Deep Learning Project Report

(2025W) COMP-5421-WA - Deep Learning

Stephon Phillips (1208075)
Naresh Jung Shahi (1271540)
Jay Chiedozie Nwokocha (1296781)

Professor:
Dr. Saad Bin Ahmed

Addressing Translation variance in Traffic Sign Classification Using Vision Transformers

Stephon Phillips
Department of ComputerScience
Faculty of Computer Science,
Lakehead University
Thunder Bay, Canada
sphilli8@lakeheadu.ca

Naresh Jung Shahi
Department of Computer Science
Faculty of Computer Science,
Lakehead University
Thunder Bay, Canada
nshahi3@lakeheadu.ca

Jay Chiedozie Nwokocha
Department of Computer Science
Faculty of Computer Science,
Lakehead University
Thunder Bay, Canada
cnwokoch@lakeheadu.ca

***Abstract*—This paper investigates the use of Vision Transformers (ViTs) to address the translation variance problem in image classification tasks, specifically in traffic sign recognition. While convolutional neural networks (CNNs) have been the standard approach for handling translation-invariant features, their ability to cope with positional shifts in images is limited. By leveraging the self-attention mechanism of transformers, we propose a method that enhances translation robustness. Our approach involves preprocessing traffic sign images through pixel embedding using a bidirectional mechanism, followed by a series of transformer blocks and fully connected layers. The model's performance is evaluated on a publicly available traffic sign dataset, demonstrating promising results in classification accuracy and robustness to image translations. The results highlight the potential of transformers in overcoming translation variance and improving the generalization of deep learning models in computer vision tasks.**

Keywords—Vision Transformer, Translation Invariance, Traffic Sign Classification, Image Classification

I. INTRODUCTION (*HEADING 1*)

In deep learning, translation invariance refers to the model's ability to correctly classify or predict objects regardless of their position in an image. Translation invariance is a critical challenge in image classification tasks, particularly in applications like traffic sign recognition, where signs may appear at different positions within an image. While Convolutional Neural Networks (CNNs) are commonly used for image classification tasks, their local receptive fields can make them vulnerable to shifts in image features, affecting model robustness [1]. This paper explores the use of Vision Transformers (ViTs) to address this issue. Transformers, originally developed for Natural Language Processing (NLP), have recently been applied to image classification tasks. Their self-attention mechanism provides the ability to focus on relevant features regardless of their spatial location, making them ideal candidates for handling translation invariance. The objective of this work is to evaluate the performance of a transformer-based model in traffic sign recognition, with a focus on addressing translation invariance. The proposed method is evaluated on a publicly available traffic sign dataset, and its performance is analyzed in terms of accuracy and robustness to translations.

II. LITERATURE REVIEW

Translation invariance has long been a challenge in image classification, with early solutions involving techniques such as pooling layers in CNNs [2]. These methods, while effective in some cases, can still suffer from positional shifts that affect model performance. Recent advancements in transformer-based models have shown promise in addressing these issues due to their self-attention mechanism, which allows for the modeling of long-range dependencies [5]. Vision Transformers (ViTs) have demonstrated state-of-the-art performance in several image classification benchmarks, surpassing CNNs in tasks requiring global contextual understanding [6]. Additionally, various techniques like data augmentation, positional encoding, and pixel embedding have been proposed to enhance the robustness of deep learning models to translation invariance [3]. This paper builds upon these advancements by applying ViTs to traffic sign classification and evaluating their performance under translation shifts.

With recent advancements in transformer architectures, such as Vision Transformers (ViTs), it has demonstrated superior performance in image classification tasks by leveraging self-attention mechanisms. ViTs divide images into patches, embed them into a high-dimensional space, and process them using transformer blocks. This approach allows the model to capture global context and spatial relationships, making it inherently more robust to translations [7]. Several studies have explored the application of ViTs in various domains, but their effectiveness in traffic sign classification remains understudied. This paper builds on these insights to propose a transformer-based solution for traffic sign classification.

III. METHODOLOGY

A. Dataset Description

The Traffic Sign Classification dataset used in this study consists of various traffic sign images collected from the real-world traffic environments. The dataset is divided into training, validation, and testing sets. Each image is categorized into one of the 58 classes, representing different traffic sign types. The dataset is divided into training, validation, and test sets, with images resized to 224x224 pixels for consistency.

B. Preprocessing

The images are resized to a uniform size of 224x224 pixels. A rescaling operation normalizes the pixel values to the range [0, 1]. Data augmentation techniques such as translations are applied to simulate real-world shifts and test the model’s translation invariance.

C. Model Architecture

A Vision Transformer (ViT) model is constructed with an embedding dimension of 768 and 12 transformer blocks. The images are first converted into patches, and a special token is added to the patch sequence to facilitate classification.

Positional encoding is applied to each patch to preserve spatial relationships. The transformer block consists of multi-head self-attention and feed-forward layers, followed by a classification head for the final prediction. The model is

trained using the Adam optimizer and categorical cross-entropy loss [4].

IV. Performance Evaluation

A. Training and Validation

The model is trained for 32 epochs, and its performance is evaluated on the validation set. Training and validation loss curves are plotted to assess overfitting or underfitting.

B. Testing

Our test dataset does not have labels; therefore, we relied on predictions combined with human judgment. This means we generated predictions and manually verified them. Additionally, the model's robustness to translations is evaluated by applying horizontal and vertical shifts to the test images and measuring their impact on classification accuracy.

C. Results

- Accuracy: The model achieves an validation accuracy of 77% .

Translation Invariance: To test translation variance, we first made a prediction on a random image. Then, we modified the image by shifting it 60 pixels up, down, right, and left, creating four different versions. We then tested all the translated images and found that only one prediction differed from the original test image's prediction. This result demonstrates that translation variance can be addressed using a transformer architecture.

Fig. 1. Showing original prediction without shifting

Fig. 2. Showing prediction with image shifted to the right

Fig. 3. Showing prediction with image shifted to the left

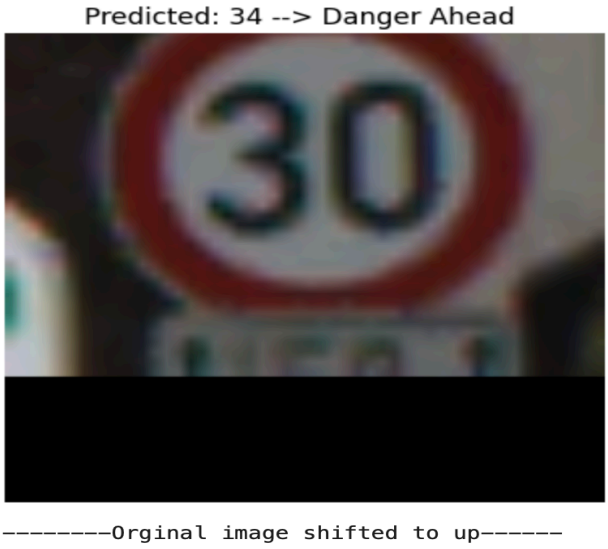

Fig. 4. Showing prediction with image shifted up

Fig. 5. Showing prediction with image shifted down

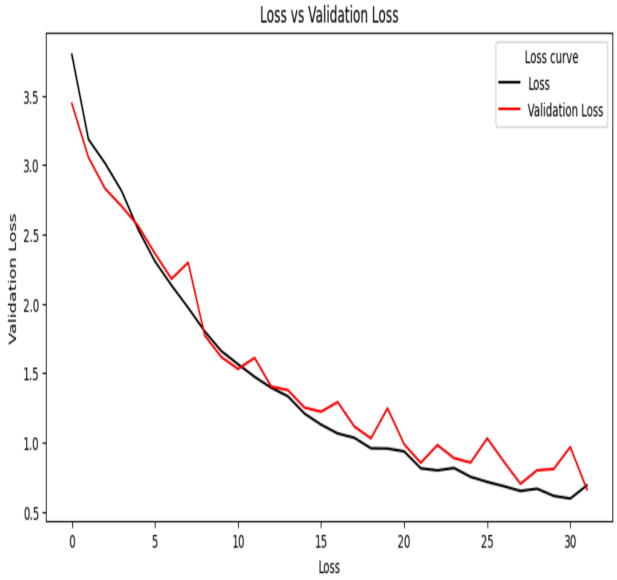

Fig. 6. Showing Loss VS Validation Loss Checking for Overfitting

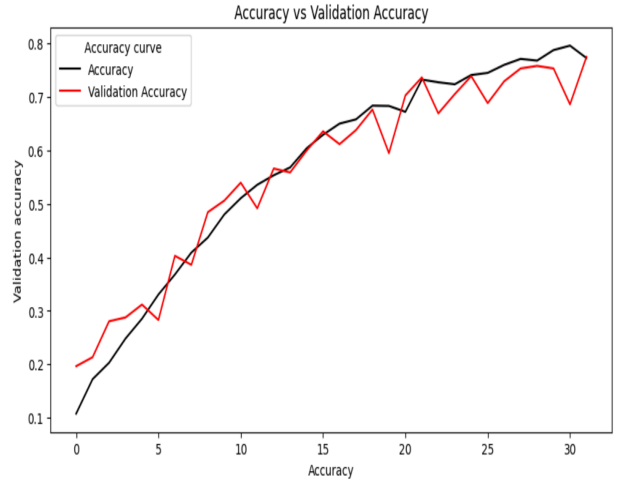

Fig. 6. Showing Accuracy Plot

V. Analysis and Findings

The results demonstrate that the Vision Transformer model is effective in handling translation variance. The self-attention using bidirectional mechanism allows the model to capture global context, making it less sensitive to the position of objects within the image.

VI. Conclusion

This paper presented a Vision Transformer-based model for traffic sign classification and demonstrated its robustness to translation shifts. The results indicate that transformers, with their self-attention mechanism, can effectively mitigate the translation invariance problem that traditionally affects CNN-based models. Future work could explore techniques to reduce the model's computational complexity and extend its application to other image classification tasks. It can also explore further optimizations, such as hybrid models combining CNNs and transformers, or the use of larger, more diverse datasets to evaluate the generalizability of the approach. The success of this model in traffic sign recognition suggests potential applications in autonomous driving systems, where translation invariance is crucial for reliable performance.

References

1. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, Massachusetts: The MIT Press, 2016. Available: https://www.deeplearningbook.org/
2. Rajalingappaa Shanmugamani, Deep learning for computer vision : expert techniques to train advanced neural networks using
3. TensorFlow and Keras. Birmingham, Uk: Packt Publishing, 2018.
4. R. Szeliski, COMPUTER VISION : algorithms and applications. S.L.: Springer Nature, 2020.
5. M. Ekman, Learning Deep Learning. Addison-Wesley Professional, 2021.
6. J. Bi, Z. Zhu, and Q. Meng, "Transformer in Computer Vision," 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Sep. 2021, doi: https://doi.org/10.1109/cei52496.2021.9574462.
7. M. Elgendy, Deep learning for vision systems. Shelter Island, Ny Manning Publications Co, 2020.
8. D. Rothman, Transformers for Natural Language Processing and Computer Vision. Packt Publishing Ltd, 2024.