Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18681-3

Arabic sign language letters recognition using Vision Transformer

Aya F. Alnabih[1] · Ashraf Y. Maghari[1]

Received: 31 August 2023 / Revised: 25 January 2024 / Accepted: 19 February 2024

Abstract

Sign languages, as means of communication, enable individuals to convey messages through hand gestures, body movements, and facial expressions. It is primarily utilized by those with hearing impairments to communicate with others. Currently the recognition of static sign language predominantly relies on the Convolutional Neural Network (CNN) approaches and transfer learning for classifying hand sign images. While researchers have been actively working on this issue, only a few have explored the potential of Vision Transformers (ViT) in addressing the sign language recognition problem, particularly Arabic sings. Where, no record or documentation aimed at identifying Arabic sign language letters using ViT model. Vision Transformers is a new addition to the world of deep neural networks and have shown promising performance with less computational power required compared to existing methods. This paper aims to leverage the capabilities of Vision Transformers to improve the accuracy of Arabic sign language recognition. In that context, a ViT-based model is proposed in which a pre trained ViT model is fine tuned to be adapted for recognizing Arabic sign language letters. To finetune and evaluate our ViT-based model, we utilize the ArSL2018 dataset which consists of 54,049 images of Arabic sign language letters with 32 classes. Accuracy, F1 score, recall and precision are used for evaluation. The proposed model achieved an accuracy of 99.3% on the ArSL2018 dataset, which outperforms some recent CNN based models. Additionally, we conducted evaluations on real case images to assess the practicality and real-world applicability of our model.

Keywords Deep learning - Vision Transformer · Sign language recognition · Arabic sign language

✉ Ashraf Y. Maghari
amaghari@iugaza.edu.ps

Aya F. Alnabih
anabih1997@gmail.com

[1] Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine

Published online: 06 March 2024

1 Introduction

Individuals with hearing impairments use Sign Language (SL) as a nonverbal communication method to communicate through important body movements known as gestures or signs [1]. Sign language recognition (SLR) is the process of interpreting and translating sign languages, such as British Sign Language (BSL) or American Sign Language (ASL), into text or speech using technology [2]. Communication issues arise as a result of different sign languages or a lack of understanding of sign languages by people who use spoken languages. These issues drew the attention of researchers, who set about developing applications to help people communicate more effectively and eliminate the communication gap between communities by using a variety of techniques and methods for sign language recognition.

SLR research is particularly interested in how to clearly and unambiguously interpret the hand and body movements associated with sign language. Research efforts can be categorized into two primary groups based on the type of data utilized: sensor-based methods and vision-based methods [3]. SLR systems that utilize sensors rely on external devices, like data gloves worn by the signer, to collect data about their actions [4]. However, sensor-based approaches may come with drawbacks such as potential discomfort or limitations in movement due to sensor configuration. On the other hand, vision-based systems rely on images and videos to understand the meaning of hand signs.

Vision based research efforts can be categorized into two primary groups: traditional machine learning and deep learning [1, 3]. Figure 1 shows deep learning system architecture for static sign language gesture recognition. As mentioned in reference [5] the architecture of a Convolutional Neural Network (CNN) includes components that incorporate types of layers and activation functions. In general, the CNN architecture is organized, around four layers: the convolutional layer, the pooling layer, the ReLU layer and the fully connected or output layer.

The development made in deep learning has significant implications for sign language recognition. This is evident through the application of essential deep learning techniques, including convolutional neural networks (CNN) [6, 7], Recurrent neural networks (RNN) [8], and Recurrent convolutional neural networks (RCNN), which have been utilized in experiments focusing on this domain. In recent years, one particularly promising approach that has gained considerable traction is the utilization of vision transformers (ViT) [9]. Vision transformers are deep learning models that uses attention mechanism which is

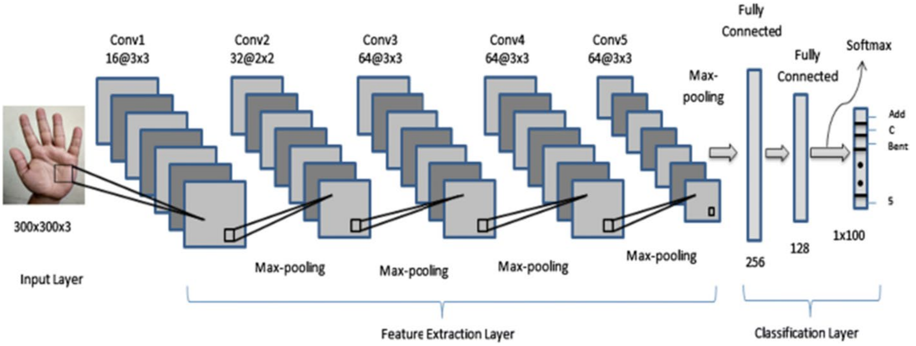


Fig. 1 Deep learning recognition system [5]

known for their effectiveness in tasks such as image classification and object detection. It has also been used in the field of American Sign Language classification and has proven its efficiency Tan, Lim [15].

In this paper, we consider the problem of recognition Arabic sign language letters using vision transformer (ViT). A ViT-based model is proposed in which a pre trained ViT model is finetuned with an Arabic sign dataset (ArSL2018) and tested for its recognition performance. Hence, a framework with many steps for Arabic sign recognition is proposed. In the framework, a hand detector module checks if there is a hand in the image. Then, the detected hand is isolated from the image and fed into the classifier which categorizes the class of hand sign. ArSL2018 dataset [16] which consists of 54,049 images of Arabic sign languages letters with 32 classes has been used for training and testing the ViT-based model. Our experimental results indicate that the proposed ViT-based model outperforms some recent CNN based models. Moreover, the experiments were carried out on real case images to evaluate and assess the practicality and real-world applicability of our model.

The rest of paper is structured as follows; Section 2 discusses related research and studies. Section 3 presents the proposed ViT-based model and the dataset. In Section 4 we report on the experiments conducted and discuss their corresponding results. Finally in Section 5 we conclude our work and present directions for further research.

2 Related works

Researchers have dedicated efforts, to the development of systems for identifying and comprehending sign language. This task necessitates algorithms that can recognize and classify signs. To address this challenge, researchers have explored two approaches to sign language recognition: machine learning and deep learning [1]. Both traditional machine learning and deep learning methods have been employed in sign language recognition. Traditional machine learning methods, such as k neighbors, support vector machines and random forests have been combined with feature extraction techniques like SURF, SIFT, PCA, LDA and HOG to achieve recognition tasks [1, 3]. On the hand learning techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models have also exhibited potential in recognizing sign language. CNNs excel at capturing relationships and hierarchies of image features while RNNs are effective in processing data with dynamics.

Regarding traditional machine learning, Aryanie and Heryadi [10] utilized a k-NN classifier for recognizing finger spelling in American Sign Language (ASL) and achieving acceptable accuracy. Similarly, Yasir, Prasad [11] applied the SIFT based approach to recognize Bangla sign language (BdSL) by training SVM classifiers, for each signed word using a Bag of Words model. However, this paper focuses on deep learning methods.

Many researchers have made progress in sign language recognition by deep learning methods Lee, Ng [8] successfully implemented a real time sign recognition system using LSTM RNN combined with k-NN classification. Their approach achieved an accuracy rate of 99.44% when applied to American Sign Language (ASL) alphabets. Furthermore, Mujahid, Awan [12] proposed a model based on YOLO v3 and DarkNet 53 for gesture recognition delivering excellent results across various performance metrics such, as accuracy, precision, recall and F1 scores. In this paper, we work on recognizing Arabic sign language. However, this paper focuses on deep learning methods in field of Arabic sign language recognition.

In the field of Arabic sign language recognition, Alnuaim, Zakariah [13] focused their research on training neural network (CNN) models to classify Arabic sign language. Their study specifically aimed to recognize 32 alphabet sign categories. Employed ResNet50 and MobileNetV2 models, which were fed preprocessed images. To boost performance, they applied techniques performed hyperparameter tuning and utilized data augmentation methods. As a result, their final model achieved an accuracy of 97% on the test set. Recently, Duwairi and Halloush [14] employed transfer learning on learning models such as AlexNet, VGGNet, GoogleNet/Inception. They also tested shallow learning approaches using SVM and K-nearest neighbors' algorithms as baselines. The proposed model achieved an accuracy of 97% in recognizing Arabic sign language alphabets. The evaluation was conducted on a dataset consisting of 54,049 labeled images, which is a large-scale dataset for Arabic sign language. However, this paper focuses Vision Transformer method in field of sign language recognition.

In the domain of applying Vision Transformer for the recognition of sign language, Tan, Lim [15] developed a system to identify hand gestures in American Sign Language. They trained the model using three datasets; ASL dataset, ASL, with Digits dataset and NUS hand gesture dataset. The outcomes were highly remarkable with accuracy rates of 99.98%, 99.36% and 99.85% achieved for each dataset. This showcases the systems effectiveness, in accomplishing hand gesture recognition tasks. However, in this work, we consider the problem of recognition Arabic sign language using vision transformer (ViT). Hence, the aim is to fine tune the ViT model with an Arabic sign dataset (ArSL2018) and test its recognition performance.

3 The proposed Arabic sign recognition framework

Figure 2 illustrates the main steps of the proposed framework for Arabic sign recognition. Initially an image is inputted into the system followed by the hand detector module which checks if there is a hand in the image. Once a hand is detected it is isolated from the image. Subsequently fed into the classifier. The classifier then categorizes the class of hand sign.

3.1 ViT-based model

In this paper, a ViT-based model is proposed in which a pre trained ViT model is fine tuned to be adapted for recognizing Arabic sign language letters. Figure 3 shows the steps of finetuning the pre-trained ViT model on ArSL dataset. Initially, we prepare a dataset

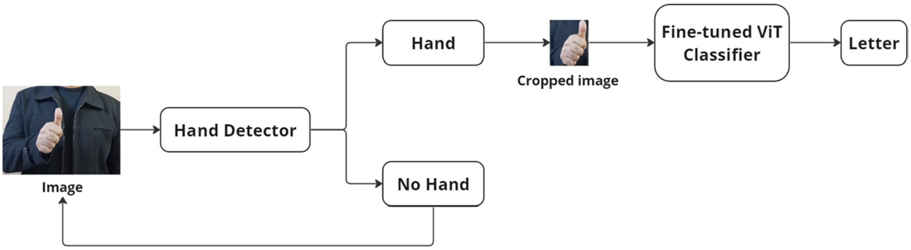


Fig. 2 The proposed Arabic sign recognition framework

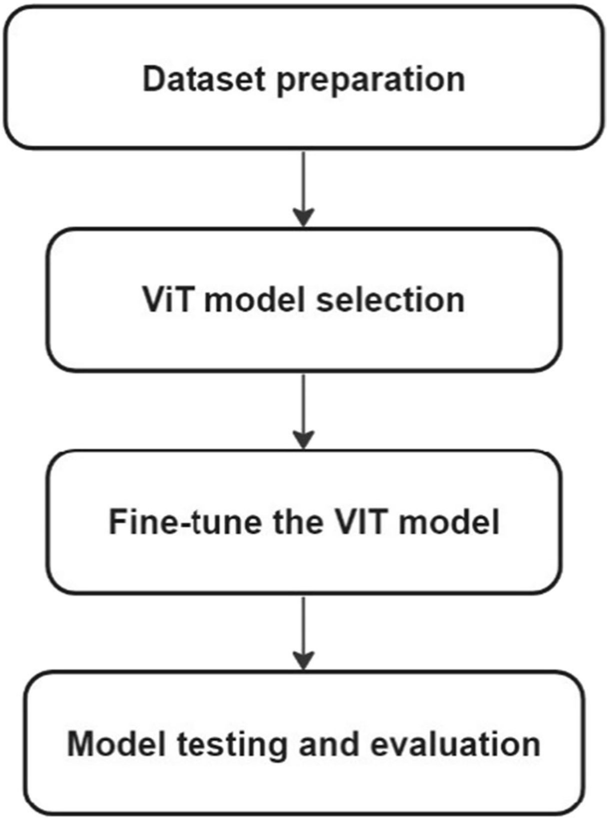


Fig. 3 Our ViT-based model steps

comprising numerous images showcasing signs from Arabic Sign Language letters. The next step involves selecting the most suitable pre-trained Vision Transformer (VIT) model for our specific task. In this step, we fine-tune the selected pretrained VIT model by retraining it with our collected dataset. Finally, we evaluate the performance of the model on a test suite using evaluation metrics to measure its effectiveness. Additionally, we assess how well the model performs on real world images to determine its practicality and reliability in real time sign language recognition, which's crucial, for real world applications.

3.2 Dataset preparation

In the field of Arabic sign language letter recognition, many researchers as Alnuaim et al. [13], Duwairi and Halloush [14] have commonly employed a well-known dataset called ArSL2018 [16] which consists of 54,049 images of Arabic sign languages letters of 32 classes and other standard signs. Each class contains a number of images of the sign in different positions. The images were performed by more than 40 different people, providing a diverse and representative sample of the signs used in Arabic sign languages. Figure 4 shows a sample image of all Arabic language signs. We used this dataset for training and testing the ViT model. This allows us to make comparisons between our work and similar studies in the field.

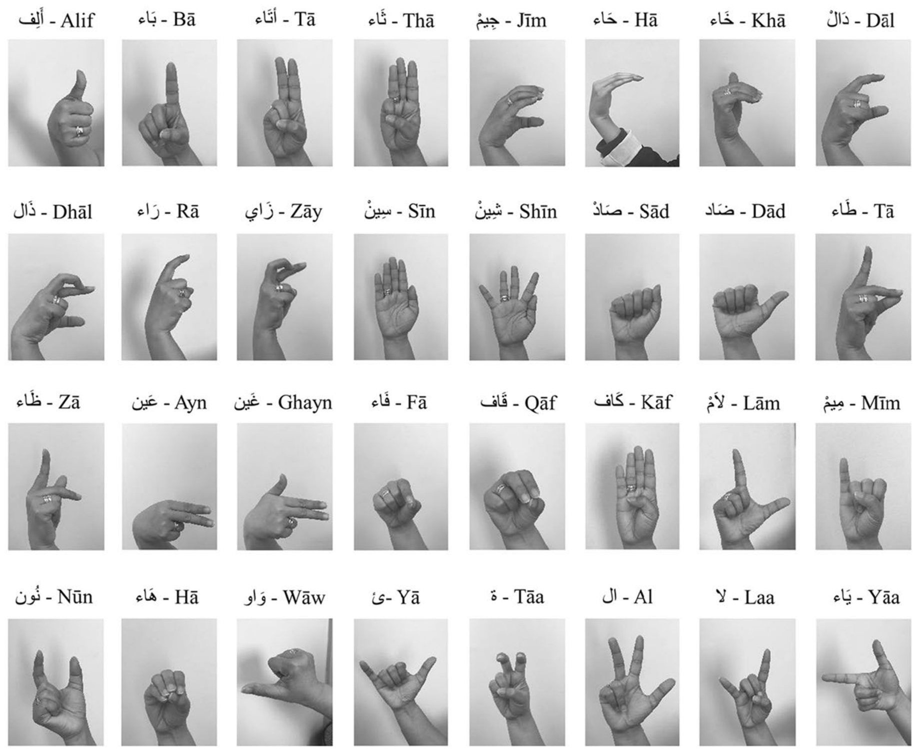


Fig. 4 Sample of 32 classes from ArSL2018 dataset [16]

3.3 ViT model selection

For our experiment, a pre-trained ViT model is selected. We discovered the most used models for image classification tasks, which have shown effectiveness in the domains of deep learning. Two of the pre trained Vision Transformer (ViT) models created by Google are "google/vit large patch16 224 in21k" and "google/vit base patch16 224 in21k". Additionally, the Facebook team has previously trained a Vision Transformer (ViT) model called "facebook/dino vitb16".

All of the models have a similar architecture, as they are transformer-encoding models that are pre-trained on a large set of images in a supervised manner, specifically the ImageNet-21k dataset. This dataset consists of huge images with 21,000 different categories, and the models are trained at a resolution of 224 × 224 pixels. ViT models follow a BERT-like structure, treat images as sequences of fixed-size spots, apply linear modulation. Table 1 shows the difference between the hyperparameters of the three models.

3.4 Fine-tune the ViT model

The model training and fine-tuning phase involve selecting a pre-trained Vision Transformer (ViT) model and fine-tuning it using the ArSL2018 dataset.

Table 1 Different hyperparameters of the selected three pretrained ViT models

Parameter	‘google/vit-base-patch16-224-in21k’ model	‘google/vit-large-patch16-224-in21k’ model	facebook/dino-vitb16 model
Trainable parameters _number	85.8 million	303 million	85.8 million
patch_size	16	16	16
image_size	224	224	224
hidden_size	768	1024	768
initializer_range	0.02	0.02	0.02
intermediate_size	3072	4096	3072
num_attention_heads	12	16	12
num_channels	3	3	3
num_hidden_layers	12	24	12
hidden_act	gelu	gelu	gelu

3.4.1 Vision Transformer architecture

ViT is a deep learning system created specifically for computer vision applications such as object detection, and image classification. It is tailored to analyse and interpret visual data, making it highly effective in tasks related to computer vision [17]. ViT is developed by researchers at Google and was originally introduced for natural language processing tasks [18]. It based on the Transformer architecture as shown in Fig. 5. Unlike traditional CNNs, ViT processes images as arrays of pixels through a series of convolutional and pooling layers, it processes images by dividing them into a grid of patches and feeding these patches directly into the transformer architecture [17]. The transformer architecture is composed of two main components: the encoder and the decoder.

The encoder component receives an image as input and transforms it into a collection of feature vectors. These feature vectors contain essential information about the image.

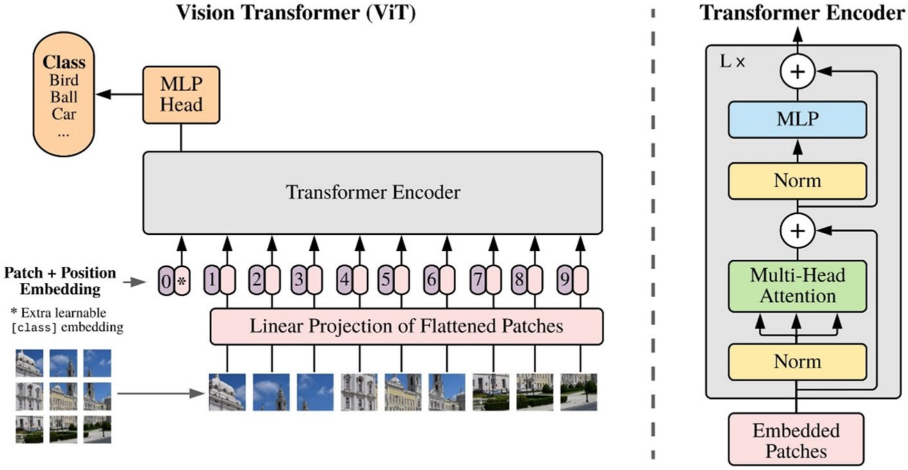


Fig. 5 ViT architecture [17]

Subsequently, the decoder component takes these feature vectors as input and processes them to generate a prediction or output for the image. The decoder leverages the information encoded in the feature vectors to make its prediction.

Here's a comprehensive breakdown of the elements and stages in the Vision Transformer architecture:

2.1.1 Embedding Patches:

- The input image is divided into fixed size patches, where each patch represents a region of the image.
- These patches are then transformed into sequences of vectors through a linear projection. This transformation helps convert image details into data sequences.

2.1.2 Positional Encoding:

- Similar to language-based transformers the Vision Transformer lacks information found in Convolutional Neural Networks (CNNs) due to the absence of convolutional operations.
- To address this positional encoding are added to the patch embeddings to provide spatial arrangement information about patches within the image.

2.1.3 Transformer Encoder:

- The patch embeddings, along with the positional encodings, are fed into a standard transformer encoder.
- Self-attention mechanisms within the transformer capture relationships between different patches, allowing the model to understand global context and dependencies in the image.
- Multi-layer perceptrons (MLPs) are applied after the self-attention layers to process and refine the contextual information.

2.1.4 Classification Head:

- The final output, from the encoder is directed towards a classification head.
- Based on the nature of the task, the classification head can comprise connected layers that generate predictions, for classes within an image classification problem.

This approach fundamentally alters our perception of comprehending images. The ViT is capable to handle both global context in images without depending on convolutional operations highlights the effectiveness of self-attention mechanisms. This change has resulted in the development of models that modify transformer-based architectures, for computer vision tasks thereby pushing the limits of what can be achieved in this field.

3.4.2 Fine-tuning steps

In our experiments, we used one of the pre-trained ViT models to recognize the Arabic sign language letters by fine-tune the model on the ArSL dataset datasets. This process aims to adapt the model specifically for the task of recognizing Arabic sign language letters. Fine-tuning a pre-trained ViT model typically involves the following steps:

1. **Initialization**: Load the selected pre-trained ViT model with its learned parameters. These parameters capture general features from a large dataset, such as ImageNet, and

provide a strong starting point for the fine-tuning process. Initialize the model weights, biases, and other necessary components.

2. **Freeze Layers**: it's common practice to freeze the last layer, preserving the learned representations while tailoring the model for a task-specific dataset.
3. **Hyperparameter tuning**: Adjust various hyperparameters to optimize the model's performance. Important hyperparameters include the learning rate, batch size, weight decay, and dropout rate. Conduct experiments by trying different configurations and evaluate their impact on the model's accuracy and convergence.
4. **Fine-tuning process**: Train the pre-trained ViT model on the dataset of Arabic sign language letters using backpropagation. Feed the images as inputs to the model and compute the loss using a suitable loss function. Use an optimizer to update the model's parameters iteratively and minimize the loss.
5. **Training iterations**: Perform multiple training iterations or epochs. In each iteration, the model processes the entire dataset, adjusting its parameters to improve performance. Monitor the training progress by evaluating metrics such as training loss and accuracy.

By following these steps, a pre-trained ViT model can be finetuned effectively using a dataset of Arabic sign language letters, allowing the model to adapt to the task of recognizing Arabic sign language letters accurately.

3.5 Model evaluation

Evaluating the effectiveness of a machine learning model is crucial for its development and practical implementation. After the training phase, the model's performance can be assessed using various evaluation methods, including: performance metrics, comparison with related models, and real case testing on webcam. Accuracy, precision, recall, and F1-score have been used to evaluate the proposed ViT-based model. Accuracy is a metric that assesses the overall correctness of the classification model. Precision is a metric that measures the accuracy of positive predictions. Recall measures the ability of a model to correctly identify positive instances. Where, F1 score combines precision and recall into a single value, providing a balanced assessment of the model's performance.

4 Experimental results and discussion

4.1 Dataset

The ArSL dataset [16] was created in 2018 and consists of 54,049 images of signs from Arabic sign languages, including individual letters of the Arabic alphabet and other standard signs. These images were performed by more than 40 different people, providing a diverse and representative sample of the signs used in Arabic sign languages. The dataset includes 32 different classes of signs, and the number of images per class varies. Each class contains a number of images of the sign in different positions. Figure 6 shows sample classes from Arabic sign language dataset.

We divided the dataset into three distinct subsets: training, validation, and test sets. As shown in Table 2, the dataset was split using a ratio of 70% for training, 15% for validation, and 15% for testing. This partitioning allowed us to utilize a majority of the

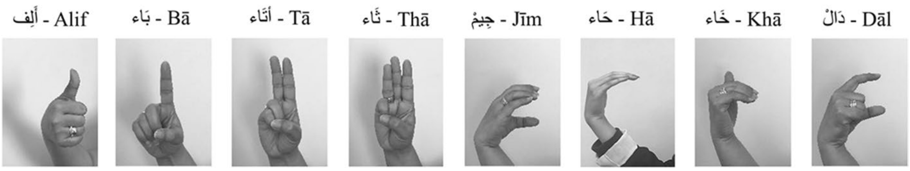


Fig. 6 Sample classes from ArSL dataset [16]

Table 2 ArSL dataset splitting for training, validation, and testing the proposed model

Training images	Validation Images	Test Images	Total
37853	8111	8111	54075

Table 3 Experimental result performances of the 3 ViT models

ViT Models	Precision	Recall	F1	Accuracy
google /vit- base-patch16-224-in21k	99.0	99.1	99.0	99.0
google /vit- large-patch16-224-in21k	**99.3**	**99.3**	**99.4**	99.3
facebook/dino-vitb16	98.9	98.9	99.0	98.9

data for training the model while reserving separate subsets for validating its performance during development and evaluating its generalization abilities on unseen data.

4.2 Results

In the first experiment, the goal is to identify the most suitable pre-trained Vision Transformer (ViT) model for the Arabic sign language letters recognition task. This involves evaluating and comparing the performance of different ViT models on the dataset. These models have been trained on large-scale datasets and have learned to extract meaningful features from images.

The experiment involves fine-tuning various pre-trained Vision Transformer (ViT) models on the ArSL2018 dataset. The first is the ‘google/vit-large-patch16-224-in21k’ model, which has a larger architecture and more trainable parameters. The ‘google/vit-base-patch16-224-in21k’ and ‘facebook/dino-vitb16’ models share the same model architecture and number of trainable parameters, but have a smaller architecture and fewer trainable parameters than the ‘google/vit-large-patch16-224-in21k’ model. All models have been pre-trained on a large collection of images and can capture meaningful features for image recognition tasks. The performance of these models on the ArSL2018 dataset has been evaluated using metrics such as recall, precision, accuracy, and F1 score, as presented in Table 3. These metrics offer valuable insights into the models’ ability to classify Arabic sign language letters accurately. Furthermore, the computational requirements and training convergence of each model will be taken into account to assess their practical applicability.

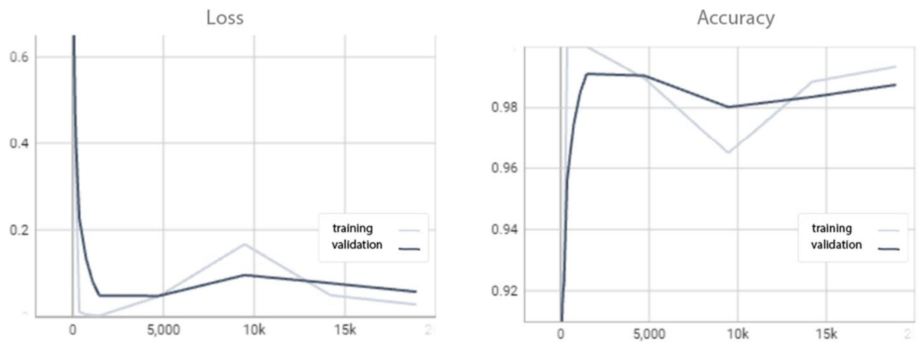


Fig. 7 The accuracy and loss of ‘google/vit-large-patch16-224-in21k’ model

Table 4 Accuracy comparison of our Vit-based model with related work

Papers	Model	Accuracy
Our work	Vision Transformer (ViT)	**99.3%**
Latif, Mohammad [19]	Convolutional Neural Networks (CNN)	97.6%
Alsaadi, Alshamani [20]	AlexNet	94.81%
Zakariah, Alotaibi [21]	EfficientNetB4	98%
Alnuaim, Zakariah [13]	ResNet50 and MobileNetV2	95%
Duwairi and Halloush [14]	VGGNet	97%

4.2.1 Model evaluation metrics

As shown in Table 3, the results of the experiments indicate that the 3 models have very similar accuracies where ‘google/vit-large-patch16-224-in21k’ model achieved slightly higher accuracy compared to other models on the ArSL2018 dataset for Arabic sign language letters recognition.

Figure 7 shows the ‘google/vit-large-patch16-224-in21k’ model accuracy and loss for training and validation. The ‘google/vit-large-patch16-224-in21k’, is selected for the ViT-based sign recognition model.

4.2.2 Comparison with some related deep learning models

To accurately assess the performance of our Vit-based model, it is essential to compare its accuracy with other similar related works that have utilized the same dataset. This comparison provides valuable insights into the relative effectiveness and performance of our Vit-based model within the context of the given dataset. Table 4 shows the comparison with other related deep learning models.

Our Vit-based model achieved a remarkable accuracy of 99.3% on the task of recognizing Arabic sign language letters. This performance surpasses the results reported by other studies. These comparisons demonstrate that our Vit-based model outperformed these approaches in terms of accuracy, highlighting its effectiveness and superiority in recognizing Arabic sign language letters.

4.2.3 Testing on images from web camera

Testing the model on real life images confirms the importance of the model and its ability to provide results thereby strengthening its significance in our research and classification efforts. We integrated a hand detector pretrained model "HandDetection" [22] into our existing framework after making adjustments. The purpose of this step was to locate and identify hands in the images and videos of our dataset. By using the pretrained model we ensured effective hand detection, which was essential, for the classification tasks. To test the proposed model on real world effectiveness, we followed a series of steps outlined in Fig. 8. These steps involved image capturing with web camera, hand frame detection, and sign recognition.

1. **Capture an image from a webcam**: Obtain an image containing sign language gestures using a webcam.
2. **Frame detection**: Apply a detector to identify frames within the captured image that contain sign language gestures. This step helps isolate relevant regions of interest for further analysis.
3. **ViT classifier application**: Utilize the fine-tuned ViT classifier on the frames identified in the previous step. The classifier processes these frames and make predictions regarding the recognized sign language letters.

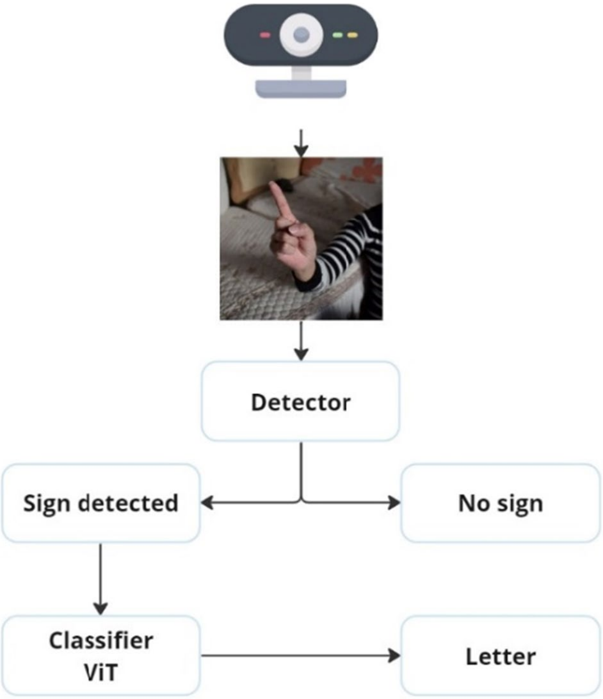
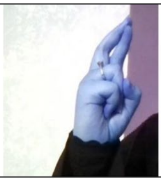
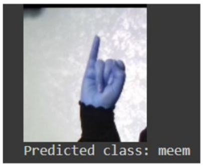
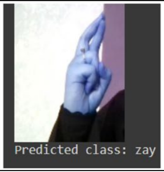
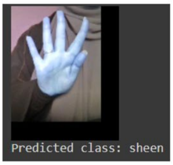
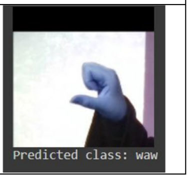


Fig. 8 The model test methodology on real case image from webcam

Table 5 Test our model on images captured from webcam

Image				
Class	Meem / ميم	Zay/زاي	sheen / شين	Waw / واو
Prediction class	 Predicted class: meem	 Predicted class: zay	 Predicted class: sheen	 Predicted class: waw

We tested our model on 20 samples from different classes, and the result is that all samples are 100% correctly detected. Table 5 presents some sign images of these experiments.

4.3 Discussion

Our experimental results clearly show that the Vision Transformer (ViT) model is the most accurate model compared to other models at recognizing hand gestures. This aligns with a study conducted Karna, Kode [23], where they developed a system to interpret American Sign Language letters. Their system was extensively tested and achieved an accuracy rate of 99.99% for nearly all the letters. Furthermore, the work of Tan, Lim [15], introduced another noteworthy system for American Sign Language hand gesture recognition. They conducted their training using three distinct datasets: the ASL dataset, the ASL with Digits dataset, and the NUS hand gesture dataset. The results were quite impressive, achieving accuracy rates of 99.98%, 99.36%, and 99.85% for each dataset. The remarkable accuracy of the ViT model in hand gesture recognition tasks can be attributed to the advantages it possesses. Firstly, ViT excels at capturing context and effectively handling images while maintaining a consistent architecture. Secondly, its self-attention mechanism enables the model to focus on aspects within an image, enabling it to comprehend relationships within input data without relying heavily on extensive convolutions. As a result, ViT demonstrates improved understanding of context, which proves valuable for tasks such as object recognition. Additionally, this self-attention mechanism enables the ViT model to perform well when working with image datasets. On the other hand, the proposed ViT-based mode has been tested for real life images from web camera. The model confirmed its ability to provide accurate results on tested web camera images, which thereby showed its significance in our research and classification efforts.

5 Conclusion

In this paper, we show how to apply the Vision Transformer (ViT) model in recognizing Arabic Sign Language characters from images and convert them effectively into written forms. By fine-tuning an existing pre-trained ViT model through train it on large ArSL2018 dataset composed entirely of 54,049 Arabic Sign Language character images, we achieved noteworthy results. The proposed ViT-based model achieved an impressive accuracy rate of 99.3%. At the same time, the precision, recall and F1 score results also showed a distinction of 99.4%, 99.3% and 99.3%, respectively.

These results highlight the effectiveness of the ViT model in terms of accurately recognizing and classifying Arabic sign language letters. Furthermore, we made comprehensive performance comparisons between the developed ViT model and other related deep learning models that used the same dataset. For Arabic Sign Language recognition, the comparison showed that our model performed better than the state-of-the-art models in terms of recognition accuracy. Moreover, the experiments were carried out on real case images, captured from web camera, to evaluate the applicability of ViT model on real-world images.

In the future, for enhancing the sign language recognition system, we can include integrating facial expression analysis and expanding the dataset to include numbers and words. Further, the model can contribute in creating a real time recognition system specifically designed for sign language gestures. These advancements hold promise for improving our Arabic sign language recognition system.

In this context, we can conclude our future directions in the following 5 points:

1. Facial expressions: Incorporate facial expression analysis into the sign language recognition system.
2. Expand the data set: Collect a larger data set consisting of new categories such as numbers and words to improve model performance and generalization capabilities.
3. Real-time Recognition: Develop a system that can process and classify Arabic sign language gestures in real time.
4. User interface and accessibility: Integrate the ViT-based Arabic sign language recognition system into user-friendly applications or devices.
5. Signs generation: Explore the domain of sign generation as part of our research focus.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Madhiarasan DM, Roy P, Pratim P (2022) A comprehensive review of sign language recognition: different types, modalities, and datasets arXiv preprint arXiv:2204.03328. https://doi.org/10.48550/arXiv.2204.03328
2. Sharma S, Singh S (2020) Vision-based sign language recognition system: a comprehensive review. In: 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE
3. Al-Qurishi M, Khalid T, Souissi R (2021) Deep learning for sign language recognition: current techniques, benchmarks, and open issues. IEEE Access 9:126917–126951
4. Kudrinko K et al (2020) Wearable sensor-based sign language recognition: a comprehensive review. IEEE Rev Biomed Eng 14:82–97
5. Wadhawan A, Kumar P (2020) Deep learning-based sign language recognition system for static signs. Neural Comput Appl 32:7957–7968

6. AlKhuraym BY, Ismail MMB, Bchir O (2022) Arabic sign language recognition using lightweight CNN-based Architecture. Int J Adv Comput Sci Appl 13(4):319–328
7. Liu Y et al (2021) Dynamic gesture recognition algorithm based on 3D convolutional neural network. Comput Intell Neurosci 2021
8. Lee CK et al (2021) American sign language recognition and training method with recurrent neural network. Expert Syst Appl 167:114403
9. Dosovitskiy A et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
10. Aryanie D, Heryadi Y (2015) American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier. In: 3rd International Conference on Information and Communication Technology (ICoICT), IEEE
11. Yasir F et al (2015) Sift based approach on bangla sign language recognition. In: 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA), IEEE
12. Mujahid A et al (2021) Real-time hand gesture recognition based on deep learning YOLOv3 model. Appl Sci 11(9):4164
13. Alnuaim A et al (2022) Human-computer interaction with Hand gesture recognition using ResNet and MobileNet. Comput Intell Neurosci 2022. https://doi.org/10.1155/2022/8777355
14. Duwairi R, Halloush ZA (2022) Automatic recognition of Arabic alphabets sign language using deep learning. Int J Electric Comput Eng (2088-8708) 12(3):2996–3004
15. Tan CK et al (2023) HGR-ViT: hand gesture recognition with vision transformer. Sensors 23(12):5555
16. Latif G et al (2019) "ArASL: Arabic alphabets sign language dataset." Data in brief 23:103777
17. Deng J et al (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference On Computer Vision and Pattern Recognition, IEEE
18. Vaswani A et al (2017) Attention is all you need. Advances in neural information processing systems (NIPS 2017), 30
19. Latif G et al (2020) An automatic arabic sign language recognition system based on deep CNN: an assistive system for the deaf and hard of hearing. Int J Comput Digit Syst 9(4):715–724
20. Alsaadi Z et al (2022) A real time arabic sign language alphabets (ArSLA) recognition model using deep learning architecture. Computers 11(5):78
21. Zakariah M et al (2022) Sign language recognition for Arabic alphabets using transfer learning technique. Comput Intell Neurosci 2022. https://doi.org/10.1155/2022/4567989
22. Work Hand Detection (2023) Available from: https://universe.roboflow.com/work-tbypc/handdetection-qycc7/model/1. Accessed 20 May 2023
23. Karna SNR et al (2021) American sign language static gesture recognition using deep learning and computer vision. In: 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC). IEEE

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.