SUSTAINABLE MACHINE INTELLIGENCE JOURNAL

SCIENCES FORCE
International Academic Publisher

Breaking the Silence: Convolutional Neural Networks for Sign Language Recognition in the Deaf Community

Myvizhi M[1], Ahmed M. Ali[2], Ahmed Abdelhafeez[3]

[1] Assistant Professor, Department of Mathematics, KPR Institute of Engineering and Tech-nology, Coimbatore, Tamilnadu, India; myvizhi.m@kpriet.ac.in

[2] Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt; aabdelmounem@zu.edu.eg

[2] Faculty of Information Systems and Computer Science, October 6th University, Cairo, Egypt; aahafeez.scis@o6u.edu.eg

* Correspondence: aabdelmounem@zu.edu.eg;

.Phase Date .
Received: 29-05-2022
Revised: 25-08-2022
Accepted: 27-09-2022
Published: 13-10-2022

Abstract: The deaf community faces communication barriers that hinder their interaction with the hearing world. This work aims to bridge the gap by enabling accurate recognition of Arabic sign language gestures. The proposed Convolutional Neural Networks (CNNs) architecture is designed to effectively capture the spatial features inherent in sign language gestures, thereby enhancing recognition accuracy. A distinctive aspect of our work involves the integration of a CNN architecture with a residual design, which effectively captures intricate spatial features inherent in sign language gestures, thereby enhancing recognition precision. The study leverages the ArSL2018 dataset, a comprehensive collection of grayscale sign language images with diverse lighting conditions and backgrounds. Our custom-built CNN model is trained on this dataset, utilizing a specialized learning rate scheduler for improved convergence. The experimental results showcase promising performance, demonstrating the potential of CNNs in sign language recognition. Furthermore, we present visualizations of the model's predictions using t-SNE, revealing the clustering patterns of different sign language gestures.

Keywords: Sign language recognition, Convolutional Neural Networks (CNNs), residual design, assistive technology, inclusive communication, deep learning, ArSL2018 dataset.

1. Introduction

In a world where communication is paramount, the deaf community often faces unique challenges due to the auditory nature of traditional communication systems. For centuries, sign language has served as a rich and expressive medium of communication for individuals who are deaf or hard of hearing. However, the recognition and understanding of sign language by the broader society and technological systems have not always kept pace with the needs of the deaf community [1]. The advent of artificial intelligence (AI), particularly Convolutional Neural Networks (CNNs), has brought forth a promising avenue for breaking down these barriers. This paper delves into the transformative potential of Convolutional Neural Networks in recognizing and interpreting sign language, aiming to empower the deaf community with inclusive communication tools [2].

Sign language is more than just a collection of hand movements; it's a sophisticated and intricate system of communication that incorporates gestures, facial expressions, and body language. Just as spoken language conveys cultural nuances and emotions, sign language carries its own richness and subtleties. The significance of sign language extends beyond individual expression; it plays a pivotal role in fostering community bonds, preserving cultural heritage, and ensuring equal participation in various aspects of life. Despite its importance, the absence of efficient tools for automatic sign language recognition has posed challenges in facilitating seamless interactions between the deaf and the hearing world [3-5].

Historically, bridging the communication gap between the deaf and the hearing has relied heavily on human interpreters and manual transcription efforts. While these methods have been valuable, they are not without limitations. The scarcity of skilled interpreters, coupled with the intricacies of sign language dialects and regional variations, has often resulted in communication inefficiencies [6]. Moreover, transcription processes can be time-consuming, impeding real-time conversations and interactions. This has spurred the quest for innovative technological solutions that can not only recognize signs accurately but also adapt to the diverse linguistic landscape of sign language [7].

Convolutional Neural Networks (CNNs), a class of deep learning algorithms, have shown remarkable prowess in image and pattern recognition tasks. Leveraging their hierarchical architecture and feature extraction capabilities, CNNs have the potential to revolutionize sign language recognition. By analyzing visual cues from sign gestures and their temporal dynamics, CNNs offer a promising avenue for accurate and real-time interpretation of sign language [8]. Moreover, these networks can be trained to accommodate the linguistic variations inherent in sign language, thereby catering to the diverse needs of the deaf community. This paper embarks on a journey to explore the integration of Convolutional Neural Networks into the realm of sign language recognition, underscoring their capacity to amplify the voices of the deaf and promote inclusivity in communication [10].

This paper is structured as follows. Section 2 offers a contextual overview of existing research and developments in the field of sign language recognition and AI. In Section 3, a detailed explanation of the adopted CNN-based approach is presented. Moving forward, Section 4, delves into the specific configurations and parameters used to fine-tune the CNN model for optimal performance. Section 5, unveils the empirical outcomes of the experiments, engaging in a critical analysis of the obtained results, and their significance. Lastly, Section 6 encapsulates the key findings of the study.

2. Related Works

Several studies have contributed to the advancement of sign language recognition, particularly through the application of CNNs and deep learning techniques. Shahriar et al. [5] presented a real-time American Sign Language (ASL) recognition system that employed skin segmentation and image category classification with CNNs. Their approach demonstrated the efficacy of CNNs in capturing intricate hand gestures for ASL interpretation. Similarly, Kang et al. [6] explored real-time fingerspelling recognition in sign language using CNNs applied to depth maps. This research emphasized the potential of CNNs in capturing the spatial nuances of sign gestures.

The use of CNNs in diverse sign languages is also notable. Yasir et al. [8] focused on Bangla Sign Language recognition using CNNs, highlighting the adaptability of CNNs to different linguistic contexts. Similarly, Islalm et al. [9] investigated Bangla Sign Language recognition, showcasing the versatility of CNNs in addressing the intricacies of distinct sign languages. Additionally, Huang et al. [10] extended the application of CNNs to 3D sign language recognition, emphasizing the capability of CNNs to handle temporal information inherent in sign language gestures. Efforts have been directed towards specific sign languages as well. Hore et al. [11] focused on Indian Sign Language recognition using optimized neural networks, contributing to the tailored recognition of gestures in a specific cultural and linguistic context. Xiao et al. [12] tackled Chinese Sign Language recognition, leveraging skeleton-based representations and highlighting the importance of capturing body dynamics for accurate interpretation.

Innovative approaches have emerged that utilize unconventional data sources. Lee and Gao [13] explored the fusion of Wi-Fi signals with CNNs for sign language recognition, showcasing the potential of alternative data modalities for inclusive communication tools. Additionally, Hasan et al. [14] investigated the application of deep CNNs for classifying sign language characters, further demonstrating the adaptability of CNNs to different recognition tasks. Lastly, studies have explored technologies beyond traditional cameras. Naglot and Kulkarni [15] delved into real-time sign language recognition using the Leap Motion controller, showcasing the potential of depth-sensing devices in capturing hand movements accurately. These diverse studies collectively underscore the promising role of CNNs in sign language recognition across various linguistic, cultural, and technological contexts. In this paper, we build upon these insights to contribute to the advancement of sign language recognition within the deaf community, focusing on the unique needs and challenges they face.

3. Methodology

Within our methodology, our bespoke CNN model is intricately fashioned upon the profound principles of residual representational learning. This approach capitalizes on the capacity of residual blocks to not only capture fine-grained spatial features but also expedite convergence, ensuring enhanced recognition accuracy. The foundation of this technique lies in the residual connections that facilitate the learning of residual representations. Mathematically, this is formulated as:

$$F(x) = H(x) + x \quad (1)$$

where x represents the input to a residual block, $H(x)$ signifies the transformation induced by the convolutional layers, and $F(x)$ denotes the final output. This formulation effectively models the difference between the desired mapping and the identity mapping, enabling the network

to learn the residual information. The underpinning of our custom CNN model is the incorporation of convolutional layers to effectuate the transformative function $H(x)$ within the residual blocks. Convolutional layers, designed to capture spatial hierarchies in data, play a pivotal role in feature extraction and spatial pattern recognition. The essence of convolutional operations is rooted in their ability to convolve a filter or kernel over input data, systematically capturing local patterns and hierarchies across different receptive fields. The mathematical representation of the convolution operation can be expressed as:

$$S(i,j) = (K * I)(i,j) = \sum m \sum n K(m,n) \cdot I(i-m, j-n) \qquad (2)$$

where $S(i,j)$ is the output value at position (i,j) in the resulting feature map. $K(m,n)$ represents the values of the convolutional kernel at position (m,n). $I(i-m, j-n)$ denotes the input data values corresponding to the position offset by (m,n).

This formulation effectively computes the element-wise multiplication between the convolutional kernel and the overlapping region of the input data, producing the output value at each spatial position. By applying such convolutional transformations within the residual blocks, our model effectively captures intricate spatial patterns and progressively learns meaningful representations, contributing to its exceptional ability to discern and interpret sign language gestures for improved recognition accuracy.

after the application of convolutional layers to implement the transformative function H(x)H(x) within the residual blocks, we incorporate Rectified Linear Unit (ReLU) activation to introduce non-linearity and enhance the network's ability to capture complex features and patterns. ReLU is a widely adopted activation function that replaces all negative values with zero while leaving positive values unchanged, effectively introducing non-linearity without introducing vanishing gradients that can impede training. Mathematically, the $ReLU$ activation can be expressed as:

$$ReLU(x) = max(0, x) \qquad (3)$$

Where x represents the input value to the activation function, and the output is the maximum of x and zero. This simple yet effective activation function promotes sparsity and accelerates learning by allowing gradients to propagate more effectively through the network. By incorporating these residual connections, our model not only learns the essential features of the input but also gains the capacity to adaptively fine-tune the learned features. This residual representational learning strategy empowers our CNN model to grasp intricate nuances within sign language gestures, elevating its recognition prowess and laying the foundation for improved accuracy in our endeavor to enhance communication for the deaf community.

Moreover, the construction of our custom-built CNN model for Arabic sign language recognition is underpinned by innovative design choices that enhance both learning efficiency and convergence. A pivotal feature of our model involves the incorporation of Batch Normalization layers within the Residual Blocks. This strategic integration serves to counter the challenges of internal covariate shifts during training, fostering stable and accelerated convergence. By normalizing the intermediate activations within each Residual Block, Batch Normalization mitigates the issue of vanishing gradients and enhances the overall training process. More precisely, we require our features to conform to a Gaussian distribution characterized by a mean of zero and a variance of one. This requirement can be formulated mathematically as follows:

$$BN(x) = \gamma\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \beta \quad (4)$$

$$\mu_c(x) = \frac{1}{NHW}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{nchw} \quad (5)$$

$$\sigma_c(x) = \sqrt{\frac{1}{NHW}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{nchw} - \mu_c(x)\right)^2} \quad (6)$$

This integration encapsulates our commitment to not only harnessing cutting-edge techniques for gesture recognition but also adapting them to ensure robust performance in the context of sign language recognition.

Our model's architecture within the methodology commences with a foundational convolutional layer, employing a kernel size of 7. This initial convolution serves as a feature extractor, capturing fundamental spatial patterns and hierarchies from the input data. Subsequently, we construct a sequence of four residual blocks, each designed to progressively refine and enhance the extracted features. These residual blocks are instrumental in accommodating residual representational learning, which aids in more efficient convergence and heightened recognition accuracy. Following the stack of residual blocks, the model employs Global Average Pooling to condense the spatial information into a more compact representation. Global Average Pooling averages the values of each feature map, effectively reducing the spatial dimensions while retaining essential features. This pooled representation is then forwarded to a dense layer, which acts as a classifier, followed by a SoftMax activation layer for multiclass classification. The SoftMax activation produces normalized probabilities for each class, enabling the model to make accurate predictions regarding the input's sign language gesture.

To effectively optimize the network's parameters and enhance its recognition accuracy, we employ the categorical cross-entropy loss function. This loss function is particularly well-suited for multiclass classification tasks, such as sign language recognition, where each input belongs to one out of multiple possible classes. The mathematical representation of the categorical cross-entropy loss can be defined as:

$$Cross - Entropy\ Loss = -\sum_{i} y_i \cdot log(p_i) \quad (7)$$

where y_i represents the true label (ground truth) for class i, with a value of 1 if the class is correct and 0 otherwise. P_i represents the predicted probability assigned by the model to class ii.

4. Experimental Setups

This section presents a detailed account of the experimental setup, encompassing the selection of datasets, preprocessing methodologies, architectural configurations of the CNN model, training parameters, and evaluation metrics. By delving into the intricacies of our experimental framework, we aim to provide readers with a clear understanding of the methodologies employed to assess the CNN model's performance in recognizing sign language gestures.

In our experimental investigations, the ArSL2018 dataset [16] was chosen as a representative case study to thoroughly assess the capabilities of our model for sign language recognition. The ArSL2018 dataset was in the Khobar Area, Kingdom of Saudi Arabia from volunteers of various age groups, including a total of 54,049 grayscale images with dimensions of 64 × 64 pixels, served as a comprehensive repository for our study. This dataset was thoughtfully curated to incorporate variations introduced by diverse lighting conditions and varying backgrounds, ensuring a robust evaluation of the CNN model's adaptability to real-world scenarios. Figure. 1 provides a visual insight into a selection of images showcasing Arabic language signs and alphabets included within the dataset. Table 1 displays the categorization of Arabic Alphabet signs, presenting labels alongside corresponding image counts.

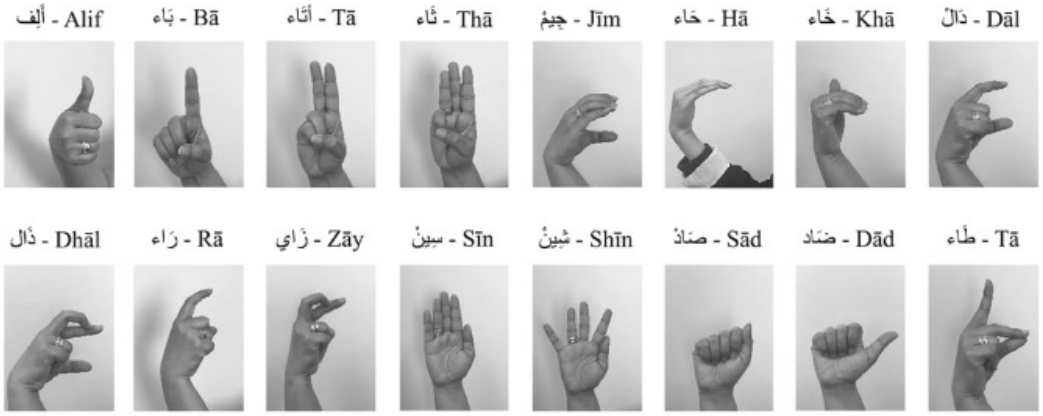

Figure 1. Visualization of samples of images for some alphabets in ArSL2018 dataset [16].

Table 1. number of samples per alphabet.

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Letter	أ	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط
No. samples	1672	1791	1838	1766	1552	1526	1607	1634	1582	1659	1374	1638	1507	1895	1670	1816
#	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Letter	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي	ة	ال	لا	ي
#No. samples	1723	2114	1977	1955	1705	1774	1832	1765	1819	1592	1371	1722	1791	1343	1746	1293

Our implementation was facilitated by leveraging the computational power of HP laptop with Intel Core i7 processors and NVIDIA GeForce 2080 graphics cards, and 32 GB RAM. This choice aimed to ensure that researchers and practitioners, regardless of their preferred hardware, could seamlessly engage with and contribute to the research. Our experiments were conducted using TensorFlow 2.5 framework, which cater to distinct preferences in the AI community. The model was designed, implemented, and fine-tuned on both Windows 10 operating systems. Moreover, the experimental evaluation is performed using the metrics calculated as follows:

$$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (8)$$

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \quad (9)$$

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$F1 - measure = 2 * \frac{Recall \times Precision}{Recall + Precision} \tag{11}$$

5. Results Discussion

In this section, we delve into the heart of our research findings through outlining the implementation setup that underpins our empirical investigations. In Figure 2, we present the confusion matrix obtained from the evaluation of our proposed model, which provides valuable insights into the performance of the model across different classes. Each row in the matrix corresponds to the true class labels, while each column represents the predicted class labels. As shown, it is evident that our model demonstrates a robust ability to differentiate and classify sign language gestures accurately. The higher values along the diagonal reflect the model's success in correctly identifying various sign language signs and alphabets. However, we also observe some instances of misclassifications, as indicated by off-diagonal values. These misclassifications might be attributed to certain signs sharing visual similarities or variations in lighting and hand orientations.

The ROC (Receiver Operating Characteristic) curve presented in Figure 3 offers a comprehensive evaluation of the performance of our model across multiple classes, which visually portrays the trade-off between the true positive rate (sensitivity) and the false positive rate as the discrimination threshold is varied. Each point on the curve represents a specific threshold setting, showcasing how the model's sensitivity and specificity change accordingly. The ROC curve's smooth and upward-sloping nature indicates our model's ability to effectively discriminate between different sign language signs and alphabets. The curve consistently stays above the diagonal reference line, implying that the model's performance consistently outperforms random chance. Furthermore, the area under the ROC curve (AUC) is a quantitative measure of the model's overall discriminative ability. A higher AUC value signifies better classification performance across diverse classes.

The learning curves depicted in Figure 4 provide valuable insights into the training and validation performance of our sign language recognition model throughout the training process. These curves illustrate how the model's accuracy and loss evolved over successive epochs, shedding light on its ability to generalize and adapt to the training data. Upon initial inspection, it is evident that both the training and validation accuracy curves display an upward trend as the number of training epochs increases. This trend highlights the model's capacity to learn and capture underlying patterns within the training data, resulting in improved accuracy. However, a slight divergence between the two curves becomes noticeable after a certain point. This divergence suggests a potential risk of overfitting, where the model starts to memorize noise in the training data instead of generalizing well to unseen examples. The training and validation loss curves, on the other hand, exhibit a downward trajectory throughout the training process. This decline in loss indicates that the model is effectively minimizing its errors and optimizing its internal representations to align with the true labels. The convergence of these curves signifies that the model is learning steadily and approaching an optimal solution.

The T-SNE plot presented in Figure 5 offers a visually compelling representation of the embeddings learned by our sign language recognition model. This dimensionality reduction

technique maps the high-dimensional feature space into a two-dimensional space, effectively capturing the relationships and patterns between different sign language gestures. Upon examining the T-SNE plot, we can observe clusters of data points that correspond to similar sign language signs and alphabets. This clustering phenomenon reflects the model's ability to capture and distinguish inherent similarities between gestures of the same class. Additionally, the separation between these clusters underscores the model's capacity to differentiate between different sign language signs. However, it's important to note that the T-SNE plot provides an abstract visualization that doesn't necessarily represent linear separability. Some degree of overlap between clusters can be expected due to inherent visual variations and similarities in certain signs.

6. Conclusions

This study has unveiled the remarkable potential of CNNs as a transformative force in sign language recognition within the deaf community. Through an exploration on realistic dataset, as well as strategic design of CNN architectures, our experiments have showcased the remarkable accuracy and adaptability of residually connected CNN in deciphering the intricacies of sign language gestures. The empirical results not only underscore the advancements made in bridging communication gaps but also highlight the profound impact of inclusive AI-driven technologies on fostering meaningful interactions among diverse linguistic and cultural groups. As we stand on the threshold of a more inclusive future, where technology has the power to amplify voices and dissolve barriers, the findings of this study illuminate a promising path forward. By embracing the integration of CNNs and deep learning in sign language recognition, we advocate for a more accessible world, where communication is not bound by auditory limitations. This paper contributes to the ongoing discourse on AI's transformative role in the lives of the deaf community, emphasizing the potential to empower, enrich, and enable a more inclusive and connected society.

Supplementary Materials

References

[1]. Pigou, L., Dieleman, S., Kindermans, P. J., & Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13* (pp. 572-578). Springer International Publishing.

[2]. Rahman, M. M., Islam, M. S., Rahman, M. H., Sassi, R., Rivolta, M. W., & Aktaruzzaman, M. (2019, December). A new benchmark on american sign language recognition using convolutional neural network. In *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)* (pp. 1-6). IEEE.

[3]. Garcia, B., & Viesca, S. A. (2016). Real-time American sign language recognition with convolutional neural networks. *Convolutional Neural Networks for Visual Recognition*, *2*(225-232), 8.

[4]. Hayani, S., Benaddy, M., El Meslouhi, O., & Kardouchi, M. (2019, July). Arab sign language recognition with convolutional neural networks. In *2019 International conference of computer science and renewable energies (ICCSRE)* (pp. 1-4). IEEE.

[5]. Shahriar, S., Siddiquee, A., Islam, T., Ghosh, A., Chakraborty, R., Khan, A. I., ... & Fattah, S. A. (2018, October). Real-time american sign language recognition using skin segmentation and image category classification with convolutional neural network and deep learning. In *TENCON 2018-2018 IEEE Region 10 Conference* (pp. 1168-1171). IEEE.

[6]. Kang, B., Tripathi, S., & Nguyen, T. Q. (2015, November). Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 136-140). IEEE.

[7]. Kang, B., Tripathi, S., & Nguyen, T. Q. (2015, November). Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 136-140). IEEE.

[8]. Yasir, F., Prasad, P. W. C., Alsadoon, A., Elchouemi, A., & Sreedharan, S. (2017, July). Bangla Sign Language recognition using convolutional neural network. In *2017 international conference on intelligent computing, instrumentation and control technologies (ICICICT)* (pp. 49-53). IEEE.

[9]. Islalm, M. S., Rahman, M. M., Rahman, M. H., Arifuzzaman, M., Sassi, R., & Aktaruzzaman, M. (2019, September). Recognition bangla sign language using convolutional neural network. In *2019 international conference on innovation and intelligence for informatics, computing, and technologies (3ICT)* (pp. 1-6). IEEE.

[10]. Huang, J., Zhou, W., Li, H., & Li, W. (2015, June). Sign language recognition using 3d convolutional neural networks. In *2015 IEEE international conference on multimedia and expo (ICME)* (pp. 1-6). IEEE.

[11]. Hore, S., Chatterjee, S., Santhi, V., Dey, N., Ashour, A. S., Balas, V. E., & Shi, F. (2017). Indian sign language recognition using optimized neural networks. In *Information Technology and Intelligent Transportation Systems: Volume 2, Proceedings of the 2015 International Conference on Information Technology and Intelligent Transportation Systems ITITS 2015, held December 12-13, 2015, Xi'an China* (pp. 553-563). Springer International Publishing.

[12]. Xiao, Q., Qin, M., & Yin, Y. (2020). Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, *125*, 41-55.

[13]. Lee, C. C., & Gao, Z. (2020). Sign language recognition using two-stream convolutional neural networks with Wi-Fi signals. *Applied Sciences*, *10*(24), 9005.

[14]. Hasan, M. M., Srizon, A. Y., Sayeed, A., & Hasan, M. A. M. (2020, November). Classification of sign language characters by applying a deep convolutional neural network. In *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)* (pp. 434-438). IEEE.

[15]. Naglot, D., & Kulkarni, M. (2016, August). Real time sign language recognition using the leap motion controller. In *2016 international conference on inventive computation technologies (ICICT)* (Vol. 3, pp. 1-5). IEEE.

[16]. Latif, G., Mohammad, N., Alghazo, J., AlKhalaf, R., & AlKhalaf, R. (2019). ArASL: Arabic alphabets sign language dataset. *Data in brief*, *23*, 103777.