a series of data augmentation techniques to each sample. The data is then flattened to a single dimension. Finally, this one-dimensional data is fed into the ConvAtt model, which makes the final prediction.

ple 30 contiguous frames from each clip. For clips comprising fewer than 30 frames, we employ circular padding to extend the sequence. This padding technique involves using the initial values of the dimension to pad the terminal portion and vice versa, ensuring continuity and completeness of the data.

Figure 2: Frame from the LSFB dataset with its extracted pose.

4 Model

This section outlines the ConvAtt model and its training process. We detail the model architecture and compare the outcomes obtained with various hyperparameter configurations. Additionally, we assess each regularization method and the enhancements each contributes.

4.1 Model architecture

Our model, summarized in Figure 3, combines convolutions and self-attention mechanisms to extract local and global input information. We employ 1D depthwise convolutional layers to leverage the local information of adjacent keypoints through its sliding window operation. Conversely, a self-attention module [18] allows our model to discern position-wise local features and engage in content-based global interactions.

The initial processing of our model's input is through a linear encoding layer. 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) [26] module and Batch Normalization [27]. The ECA module introduces channel attention into the model efficiently. The inputs and outputs of each convolutional module are managed by fully-connected layers with Gaussian Error Linear Unit (GELU) activation functions [28] and incorporate residual connections.

Given the necessity for compact and rapid models in real-life SLR tasks, our model is designed with 538,617 parameters. This model uses fewer parameters than other leading models, which often count in the millions, yet it achieves comparable results [5, 6, 4].

4.2 Setup configuration

For the optimization of our model, we employ RAdam with 1×10^{-3} weight decay [29] complemented by the lookahead strategy [30] which facilitates faster convergence and reduced variance by using a second set of parameters that are periodically updated using k future steps. To achieve "super-convergence" [31] 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. The model undergoes training over 50 epochs with a batch size of 1024. Regarding hyperparameters, we opt for an embedding size of 32, 128 channels for the convolutional layers, and a depth of 4 blocks. We initialize the weights us-

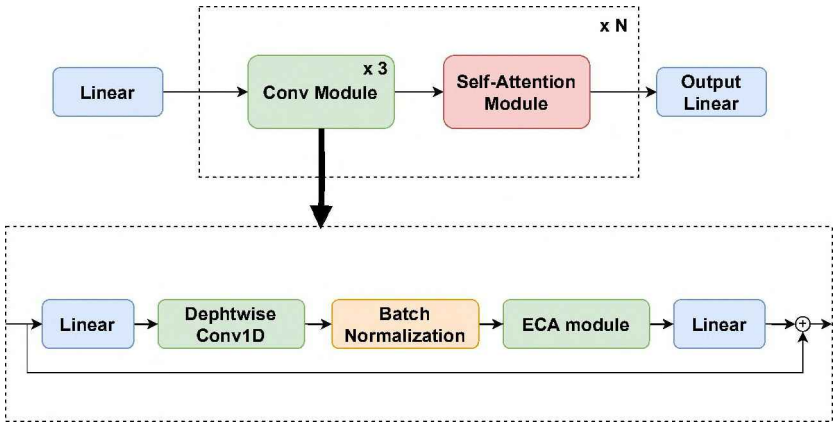

Figure 3: Our model is composed of N ConvAtt blocks with 3 convolutional modules and a self-attention module. These convolutional modules are structured as follows: a linear layer with GELU activation, succeeded by a depthwise 1D convolution and batch normalization, and an ECA module, culminating in a linear layer that maintains a residual connection.

ing orthogonal initialization. Throughout the training phase, we apply categorical cross-entropy as our loss function.

4.3 Regularization

The inherent low variance within the domain predisposes models to overfitting. To mitigate this, we have implemented a variety of regularization techniques. Dropout[32] and DropPath [33] 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.

4.4 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%, simulating left-handed and right-handed signers and accounting for potential mirror images in real-world applications. We also implemented scaling by factors ranging from 0.95 to 1.05, and rotations within ± 5 degrees, which helps the model adapt to variations in signer size, camera angle, and slight posture changes. These transformations introduce variability and simulate different perspectives and signer sizes, crucial for developing a robust SLR system that can generalize across diverse signing styles and recording conditions.

To further prevent overfitting and bolster the model's robustness—thereby reducing its dependence on specific keypoints or frames—we employed random frames masking and random cutout [34]. Random masking involves randomly removing entire frames from the input sequence, forcing the model to learn temporal coherence and reducing its reliance on specific frames. The random cutout technique, applied to blocks of 9 adjacent keypoints, simulates occlusions or tracking errors that might occur in real-world scenarios, such as when parts of the signer's body are temporarily obscured. Each of these transformations was applied with a probability of 20%, with the goal that the model learns to recognize signs even with partial occlusions or missing information.

5 Results

We conducted multiple experiments to analyze the performance of our model and the effect of our data augmentation pipeline. Table 1 displays the results of these experiments. All experiments were conducted using the same model backbone and hyperparameters. Also, the experiments share the same data preprocessing, using the same held-out test set for evaluation. We utilized Top-1 and Top-10 accuracy metrics to evaluate our model's effectiveness and compare it with the current state-of-the-art. To calculate the efficiency of each model we divided its accuracy by the number of parameters.

In 300 epochs, our baseline model achieved a top-1 accuracy of 35.2% and top-10 accuracy of 77.0%. With the addition of the one-cycle learning rate scheduler, our results improved to 42.7% top-1 and 81.9% top-10 accuracy. This also lets us reduce the training time significantly, requiring 50 epochs to reach convergence. These results are comparable to the current state-of-the-art while using a lower amount of model parameters and training epochs. This could be attributed to the regularizing property of higher learning rates used by the scheduler. Our best model showed higher efficiency than the state of the art, with a Top-1 efficiency of 79.3 and a Top-10 efficiency of 152.2.

We also conducted experiments with multiple data

Table 1: Comparison of SLR results on the LSFB-ISOL dataset. The table details each model's parameter count, their Top-1 and Top-10 scores, and the efficiency of each models calculated as the division between the accuracy and parameter number. It contrasts the ConvAtt model with the one-cycle learning rate scheduler (OneCycle) and data augmentation (DA) methods.

Model	Parameters	Top-1	Top-10	Top-1/#Par	Top-10/#Par
ConvAtt [Ours]	538k	35.2	77.0	65.4	143.1
ConvAtt + OneCycle [Ours]	538k	42.7	81.9	**79.3**	**152.2**
ConvAtt + OneCycle + DA [Ours]	538k	31.6	74.0	58.7	137.5
LSFB classifier [4]	782k	**54.4**	**83.4**	69.5	106.6

augmentation techniques reaching a top-1 accuracy of 31.6% and top-10 accuracy of 74.0%. This is lower than the accuracy obtained by our models trained without data augmentation, which shows that poorly defined data augmentation hyperparameters can be detrimental to the model training. This reflects the difficulty in the tuning of these hyperparameters in this domain.

6 Conclusion

We introduced ConvAtt, a model that merges 1D depthwise convolutions with self-attention layers for sequence classification. Despite using fewer parameters and training epochs, ConvAtt achieved comparable results with the current state-of-the-art SLR on the LSFB-ISOL dataset. We illustrated that improperly tuned data augmentation can negatively impact test accuracy. We propose the ConvAtt model and the associated pipeline as baselines for future SLR research, and have made the code publicly available.

7 Future Work

In our future work, we aim to develop a sign language generation model capable of creating new sequences of input poses. The goal is to train this model with both generated and real data to enhance its generalization capabilities and address the dataset's data imbalance. We plan to incorporate new regularization techniques and refine the data augmentation methods presented in this study to further boost the model's accuracy.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

The authors confirm contribution to the paper as follows: GGR: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; PDB: Conceptualization, Methodology; FR: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – review & editing; FQ: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – review & editing; OS: Methodology, Writing – review & editing; SPA: Methodology, Writing – review & editing; WH: Project administration, Resources. All authors reviewed the results and approved the final version of the manuscript.

References

[1] N. G. Education, "Sign-language," https://education.nationalgeographic.org/resource/sign-language/, 2024.

[2] S. Gan, Y. Yin, Z. Jiang, H. Wen, K. Xia, L. Xie, and S. Lu, "Signgraph: A sign sequence is worth graphs of nodes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 470–13 479.

[3] L. Jing, Y. Wang, T. Chen, S. Dora, Z. JI, and H. Fang, "Towards a more efficient few-shot learning-based human gesture recognition via dynamic vision sensors," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/0938.pdf

[4] J. Fink, P. Poitier, M. André, L. Meurice, B. Frénay, A. Cleve, B. Dumas, and L. Meurant, "Sign language-to-text dictionary with lightweight transformer models," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2023, pp. 5968–5976, aI for Good. [Online]. Available: https://doi.org/10.24963/ijcai.2023/662

[5] R. Wong, N. C. Camgoz, and R. Bowden, "Sign2GPT: Leveraging large language models for gloss-free sign language translation," 2024. [Online]. Available: https://arxiv.org/abs/2405.04164

[6] V. Skobov and M. Bono, "Making body movement in sign language corpus accessible for linguists and machines with three-dimensional normalization of MediaPipe," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, Dec. 2023, pp. 1844–1855.

[7] I. Papastratis, "Speech recognition: a review of the different deep learning approaches," *https://theaisummer.com/*, 2021.

[8] R. Sinha and M. Azadpour, "Employing deep learning model to evaluate speech information in acoustic simulations of auditory implants," *Research square*, 06 2023.

[9] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," 2023. [Online]. Available: https://arxiv.org/abs/2306.15687

[10] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[11] P. D. Bianco, G. Ríos, F. Ronchetti, F. Quiroga, O. Stanchi, W. Hasperué, and A. Rosete, "Lsa-t: The first continuous argentinian sign language dataset for sign language translation," in *Advances in Artificial Intelligence – IBERAMIA 2022*, A. C. B. Garcia, M. Ferro, and J. C. R. Ribón, Eds. Cham: Springer International Publishing, 2022, pp. 293–304.

[12] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015, pose & Gesture. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314215002088

[13] M. Oszust and J. Krupski, "Isolated sign language recognition with depth cameras," *Procedia Computer Science*, vol. 192, pp. 2085–2094, 2021, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921017129

[14] A. Tunga, S. V. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using gcn and bert," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 31–40.

[15] R. Pathan, M. Biswas, S. Yasmin, M. Khandaker, M. Salman, and A. Youssef, "Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network," *Scientific Reports*, vol. 13, 10 2023.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[19] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[20] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf

[21] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 01, pp. 172–186, jan 2021.

[22] Y. Li, H. Chen, G. Feng, and Q. Miao, "Learning robust representations with information bottleneck and memory network for rgb-d-based gesture recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 20 968–20 978.

[23] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[24] J. Ye, W. Jiao, X. Wang, Z. Tu, and H. Xiong, "Cross-modality data augmentation for end-to-end sign language translation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 558–13 571. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.904

[25] J. Fink, B. Frénay, L. Meurant, and A. Cleve, "Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition," in *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN 2021)*. IEEE Computer Society Press, 2021.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 11 531–11 539. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01155

[27] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" *Advances in neural information processing systems*, vol. 31, 2018.

[28] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023. [Online]. Available: https://arxiv.org/abs/1606.08415

[29] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020.

[30] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, *Lookahead optimizer: k steps forward, 1 step back*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[31] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[33] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," 2017. [Online]. Available: https://arxiv.org/abs/1605.07648

[34] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017. [Online]. Available: https://arxiv.org/abs/1708.04552

Citation: G. Ríos, P. Dal Bianco, F. Ronchetti, F. Quiroga, S. Ponte Ahon, O. Stanchi and W. Hasperué. *ConvAtt Network: A Low Parameter Approach For Sign Language Recognition*. Journal of Computer Science & Technology, vol. 24, no. 2, pp. 104–110, 2024.
DOI: 10.24215/16666038.24.e10.
Received: April 15, 2024 **Accepted:** September 9, 2024.