Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

ConvNeXt in Isolated Sign Language Recognition

Aryaman Darda[1]*; Mona Sheikh Zeinoddin[2]

[1]The University of California, Berkeley, California, U. S

[2]Centre of Doctoral Training, University College London, U.K

E-mail: aryaman_darda@berkeley.edu

Received Date: 20 Oct 2024, Manuscript No. IJCSMA-24-150518; **Editor assigned:** 23 Oct 2024, Pre QC No. IJCSMA-24-150518(PQ); **Reviewed:** 28 Oct 2024, QC No. IJCSMA-24-150518(Q); **Revised:** 31 Oct 2024, Manuscript No. IJCSMA-150518(R); **Published date:** 05 Nov 2024; DOI.

Abstract

This paper explores sign language, a natural mode of communication for the deaf community. However, sign language often remains challenging to learn and creates communication barriers between the deaf and the hearing. This work addresses this issue by assessing the performance of the state-of-the-art convolutional model, ConvNeXt, on the novel task of sign language recognition. The research yields compelling results, with accuracies surpassing 99% and fast training times that rival advanced Vision Transformers (ViTs). The experiments are rigorously evaluated using the publicly available Sign-Language-MNIST Dataset, an established benchmark for sign language research. A comparison of the generalizability of ConvNeXt and ViT is further undertaken using the publicly available Indian Sign Language Dataset which shows ViTs generalize better by ~3% in sign language recognition tasks. The findings of this study contribute to the broader goal of improving communication for the deaf community while also highlighting the capability of carefully constructed lightweight convolutional models that have recently fallen out of favour.

Keywords: ConvNext; Generalizability; Isolated sign language recognition; Vision transformers

1. Introduction

Hearing loss and deafness are experienced by 230 million people worldwide, two-thirds of who live in developing countries [1]. The majority of people in the deaf community, consequently, rely on some form of sign language as

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

their primary method of communication with the world. However, as effective as a particular sign language is, it can at times be difficult and time-consuming to master for the hearing community. In recent years, computer vision, which involves interpreting visual data using computers, has witnessed a remarkable transformation, owing to advancements in deep learning techniques such as convolutional and transformer based models. Convolutional Neural Networks (CNNs), in particular, pioneered breakthroughs in tasks like image classification, object detection, and semantic segmentation [2-4].

In the context of sign language recognition, these computer vision models have been instrumental in both continuous and isolated sign language recognition. This is due to the necessity of capturing fine-grained spatial information for the accurate classification of sign data. Traditional CNNs have been the go-to models for such tasks due to their efficiency in extracting local features and patterns [5]. However, the emergence of ViTs and SWIN Transformers also opened new possibilities. Their capacity to capture long-range dependencies and hierarchical features is very useful in recognizing the nuanced gestures of sign languages [6, 7].

Isolated Sign Language Recognition (ISLR) serves a critical role in the realm of computer vision's application to sign language, providing a crucial stepping stone toward advanced interpretation systems. By accurately classifying distinct signs, ISLR facilitates a foundational understanding of sign language, much like individual words in a sentence. This minute focus is essential, as each sign is a discrete gesture that conveys a specific meaning and requires precise recognition by computational models.

While ISLR focuses on these singular sign elements, Continuous Sign Language Recognition (CSLR) ventures into the more complex domain of interpreting sign language flows, where gestures evolve in a sequence akin to sentences in verbal communication. CSLR has recently been advancing rapidly, propelled by innovations such as Correlation Networks (CorrNets), which explicitly capture body trajectories across frames in order to effectively identify a sign in context [8].

Despite the immense progress in CSLR, this research zeroes in on ISLR due to its foundational importance. Mastery of ISLR is imperative in creating robust systems capable of tackling the intricacies of sign language. By dissecting and thoroughly understanding each sign's individual components, the groundwork for more complex tasks with more successful outcomes can be laid. This ensures that subsequent interpretations of continuous signing are built upon a nuanced understanding of the basic sign lexicon.

Before the advent of deep learning, which involves learning representations through networks with many layers, most methods for ISLR relied on classic machine learning models with handcrafted features. Examples involve using Sub-Units, Principal Component Analysis (PCA), and template matching for sign language recognition. However, despite achieving decent accuracy on the dataset they were constructed for, the overarching limitation of classical machine learning algorithms is that they require careful handcrafting of features and have poor scalability and generalizability. Current implementations are varied and highly sophisticated and models like MSG-3D allow for flexible spatial-temporal feature extraction and improved classification. Although these methods boast great

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

performance and effective feature extraction, they can be quite data-hungry and slow to train. This presents yet another alternate approach to the exhaustively researched task of ISLR fueled by groundbreaking research with respect to CNNs.

In 2022, a new type of convolution-based model architecture was proposed, known as "ConvNeXt" [9]. ConvNeXt is characterized by its extremely lightweight design while providing classification accuracies and scalability that can, under certain conditions, surpass those of popular transformer based architectures. While direct comparisons can be contingent on factors such as training conditions, model size, and parameter count, ConvNeXt has demonstrated impressive performance, even exceeding that of SWIN Transformers in image classification on the ImageNet Dataset when optimized training and architectural conditions are met. This suggests that the high performance typically associated with data-intensive, low-diverging Transformer-based models, which excel in many scenarios, can be achieved, or even surpassed by convolutional models like ConvNeXt. Despite this, ConvNeXt manages to retain the simplicity and efficiency inherent to standard CNNs. These pre-trained models are adept at capturing spatial information from images and transfer their learning effectively, facilitating ease in training and fine-tuning.

The work proposed in this paper aims to provide a holistic view of the ConvNeXt model on a novel task of ISLR and investigate how it transfers its learning to this task compared to in favor vision transformers. To further the exploration, we will investigate how misclassifications occur via the use of class activation maps that draw from weights in the latter stages of the classification process. We will then compare our results to those obtained using transformers as the classification model on the benchmark Sign-Language-MNIST dataset. Lastly, the generalizability of both models will be tested using the Indian Sign Language dataset which is a different sign language to the one on which the models will be fine-tuned. We hypothesize that the performance and generalizability of both architectures will be comparable in the task of ISLR, but ConvNeXt could potentially boast superior convergence times and ease of training.

2. Literature Review

2.1. Convolution-Based Models

The birth of CNNs can be traced back to the work of Yann LeCun *et al*. in 1989. Their groundbreaking paper introduced novel neural network architecture for the recognition of handwritten zip codes. The model, dubbed LeNet-5, was pioneering because it demonstrated the effectiveness of convolutional layers for spatial data processing and laid the foundation for future advancements in CNNs [10].

A significant leap in the field was marked by the introduction of AlexNet in 2012. Krizhevsky *et al*.'s won the ImageNet Recognition Challenge by some margin. This was because AlexNet was a deeper model than previous

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

CNNs and introduced key innovations such as the Rectified Linear Unit (ReLU) as the activation function and dropout for regularization [11].

The introduction of ResNet in 2015 by He *et al*. marked yet another milestone for CNNs. The architecture was able to address the "vanishing gradients" problem in very deep networks via the introduction of residual connections. The model performed impressively on the ImageNet challenge and ushered a shift towards deeper, more complex network architectures [12].

2022 marked yet another milestone for convolution-based neural networks with the introduction of "Con-vNeXt". This model surpassed all previous CNNs as it aimed to "modernize" traditional ResNets using principals that defined transformer architecture. From the use of non-local self-attention via larger kernels in the macro design to the use of non-linearity like Gaussian Error Linear Units (GeLUs) as part of the micro design, a revolutionary convolution-based architecture was created. It boasted accuracies of 88% on the ImageNet-1k dataset, surpassing state-of-the-art SWIN Transformers at the time [13].

2.2. Transformer Models

The inception of Transformers, which is a specific type of model architecture, revolutionized the field of Natural Language Processing (NLP). First introduced in 2017, the Transformer architecture replaced recurrent and convolutional layers in favor of the self-attention mechanism [14]. This mechanism allowed the model to weigh the significance of different parts of the input data relative to each other, which allowed for long-term dependencies in the data to be captured much more easily and efficiently. The Transformer's ability to process data in parallel significantly reduced training times and improved performance on a variety of NLP tasks such as machine translation, language modelling, and text summarization. This set a new standard for subsequent models.

Inspired by the success of Transformers in the NLP domain, researchers sought to extend their application to computer vision. The Vision Transformer (ViT) marked a pivotal transition. They approached image classification by partitioning image data into fixed-sized patches, linearly embedding the patches, and processing the patches through the Transformer mechanism. ViTs demonstrated that Transformers could achieve state-of-the-art results in image classification tasks, challenging the dominance of CNNs. The reason behind these successful results was the model's ability to capture global dependencies between image patches and this proved beneficial for learning complex visual data.

SWIN Transformers were introduced to deal with the quadratic time complexity of the self-attention mechanism with respect to the number of patches. They utilize a hierarchical structure that processes images in stages, reducing resolution while increasing the embedding dimension. Furthermore, they apply self-attention within local windows

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

that shift across layers, enabling efficient modelling of local and global contexts [15]. This design facilitated the application of Transformers to a broader range of vision tasks, including those with high-resolution input.

2.3. Isolated Sign Language Recognition

To kick start the exploration of isolated sign language detection, before the advent of deep learning frameworks, Kadous's utilization of Power Gloves, which, by applying PCA for dimensionality reduction of data, transformed complex hand gesture data into a manageable form of Auslan (Australian) sign recognition [16]. While the approach was a step forward, it was limited by the gloves' rudimentary sensory capabilities and the cumbersome nature of the hardware. Furthermore, Badhe *et al*. proposed a gesture learning algorithm for the translation of Indian Sign Language. A combinatorial algorithm is used to track hand movements as part of data pre-processing and recognition is done using template matching. They created templates from vast amounts of data and achieved accuracies of up to 97.5%. However, the template matching algorithm had limited generalization capability due to the handcrafting of features [17]. Another approach was the use of sub-units to disassemble signs into smaller, distinctive components to allow for more granular recognition, thereby increasing the classifier's generalization capability. Despite achieving sufficient accuracy on the classification task, the model required more data to be able to accurately learn features that were signer specific and those that were independent [18].

Thus arise the need for deep learning frameworks for sign language recognition. The research proposed by O. Koller *et al*. leverages the flexibility of standard CNNs for learning spatial filters and uses Long Short-Term Memory (LSTM) networks to capture temporal information. They use a multi-stream CNN-LSTM-HMM framework, via weak supervised learning, to discover sequential parallelism in sign language videos. Although their work revolves around CSLR and was able to achieve error rates of 5% on the Phoenix-14 dataset, it is still crucial to understand their framework for isolated recognition of signs on a per frame sampling basis. Their proposed model is computationally heavy and is more effective at capturing mid to long-range dependencies in the data [19].

Vazquez *et al*. built on these models by introducing RGB data as input into a 3D-CNN model. Besides capturing mid to long-term dependencies in the data, the model was now effective in extracting short-term dependencies which is crucial in sign language recognition. The model used was S3D and it displayed top-1% accuracies of up to 90% when used in isolation for detection. The model re-quires fewer parameters to learn and produces better accuracy results than most other 3D-CNN architectures. The work further used a skeleton-based graph approach that incorporated a flexible mechanism to understand the connected variations between nodes of any part of the graph on a predefined spatial and temporal scale by learning different levels of semantic information of the graph. Their MSG-3D approach resulted in 95.51% top-1% accuracy [20].

More recently, Pathan *et al*. introduced a multi-headed CNN which would take in a fusion of image and hand

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

landmark data as features. This custom CNN is lightweight and has architecture not unlike those found in standard CNNs. The work was able to demonstrate the effectiveness of adding hand landmark data to regular RGB data as it improved the model's overall classification accuracy from ~96% to ~99%.

3. Methods

3.1. The ConvNeXt Model

The ConvNeXt model is an extension of the standard ResNet-50 model using principles from hierarchical vision transformers and the model construction process is represented in **figure 1**.

The training procedure of the model entails using the AdamW optimizer and performing random cropping and resizing as part of data augmentations. A number of 300 epochs was suggested for training in the paper but our task involved simply fine-tuning the model pre-trained on the ImageNet-21k dataset. For this reason, and further computational constraints, fine-tuning was done for 5 epochs [21].

The macro design of the model involves changing the stage compute ratio. CNNs, by design, are multi-stage models where each stage involves convolution, pooling, and normalization operations to name a few. Traditional ResNets have (3, 4, 6, and 3) in each stage but ConvNeXt has (3, 3, 9, and 3). Furthermore, the "stem" cell dictates input image processing and comprises a 7×7 convolutional layer with stride 2 followed by pooling in standard ResNets. ConvNeXt, instead, implements a "pacify stem" cell which consists of a 4 × 4 convolution layer with stride 4 (non-overlapping).

The micro design of the model involves the use of Gaussian Error Linear Unit (GELU) which is a smoother version of ReLU and performs layer normalization instead of batch normalization. Finally, it employs fewer normalization layers and fewer activation functions than traditional ResNets while also introducing separate down sampling layers.

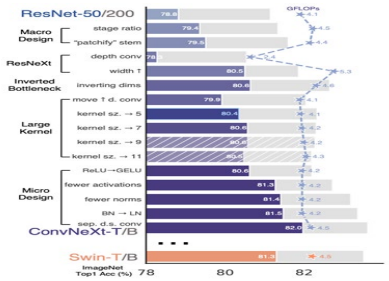

Figure 1. Showcases the design journey of ConvNeXt from a standard ResNet-50. Most of the macro and micro design changes are highlighted in this figure along with how each change affected classification accuracy.

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

For the purposes of our study and compute limitations, we found the ConvNeXt-T (iny) model to be sufficient and its architecture can be seen in **figure 2**.

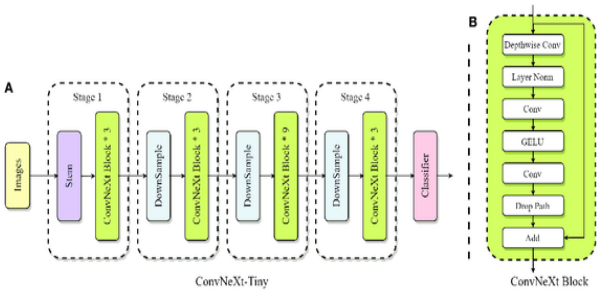

Figure 2. CovNeXt-T architecture.

3.2. The ViT Model

For the purposes of this exploration, a ViT for Image Classification model is used from Hugging Face. The model is pr-etrained using ImageNet-21k. At the core of this model lies the ViT. The architecture begins by transforming an input image into a sequence of flattened patches using convolution operation with a kernel size of 16 and stride 16. The data is then input into the positional encoder which comprises of 12 ViTLayers for the application of self-attention to extract and integrate relevant information from various parts of the image.

Following the attention mechanism is a 2-step feed-forward network which first expands the encodings to a higher dimensional space of 3072, then to the original dimension of 768. Layer normalization is applied before each attention and feed-forward operation for increased stability [22].

The output of the feed-forward layers is then passed through a final layer normalization layer and then to the classifier head. A detailed architecture of the ViT model can be seen in **figure 3**.

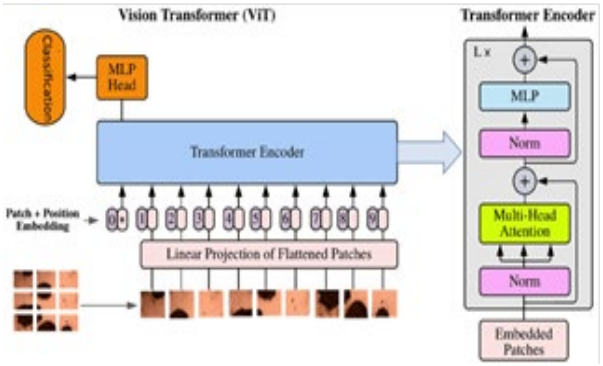

Figure 3. ViT architecture diagram.

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

3.3. Class Activation Maps

Class Activation Maps (CAMs) are a visualization technique that presents the influential regions within an input image for a model's classification decision. They generate heat maps by projecting back the weights of the output layer onto the final convolutional layer, thereby revealing areas of importance for a particular class prediction [23].

In the context of ConvNeXt in ISLR, CAMs can be particularly useful for interpreting the model's focus on spatial features critical for accurate classification. By applying CAMs, it is ensured that ConvNeXt attends to the relevant segments of the image in order to perform accurate classification. This visualization technique not only serves to deepen one's understanding of the model's internal mechanisms but also serves as a diagnostic tool to enhance model performance by identifying overlooked features in the training process.

4. Experiments

The aim of the first experiment is to gauge whether the ConvNeXt model can outperform the ViT model in terms of training efficiency and classification accuracy on the test set. The exploration begins by choosing a dataset and running a baseline classification for both models on the dataset of choice. Once the baseline is obtained, hyper parameter tuning will be conducted to get the highest possible classification accuracy for both models and their performances will be evaluated [24].

The aim of the second experiment is to compare the generalizability of the ConvNeXt to that of the ViT. The ConvNeXt model is known to transfer its learning amongst tasks well, but vision transformers are hailed for their generalizability. In order to conduct this exploration, another dataset is needed. This dataset has to be one with limited samples per class to truly test how well both models can transfer their learning onto the similar task of sign language recognition but on another signing language. Following the obtaining of baseline classification accuracies for both models, hyper parameter tuning will be conducted and final classification accuracies will be recorded for both models.

4.1. Datasets

4.1.1. Sign-Language-MNIST: The Sign-Language-MNIST dataset emerged as the optimal choice for the first experiment given its widespread use in other papers involving sign language recognition. The dataset itself consists of American Sign Language (ASL) letters and is a multi-class dataset consisting of classes 0-24 (letters 'J' and 'Z require motion and are excluded) as shown in **figure 4**. It consists of 27,455 28 × 28 grayscale training images and 7,172 28 × 28 grayscale test images.

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363

Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

Figure 4. ASL signs and their corresponding classes.

Upon doing initial data exploration, it was discovered that the distribution of training data amongst the classes was fairly even: between 900-1300 images per class. Furthermore, outlier detection was conducted using PCA and 3100 outliers were found using an image reconstruction error threshold of 0.99. The image data was then reshaped from a flattened array to a 28 × 28 array and class labels 0-25 were remapped to 0-23 to exclude classes 'J' and 'Z' for which there was no data. The data was then resized to a resolution of 224 × 224 and converted to image and label tensors [25]. Finally, the training data was split into 80% training data and 20% validation data for hyper parameter tuning **Figure 5**.

Figure 5. Examples of samples from the ISL dataset.

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

4.1.2. Indian Sign Language: The ISL dataset was selected to evaluate the generalizability of the models, considering the constraints of data availability for all classes. The dataset consists of images for 23 signs excluding the 'H', 'J', and 'Y' classes and each image is 126 × 126 × 3 pixels. There are mere 20-50 samples available for each class, but this is ideal when evaluating model generalization capabilities as in **figure 5**.

This is done to add variability and realism akin to real-world situations and serves the purpose of testing model robustness. The noise is in the form of background replacement using blurry, colorful, and messy backgrounds in order to emulate dynamic conditions. The data is loaded and processed in much the same way as the sign-language-MNIST dataset and finally divided into 60% training, 20% validation, and 20% test sets for each class.

4.2. ConvNeXt

We fine-tune the ConvNeXt model along with the custom classification head. Initially, baselines were recorded for both models using a batch size of 32 and a learning rate of 0.001 for the ConvNeXt model on the Sign-Language-MNIST dataset. After recording the baselines, the model's hyper parameters are tuned and the classification accuracies of the best model are recorded for comparison.

Following the tests on the initial dataset, the best ConvNeXt model is loaded and used to record baseline accuracies for the ISL dataset. Following this, the model undergoes hyper parameter tuning and we record the best classification accuracies for the generalization experiment [26].

4.3. ViT

The fine-tuning of the ViT model occurs in the same way as that of the ConvNeXt model using the Sign-Language-MNIST dataset. Baseline performance was recorded for the model using a batch size of 64 and a learning rate of 0.0003. After optimizing these hyper parameters, the best classification accuracy of the ViT was established for the dataset.

Following the baseline assessments, we proceed with determining the generalizability of the model on the ISL dataset. This decision is well informed due to the limited volume of data available for fine-tuning, which facilitates model adaptability for nuanced sign language features. The baseline model is established, hyper parameter tuning is conducted, and the best classification accuracies are then recorded on the ISL dataset.

5. Results

5.1. ConvNeXt

Accuracies on the Sign-Language-MNIST test set and training times were recorded for the ConvNeXt model. The best hyper parameters found for the ConvNeXt model were a batch size of 64 and a learning rate of 0.002. The loss of the tuned model over 5 epochs is presented in **figure 6.** and the corresponding classification accuracies by class are represented in **figure 7**. Via the confusion matrix.

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363

Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

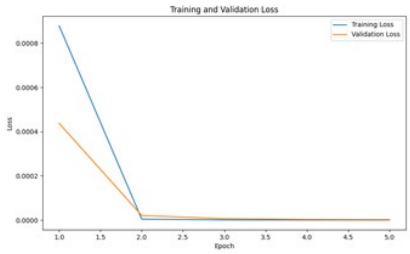

Figure 6. Training and validation loss of the tuned ConvNeXt model over 5 epochs on sign-language-MNIST.

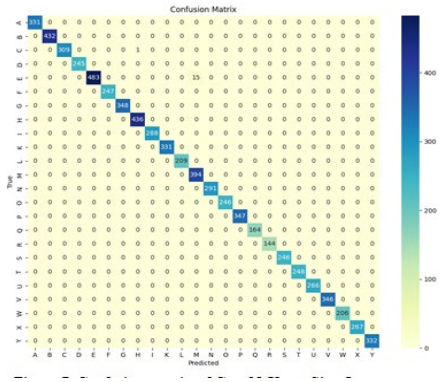

Figure 7. Confusion matrix of ConvNeXt on Sign-language-MNIST.

We then proceeded to record the accuracies of the ConvNeXt on the ISL dataset as part of the generalizability evaluation. The best hyper parameters found were a batch size of 4 and a learning rate of 0.0015. The best model is then used to record classification accuracies on the test set and these results are presented in **figure 8.**

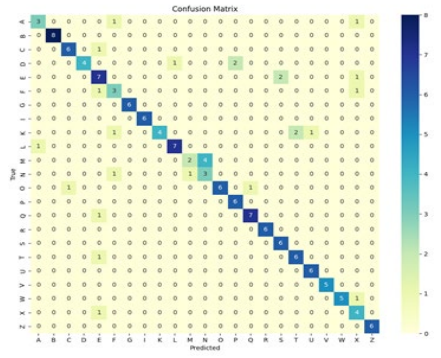

Figure 8. Confusion matrix of ConvNeXt on ISL.

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

5.2. ViT

The experimental process of the ViT was the same as that of the ConvNeXt. We found the best hyperparameters for this model to be a batch size of 32 and a learning rate of 0.001. ViT is the classification accuracy of the model on each class of the Sign-Language-MNIST dataset. Finally, represents the confusion matrix of the tuned ViT model, with batch size 1 and learning rate 3.9^{e-} 5, on the ISL dataset.

Both the ViT and the ConvNeXt achieved comparable and superior accuracies >99% on the test set and this was as hypothesized. However, ConvNeXt proved to be more robust to hyper parameter changes and consistently performed well as compared to the Vision transformer. Moreover, ConvNeXt required 6x less time to train than the ViT as seen in **table 1**. Accuracies on the ISL dataset are presented in **table 2**. The ViT model boasts superior generalizability as was hypothesized but the ConvNeXt model can be seen to transfer its learning to a similar task quite adequately. Both models achieved accuracies of >80%.

Table 1. Results on the Sign-Language-MNIST dataset.

Model	Training Time (mins)	% Accuracy
ConvNext-Tiny (Baseline)	12.3	98.4
ConvNext-Tiny (Tuned)	10.8	99.7
ViT (Tuned)	70.5	99.9

Table 2. Results on the ISL dataset.

Model	% Accuracy
ConvNext-Tiny (Baseline)	71.3
ConvNext-Tiny (Tuned)	81.3
ViT (Baseline)	62.5
ViT (Tuned)	84.6

6. Discussion

The first noticeable advantage of the ConvNeXt over the ViT was the training time. This is less significant on smaller datasets like the ISL but grows exponentially with the amount of data. The slower training times of the ViTs can be attributed to their attention mechanism which compares every image patch to every other patch and the quadratic complexity of this operation can become computationally expensive with an increasing number of patches. ConvNeXt, on the other hand, uses convolution to capture spatial information and this is a much faster operation than self-attention, boasting linear complexity. Another significant factor affecting training time could be the lack of inductive bias present in ViTs. Convolutions inherently assume locality and spatial hierarchies in an image which allows for more efficient spatial learning of features. ViT lacks these inductive biases and requires more data to

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

learn these spatial hierarchies. A final factor could be the global processing of images done in ViTs which is useful in capturing long-range dependencies but does require more compute.

Another noticeable point is the significant increase in classification accuracy between the baseline ViT and tuned ViT. This is indicative of the fact that batch sizes and learning rates are pivotal for transformer performance. This could be a result of the sequential nature of the data representation that occurs within the transformer architecture.

It becomes evident that both models have classified everything correctly except for 1-2 classes. To get a better idea of why this misclassification was occurring in the ConvNeXt model, class activation maps were used to overlay weights of the last stage of convolution onto the input image. Upon reviewing the class activation maps, it became clear that signs for 'E' and 'M' differed ever so slightly. The granularity present in the input image itself might have led to additional noise and eventual misclassification. However, an interesting question is why the model did not misclassify, for example, sign 'M' as sign 'N', both being very similar in their gesture. The generalizability of the ConvNeXt model is not as powerful as that of the ViT as there exist a 3% difference between their classification accuracies on the ISL dataset. ConvNeXt aims to emulate the non-local attention mechanism present in transformers by using larger, depth-wise convolutions. However, it is merely an approximation and is not able to represent global relationships in the data as well as the attention mechanism itself. The exhaustive comparison of image patches means that with enough data, transformers are able to learn strong patterns and transfer their learning better to similar tasks. Convolution operations inherently lack this global property and, therefore, ConvNeXt was more susceptible to the addition of controlled noise in the ISL data through the class activation maps.

7. Conclusion

The goal of this exploration was to determine the transfer learning and generalization capabilities of ConvNeXt models in a task such as sign language recognition. To further set a benchmark for comparison, the same study was conducted using ViTs. The ConvNeXt was able to achieve almost the same accuracy as a vision transformer but offered a huge boost in training time on a dataset with sufficient samples, such as the sign-language-MNIST dataset. However, it did not generalize as well as the ViT on a dataset with limited samples, such as the ISL dataset and was susceptible to misclassification due to the addition of controlled noise. This work does highlight the power of convolution based models despite their recent drop in popularity. They are lightweight and flexible models that are sufficient for tasks that do not require learning extremely long-range dependencies. Finally, to the best of our knowledge, this work is one of the first to utilize the ConvNeXt model for the task of ISLR.

8. Future Work

In the current research, the verification of model scalability was constrained due to the limited computational resources available, but this presents a promising avenue for future studies. One could focus on assessing the performance of models like ConvNeXt and ViTs under varying computational loads and with larger, more complex datasets. This would provide further insights into model robustness and behavior in more demanding scenarios like real-world situations where model efficiency is critical.

Another direction could involve optimizing these models for real-time sign language recognition. This does not only involve refining the model for speed and accuracy but also ensures the robustness of the model in dynamic

Darda A. International Journal of Computer Science and Mobile Applications, Vol. 12 Issue 10, October - 2024, pg. 01-15.

ISSN: 2321-8363

Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

environments. Future research could involve developing techniques to reduce latency and altering the models to be able to process video data, thus enabling their use in practical settings such as live translation for the deaf community.

The most exciting prospect lies in the cross-cultural adaptability of models where the goal would be to train a universal sign language recognition system capable of understanding multiple sign languages. This would require addressing challenges such as variability in sign vocabulary, syntax, and usage, making this endeavor difficult but one that is highly impactful as an accessibility application

References

1. Tucci, Debara L.,et al. "A summary of the literature on global hearing impairment: current status and priorities for action." *Otol Neurotol.* 31.1 (2010): 31-41.
2. Guo, Tianmei, et al. "Simple convolutional neural network on image classification." *2017 IEEE 2nd Int Conf Big Data Anal. (ICBDA)*. IEEE, 2017.
3. Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." *IEEE Trans Pattern Anal Mach Intell.* 44.7 (2021): 3523-3542.
4. Dhillon, Anamika, & Gyanendra K. Verma. "Convolutional neural network: a review of models, methodologies and applications to object detection." *Prog Artif Intell.* 9.2 (2020): 85-112.
5. Hu, Lianyu, et al. "Continuous sign language recognition with correlation network." *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2023.
6. Badhe, Purva C., & Vaishali Kulkarni. "Indian sign language translator using gesture recognition algorithm." *IEEE Int Conf Comput Graph Vis Inf Secur. (CGVIS)*. IEEE, 2015.
7. Vázquez-Enríquez, Manuel, et al. "Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks." *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2021.
8. Liu, Z., et al. "ConvNet for the 2020s. 10 (2022).
9. LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." *Neural Comput* 1.4 (1989): 541-551.
10. Krizhevsky, Alex,et al. "Imagenet classification with deep convolutional neural networks." *Adv Neural Inf Process Syst.* 25 (2012).
11. He, Kaiming, et al. "Deep residual learning for image recognition." *Proc IEEE Conf Comput Vis Pattern Recognit*. 2016.
12. Koller, Oscar,et al. "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs." *Proc IEEE Conf Comput Vis Pattern Recognit*. 2017.
13. Pathan, Refat Khan, et al. "Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network." *Sci Rep.* 13.1 (2023): 16975.

ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

14. Jiang, Lingjie, et al. "JujubeNet: A high-precision lightweight jujube surface defect classification network with an attention mechanism." *Front Plant Sci.* 13 (2023): 1108437.
15. Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." 2010.11929 (2020).
16. Alrfou, Khaled,et al. "Computer vision methods for the microstructural analysis of materials: the state-of-the-art and future perspectives." 2208.04149 (2022).
17. Bilgin, Metin, & Korhan Mutludoğan. "American sign language character recognition with capsule networks." *3rd Int Symp Multidiscip Stud Innov Technol (ISMSIT)*. IEEE, 2019.
18. Sabeenian, R. S.,et al. "Sign language recognition using deep learning and computer vision." *J Adv Res Dyn Control Syst* 12.5 Special Issue (2020): 964-968.
19. Mannan, Abdul, et al. "[Retracted] Hypertuned Deep Convolutional Neural Network for Sign Language Recognition." *Comput Intell Neurosci.* 2022.1 (2022): 1450822.
20. Kadous, Mohammed Waleed. "Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language." *Proc Workshop Integr Gesture Lang Speech.* 165, 1996.
21. Cooper, Helen, et al. "Sign language recognition using sub-units." *J Mach Learn Res.* 13.1 (2012): 2205-2231.
22. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proc IEEE/CVF Int Conf Comput Vis*. 2021.
23. Hendrycks, Dan, & Kevin Gimpel. "Gaussian error linear units (GELUs)." 1606.08415 (2016).
24. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *IEEE Conf Comput Vis Pattern Recognit*, 2009.
25. Ashish, Vaswani. "Attention is all you need." *Adv Neural Inf Process Syst.* 30 (2017): I.
26. Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proc. IEEE Conf Comput Vis Pattern Recognit*. 2016.