Sign Language Enhanced Image Analysis Techniques Using Multiscale Feature Extraction and Attention Mechanisms

for Arabic Sign Language Recognition

Achraf Kamni
Sakher Yaish

01

Introduction

Introduction

- Problem: Many image analysis methods miss small details which limits recognition accuracy.
- Focus: Improving image recognition for Arabic Sign Language.
- Techniques: Multiscale feature extraction, spatial-reduction attention, progressive dimensional reduction.

02

Objective

Objective

- Implement a MVTN with a custom resnet
- Capture both detailed and contextual features.
- Emphasize important regions within images.
- Compare with models like ResNet, ViT, and GoogleNet.

03

Background

Background

- Transformers: Initially for NLP, now used in image recognition.
- Multiscale Attention: Captures features at various scales for detail and context.
- Related Work: Based on recent advancements in sign language recognition.

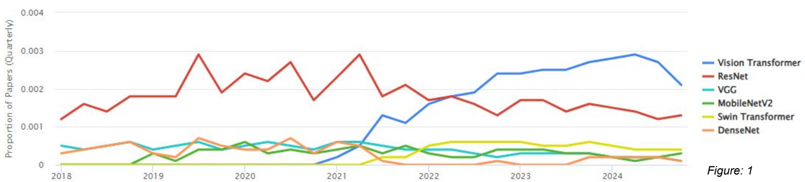


Figure: 1

04

Proposed Approach

Proposed Approach

- Multiscale Feature Extraction: Capture detailed/contextual features.
- Spatial-Reduction Attention: Focus on key regions.
- Dimensional Reduction: Preserve crucial information with reduced complexity.

05

Dataset & Implementation

Dataset & Implementation

- Dataset: ArASL Database Grayscale (Arabic Sign Language).
- Data Processing: Images resized to 224x224, and normalized.
- Custom Dataset Class: Efficiently loads, preprocesses, and labels images.

Figure: 2

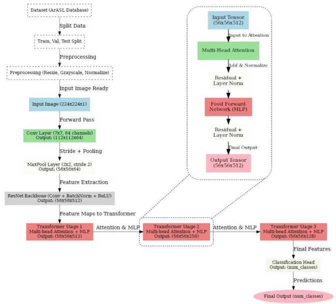
06

Model Architecture

Model Architecture

- CNN Backbone: ResNet-18 for grayscale images.
- Multiscale Transformer: Uses self-attention, multi-head attention, and dimension reduction.

Model Architecture



07

Training Setup

Training Setup

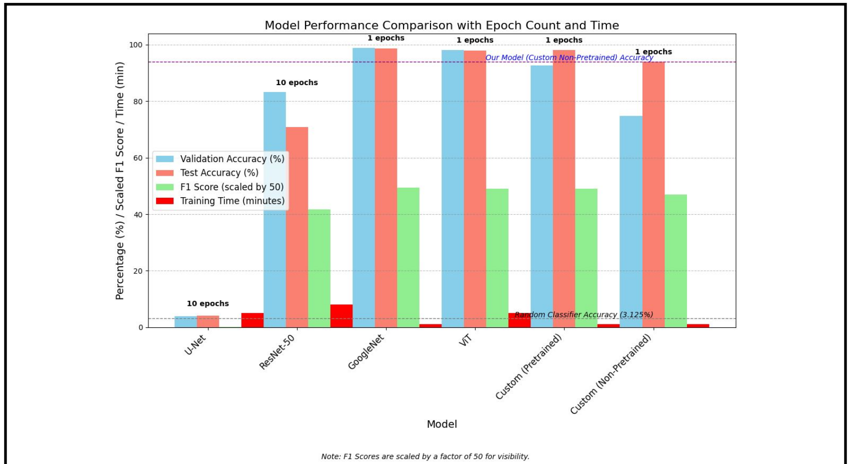
- Loss Function: Cross-Entropy.
- Optimizer: Adam with learning rate of 0.0001.
- Hardware: GPU A100.

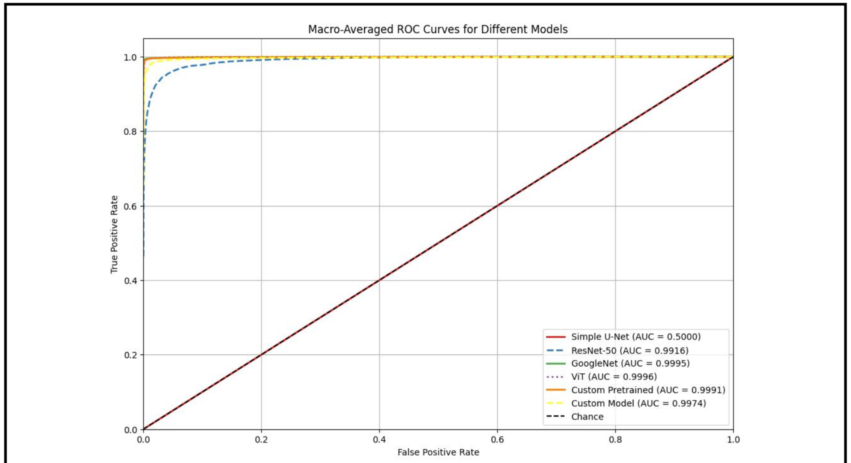
08

Results Overview

Results Overview

Model	Test Loss	Test Accuracy	F1	F2	Precision	Recall	AUC
U-net	3.4595	4.10%	0.0025	0.0055	0.0013	0.0312	0.5
Resnet-50	0.7056	82.75%	0.8286	0.8256	0.8474	0.8267	0.9916
GoogleNet	0.0492	98.99%	0.99	0.99	0.9899	0.9901	0.9995
ViT	0.1022	96.98%	0.9701	0.9699	0.9712	0.97	0.9996
Custom MVTN(pretrained resnet)	0.1119	97.65%	0.9773	0.9769	0.9787	0.9767	0.9991
Custom MVTN(non-pretraine d resnet)	0.3272	91.14%	0.9034	0.9032	0.9354	0.9069	0.9974





09

Conclusion

Conclusion

- Key Findings: Deep models with multiscale feature extraction improve accuracy for complex tasks.
- Future Work: Use of pretrained models and transfer learning for enhanced performance.

Demo

References

Figure 1:
Papers with Code. (n.d.). Vision Transformer. Retrieved from https://paperswithcode.com/method/vision-transformer

Figure 2:
Mnasri, S., Ouarda, W., & Bellil, W. (2019). Sign language recognition: A survey. Data in Brief, 25, 104255. https://doi.org/10.1016/j.dib.2019.104255

THANK YOU