Comparative Analysis of Fine-Tuning Approaches for Ship Classification Using Transformers

James Park
Stanford University
jpark22@stanford.edu

Jean Laguerre
Stanford University
jeanlag1@stanford.edu

Abstract

This paper investigates the effectiveness of Google's Vision Transformers (ViT) and Microsoft's Swin Transformers for the task of ship classification. We provide an overview of a dataset that includes 8,908 images labeled into five ship categories. The ***Methods*** *section details the architectures of the ViT and Swin models and the fine-tuning procedures used to improve their performance. In the* ***Experiments, Results, and Discussion*** *section, we present experimental results that demonstrate how fine-tuning significantly enhances classification accuracy, with the ViT model achieving a higher accuracy than the Swin Transformer. Finally, the* ***Conclusion and Future Work*** *section highlights the importance of selecting appropriate learning rates and suggests future work, including cross-validation and domain-specific pre-training, to further enhance model performance. This study advocates for the application of advanced machine learning techniques in maritime surveillance and defense systems.*

1. Introduction

The US Navy employs various methods to classify ships to ensure safe navigation, maintain situational awareness, gather intelligence, and manage logistics and supply chains efficiently. Traditional methods include the Automatic Identification System (AIS), radar and sonar, visual identification by lookouts and bridge personnel, and specialized teams like SNOOPIE (Ship's Nautical or Otherwise Photographic Interpretation and Examination). However, these methods face limitations. AIS signals can be manipulated or disabled by vessels attempting to evade detection. Radar and sonar signals are prone to clutter and interference. Visual identification relies on human accuracy, which can be compromised by limited visibility and the inherent risk of human error.

Despite extensive research in deep learning and computer vision, DOD systems have not kept up with these advancements. Machine learning can significantly improve ship classification by reducing errors caused by AIS manipulation or signal interference, automating the detection process, and integrating these capabilities into real-time systems. Our project advocates for the use of machine learning models in ship classification by fine-tuning **Google's Vision Transformer (ViT)** and **Microsoft's Swin Transformer** models on images of ships and offering a comparative analysis.

The dataset used in this study originates from a hackathon hosted by Analytics Vidhya in 2019. It comprises 8,908 ship images, of which 6,252 images with labeled categories were used for our experiments. These images are categorized into five specific ship types: cargo, military, carrier, cruise, and tankers. Each image is resized to 224x224 pixels, normalized, and converted into patches suitable for input into the ViT and Swin models. The input to our models is this collection of labeled ship images, prepared for processing by the ViT and Swin architectures.

The task of image classification involves assigning a single label to an image from a fixed set of categories. The challenges such as viewpoint variation, scale variation, deformation, occlusion, illumination conditions, background clutter, and intra-class variation are all relevant to the ship classification task. By addressing these challenges, machine learning models can improve the robustness and accuracy of ship classification systems.

The outputs include:

- **Baseline Accuracy**: The classification accuracy of the pre-trained ViT and Swin models on the test set without any fine-tuning.
- **Fine-Tuned Models**: Improved classification models achieved by fine-tuning ViT and Swin on the training dataset with different learning rates and epochs, expected to show enhanced accuracy over the baseline models.
- **Comparative Analysis**: Evaluation of the performance of both models, comparing their accuracy and

training/validation loss across a set of different learning rates to determine the most effective fine-tuning technique for ship classification.

In summary, this paper aims to contribute to the broader advocacy for the implementation of machine learning in operational systems within the DOD. By focusing on the specific application of ViT and Swin Transformer models for ship classification, we demonstrate how advanced machine learning techniques can address the limitations of traditional methods.

2. Related Work

Fine-grained image classification has been a challenging task due to the subtle differences between classes. Various methods have been proposed to improve the accuracy and efficiency of classification models in different domains, including ship classification.

2.1. Data Augmentation and Transfer Learning

Milicevic et al. [13] addressed the challenge of limited data by using data augmentation and transfer learning for ship classification. They demonstrated that artificial data creation and pre-trained models could significantly enhance classification accuracy when data is scarce. Their approach involved geometric transformations such as rotations, flips, and cropping to create additional training samples, and the use of pre-trained convolutional neural networks (CNNs) to leverage knowledge from related domains. Their results showed that these techniques could effectively reduce overfitting and improve model generalization.

2.2. Multi-Feature Region Approaches

Fayou et al. [12] proposed a multi-feature region approach combined with attention mechanisms to improve fine-grained visual recognition. Their model, MRA-CNN, incorporated Convolutional Block Attention Modules (CBAM) to enhance feature representation and the k-means clustering algorithm to select multiple discriminative regions within an image. This method allowed the model to focus on several key parts of the object, improving classification performance on benchmarks such as CUB-200-2011, Stanford Cars, and FGVC-Aircraft. The integration of multiple feature regions and attention mechanisms demonstrated significant performance gains over single-region approaches like RA-CNN.

2.3. Vision Transformers

The Vision Transformer (ViT) model introduced by Dosovitskiy et al. [1] leveraged transformer architectures for image classification, highlighting the benefits of self-attention mechanisms over traditional convolutional networks. By treating an image as a sequence of patches, ViT effectively captured long-range dependencies and global context, which are crucial for tasks requiring fine-grained details. Their experiments on ImageNet and other datasets showed that ViT could achieve competitive accuracy with fewer computational resources compared to state-of-the-art CNNs, marking a significant shift in image classification techniques.

Liu et al. [11] further extended this idea by developing the Swin Transformer, a hierarchical vision transformer that achieved excellent results on various vision tasks by incorporating shifting windows. The Swin Transformer maintained computational efficiency while capturing both local and global features through its hierarchical architecture. This approach allowed it to perform well on tasks with high-resolution images, such as object detection and semantic segmentation, making it a versatile model for different computer vision applications.

2.4. Foundational Transformer Models

The concept of using transformers in vision tasks was significantly advanced by Vaswani et al. [4], who introduced the transformer model in the context of natural language processing. Their model, which utilized self-attention mechanisms to process sequential data, laid the foundation for subsequent applications in computer vision. The ability of transformers to handle long-range dependencies and parallelize computations made them suitable for tasks beyond NLP, inspiring the development of models like ViT and Swin Transformer.

2.5. Deep Convolutional Networks

Further improvements in fine-grained classification have been explored by He et al. [8] with the ResNet model, which introduced residual learning to facilitate the training of deeper networks. ResNet's skip connections addressed the vanishing gradient problem, allowing for the construction of very deep networks that achieved state-of-the-art performance on image classification tasks.

Szegedy et al. [6] proposed the Inception architecture, which utilized multiple convolutional kernels of different sizes to capture diverse features within images. This multi-scale approach improved the network's ability to detect objects at various scales and orientations, contributing to its success in competitions like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

Simonyan and Zisserman [17] worked on VGGNet, a deep convolutional network that emphasized simplicity and depth. VGGNet's uniform architecture, consisting of repeated 3x3 convolutional layers, demonstrated that increasing network depth could significantly improve performance, setting a new benchmark for deep learning models in image classification.

2.6. Efficient Models for Mobile Applications

Howard et al. [2] developed MobileNets, which aimed at efficient models suitable for mobile and embedded vision applications. These models employed depthwise separable convolutions to reduce the number of parameters and computational complexity, enabling real-time image processing on devices with limited resources. MobileNets have been widely adopted in applications requiring low-latency and high-efficiency, such as autonomous vehicles and augmented reality.

2.7. Transfer Learning Techniques

Transfer learning techniques have also been a focal point in many studies. Yosinski et al. [7] explored the transferability of features in deep neural networks, concluding that early layers capture more generic features suitable for various tasks, while later layers learn task-specific features. This insight has led to the widespread practice of using pre-trained models as feature extractors or initializing weights for new tasks, reducing training time and improving performance.

Pan and Yang [14] provided a comprehensive survey on transfer learning, categorizing different methods and applications. They discussed various strategies such as domain adaptation, multi-task learning, and self-taught learning, highlighting the versatility and effectiveness of transfer learning in scenarios with limited labeled data.

2.8. Maritime Vessel Datasets

In the context of ship classification specifically, Solmaz et al. [5] made a substantial contribution by creating the MARVEL dataset, a large collection of maritime vessel images, which facilitated the development and benchmarking of new classification models. This dataset includes a diverse range of vessel types and conditions, providing a robust testbed for evaluating the performance of different ship classification algorithms. The availability of such datasets has been crucial for advancing research in maritime surveillance and automated vessel identification.

These studies collectively illustrate the advancements in fine-grained image classification and the potential of leveraging data augmentation, transfer learning, and transformer-based models to enhance performance in specific applications such as ship classification.

3. Methods

In this study, we employed two state-of-the-art vision transformer models, namely the Vision Transformer (ViT) [1] and the Swin Transformer [11], for the task of ship classification. This section details the model architectures, training procedures, preprocessing steps, and evaluation metrics used in our experiments.

3.1. Model Architectures

Vision Transformer (ViT): The Vision Transformer (ViT) represents a departure from traditional CNNs by leveraging the transformer architecture, which was originally designed for natural language processing tasks. The ViT model used in this study, `google/vit-base-patch16-224-in21k`, utilizes a base configuration with a 16x16 patch size and 224x224 input resolution. The architecture comprises 12 transformer layers, each with 12 attention heads and a hidden size of 768. Input images are divided into non-overlapping patches, each linearly projected into a patch embedding. These embeddings are augmented with positional encodings to retain spatial information and then processed through the transformer layers. The self-attention mechanism within these layers enables the model to capture global dependencies and long-range interactions, which are crucial for fine-grained classification tasks such as ship classification [1].

Swin Transformer: The Swin Transformer introduces a hierarchical approach to vision transformers, incorporating a shifted window mechanism to balance efficiency and performance. The Swin model employed in this study, `microsoft/swin-tiny-patch4-window7-224`, features a tiny configuration with a 4x4 patch size and a 7x7 window size. The architecture is divided into four stages, each consisting of several transformer layers. Initially, images are split into non-overlapping patches, which are grouped into windows for local self-attention computation. These windows are shifted between layers to enable cross-window connections, allowing the model to capture both local and global contexts effectively. This hierarchical structure reduces spatial resolution progressively while increasing the feature dimension, enhancing the model's ability to capture fine-grained details across various scales [11].

3.2. Training Procedure

The training process involved fine-tuning the pre-trained ViT and Swin Transformer models on our ship classification dataset. We experimented with three different learning rates: **3e-3**, **5e-5**, and **8e-8**. The models were trained for 10 epochs, with a batch size of 128 and gradient accumulation steps of 4 to handle memory constraints. The Adam optimizer, known for its efficient handling of sparse gradients and adaptive learning rates, was employed with a warmup ratio of 0.1 to stabilize the training process. The chosen hyperparameters are as follows:

- **Learning Rates:** 3e-3, 5e-5, 8e-8.
- **Batch Size:** 128.
- **Epochs:** 10.

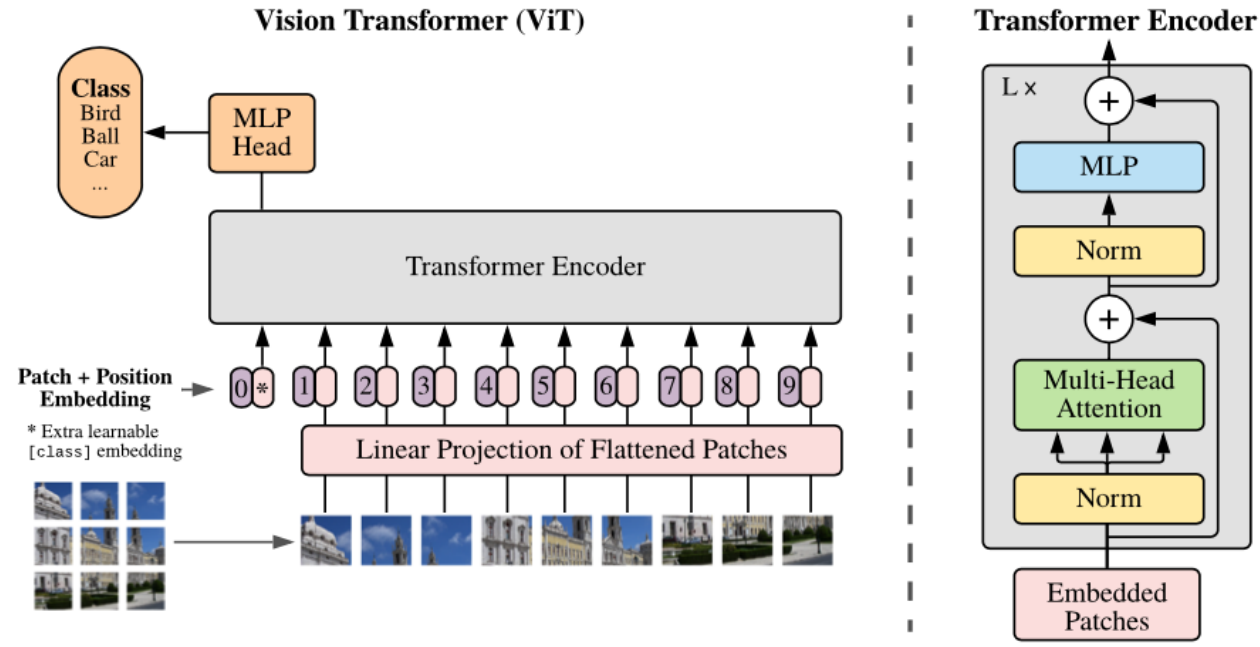

Figure 1. Architecture of Vision Transformer (ViT) [1]. The ViT model divides an image into fixed-size patches, which are linearly embedded and combined with positional encodings. These embeddings are processed through multiple transformer encoder layers, each comprising multi-head self-attention and feed-forward neural networks. The output is passed through an MLP head for final classification.

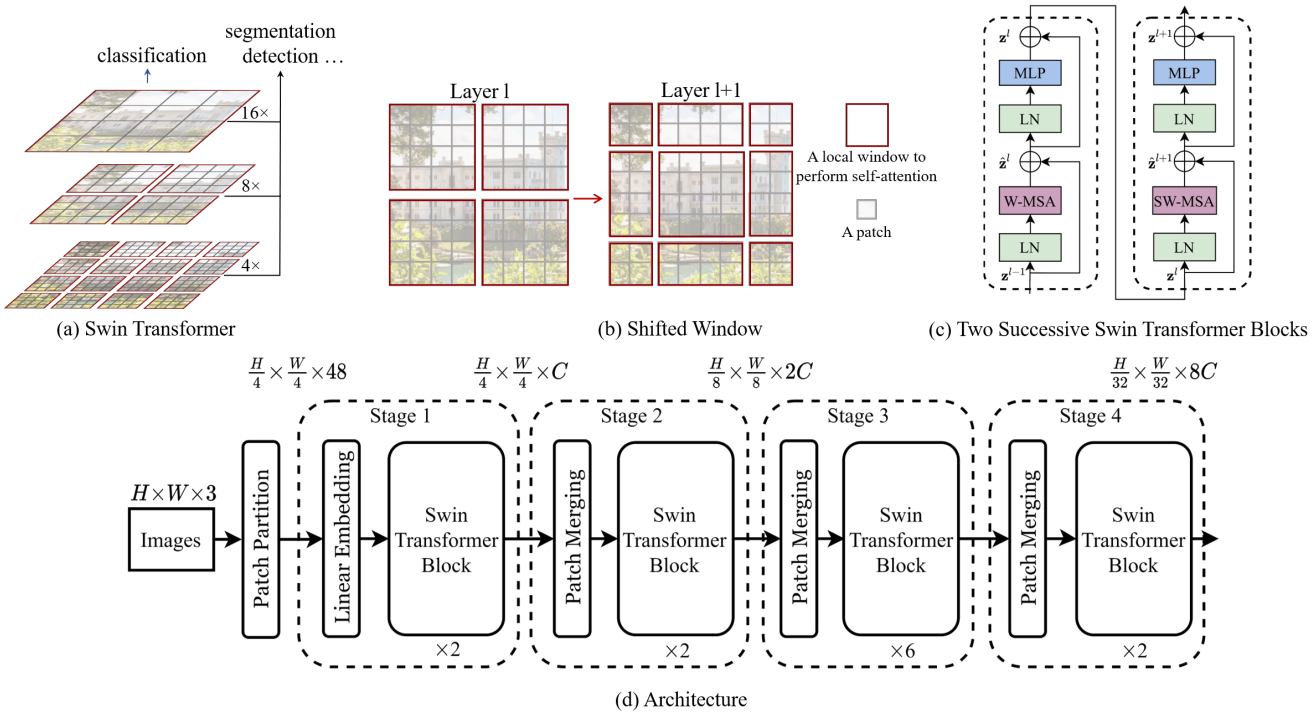

Figure 2. Architecture of Swin Transformer [11]. The Swin Transformer partitions the image into non-overlapping patches, which are processed through Swin Transformer blocks. These blocks use shifted windows for local self-attention, allowing efficient computation. Patch merging layers reduce spatial dimensions while increasing the feature dimension, enabling the model to capture multi-scale features.

- **Optimizer:** Adam with a warmup ratio of 0.1.

To manage the learning rate, a constant schedule was maintained throughout the training. The cross-entropy loss function, a standard choice for classification tasks, was used to measure the discrepancy between the predicted and true labels:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability. Initially, we tried training for 5 epochs to see if the model would overfit, but it didn't. We then extended the training to 10 epochs and found that the models still did not overfit. Consequently, we decided to go with 10 epochs, which yielded better results.

3.3. Preprocessing

We utilized the `AutoImageProcessor` from the Hugging Face Transformers library [10] to apply a series of transformations. The images were resized to 224x224 pixels to match the input resolution required by the models. This resizing step ensures uniformity across the dataset and compatibility with the model architectures. Pixel values were normalized using the mean and standard deviation values specific to each model, a step that helps in improving the convergence and generalization ability of the models. The normalized images were then converted into PyTorch tensors, the format required for training and inference in PyTorch.

3.4. Evaluation Metrics

The performance of the fine-tuned models was evaluated using accuracy (the accuracy metrics from the **evaluate** library). Accuracy measures the overall correctness of the model by dividing the number of correct predictions by the total number of predictions.

3.5. Implementation Details

The implementation of the ViT and Swin Transformer models was carried out using the PyTorch deep learning framework [3] and the Hugging Face Transformers library [10]. The training and evaluation scripts were executed on a high-performance computing cluster equipped with NVIDIA GPUs, ensuring efficient training and inference. The use of GPUs accelerated the training process, enabling the handling of large batches and complex model architectures.

4. Dataset and Features

As mentioned in the introduction, the dataset used in this study consists of 8,908 ship images sourced from a hackathon hosted by Analytics Vidhya in 2019. 6,252 of these images had labeled categories, which we used for our experiments. The images are categorized into five classes: **cargo, military, carrier, cruise**, and **tankers**. Each class represents a distinct type of ship, with varying visual characteristics and purposes.

4.1. Dataset Description

Class	Number of Images	Description
Cargo	1780	Large vessels for transporting goods
Military	1780	Naval ships with weapons and radar
Carrier	1780	Ships with flight decks for aircraft
Cruise	1784	Passenger ships for leisure travel
Tankers	1784	Ships for transporting liquids

Table 1. Summary of the ship dataset.

To ensure compatibility with the vision transformer models, all images were resized to a resolution of 224x224 pixels. The pixel values of the images were normalized using the mean and standard deviation values specific to each model. The normalization step improves the models' convergence and generalization ability.

Figure 3. Example images from the ship dataset.

4.2. Data Loading and Splitting

The labeled dataset of 6,252 images was split into a training set containing 4,376 images and a test set with 1,876 images, representing a 70%-30% train-test split. The dataset was loaded using the Hugging Face Datasets library [9], and processed in batches using the DataLoader class from the PyTorch library [3].

5. Experiments, Results, and Discussion

5.1. Baseline Performance

Before fine-tuning, we evaluated the baseline performance of the pre-trained Vision Transformer (ViT) and Swin Transformer models on the ship classification task. The baseline performance metrics provide a reference point to measure the improvement achieved through fine-tuning. For the baseline evaluation, the ViT model achieved an accuracy of 2.99%, while the Swin Transformer achieved an accuracy of 4.10%. These results indicate the models' initial performance when directly applied to the ship classification task without domain-specific fine-tuning.

Metric	Value
Correct (ViT)	56
Total	1876
Accuracy (ViT)	0.0299
Correct (Swin)	77
Total	1876
Accuracy (Swin)	0.0410

Table 2. Baseline performance of ViT and Swin Transformer.

5.2. Fine-Tuning Results

Fine-tuning the models with different learning rates significantly improved their performance. This section details the results of the fine-tuning process for both the Vision Transformer (ViT) and the Swin Transformer models.

- For the **ViT** model, three learning rates were tested: 3e-3, 5e-5, and 8e-8. The best accuracy of 0.9254 was achieved with a learning rate of 5e-5. The training and validation loss curves for each learning rate are shown in Figures 3, 4, and 5.

Learning Rate	ViT Accuracy
3e-3	0.917910
5e-5	**0.925373**
8e-8	0.313433

Table 3. ViT Accuracy across different learning rates.

- For the **Swin Transformer** model, the same three learning rates were tested: 3e-3, 5e-5, and 8e-8. The best performance was achieved with a learning rate of 3e-3, resulting in a final test accuracy of 0.903518. The training and validation loss curves for each learning rate are shown in Figures 6, 7, and 8.

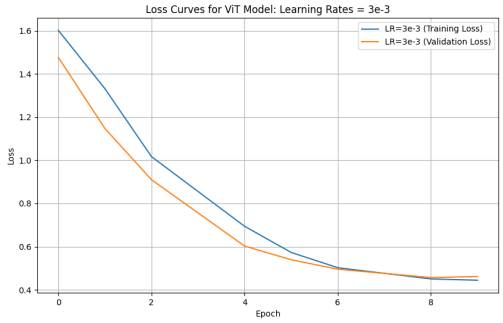

Figure 4. Training and validation loss curves for the ViT model with a learning rate of 3e-3.

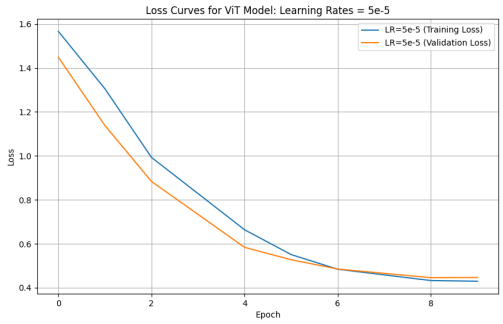

Figure 5. Training and validation loss curves for the ViT model with a learning rate of 5e-5.

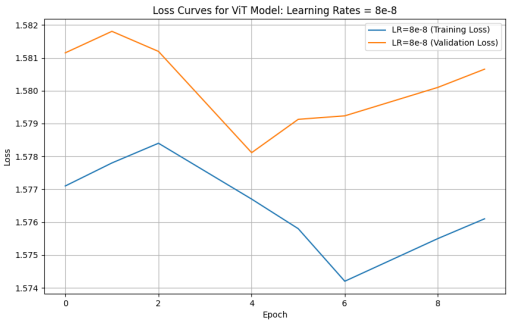

Figure 6. Training and validation loss curves for the ViT model with a learning rate of 8e-8.

Learning Rate	Swin Accuracy
3e-3	**0.903518**
5e-5	0.898721
8e-8	0.194563

Table 4. Swin Transformer Accuracy across different learning rates.

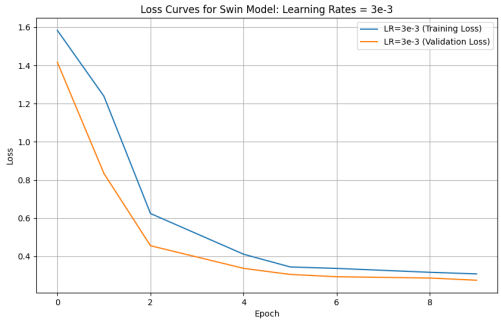

Figure 7. Training and validation loss curves for the Swin Transformer model with a learning rate of 3e-3.

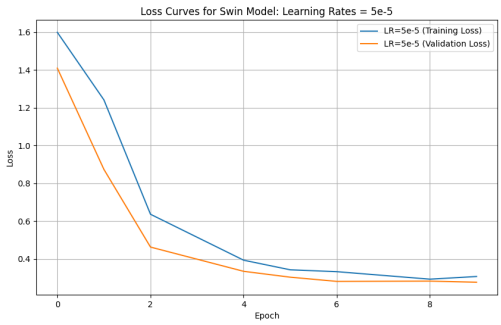

Figure 8. Training and validation loss curves for the Swin Transformer model with a learning rate of 5e-5.

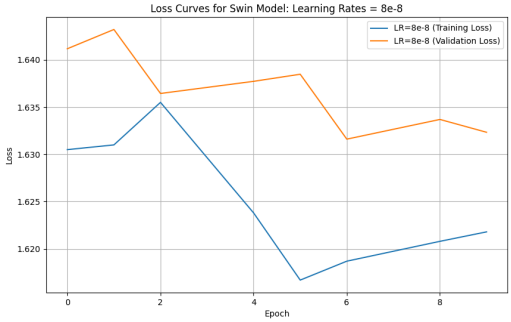

Figure 9. Training and validation loss curves for the Swin Transformer model with a learning rate of 8e-8.

5.3. Discussion

The experimental results demonstrate the importance of selecting an appropriate learning rate when fine-tuning vision transformer models for specific tasks. A learning rate that is too high (e.g., 3e-3) can lead to suboptimal performance, as the model may overshoot the optimal solution and struggle to converge. Conversely, a learning rate that is too low (e.g., 8e-8) can prevent the model from learning effectively, resulting in poor performance.

The ViT model generally outperformed the Swin Transformer in terms of test accuracy across all learning rates. This suggests that the global self-attention mechanism employed by ViT might be more suitable for the ship classification task compared to the local attention used in the Swin Transformer. The ViT model's ability to capture long-range dependencies and global context may be beneficial for distinguishing between different ship categories based on their overall appearance and structure.

It is important to note that the performance difference between the two models is relatively small, and further experiments with different model configurations and hyperparameters could potentially lead to improved results for the Swin Transformer. The Swin Transformer's hierarchical structure and shifted window attention mechanism have shown promising results in various computer vision tasks, and fine-tuning the model with a more extensive hyperparameter search may yield better performance.

6. Conclusion and Future Work

This study demonstrates the effectiveness of fine-tuning vision transformer models for ship classification. The ViT model achieved a higher accuracy compared to the Swin Transformer, with a test accuracy of 92.54% versus 90.35%, when fine-tuned with a learning rate of 5e-5 and 3e-3, respectively. These results underscore the importance of selecting appropriate hyperparameters for effective fine-tuning.

To obtain a more robust estimate of the models' performance and reduce the impact of random variations in the data split, exploring the use of cross-validation techniques is recommended. Additionally, incorporating data augmentation techniques during training could improve the models' performance and generalization. Further investigation into different variations of the vision transformer models, such as larger model sizes or increased transformer layers, may yield additional performance gains.

Another promising direction for enhancing ship classification models is the exploration of domain-specific pre-training. Training the models on a large-scale dataset of maritime images before fine-tuning on the target dataset may help the models learn more relevant and transferable features. Additionally, integrating vision transformer models with other modalities, such as radar or AIS data, could enhance ship classification performance by combining multiple sources of information.

7. Contributions & Acknowledgements

7.1. Contributions

Jean Laguerre:

- Performed data cleaning and helped with data prepro-

cessing.

- Conducted experiments, including the setup and fine-tuning of the Vision Transformer (ViT) and Swin Transformer models.
- Implemented the code modifications.
- Helped analyze experimental results.

James Park:

- Acquired the dataset and did data preprocessing.
- Designed the experiments and helped implement them.
- Analyzed the experimental results.
- Wrote the first draft of the paper and revised it to ensure clarity and coherence.

7.2. Acknowledgements

We would like to acknowledge Analytics Vidhya for the ship dataset used in this study. We also thank the open-source community for developing and maintaining the Hugging Face Transformers library [10], which facilitated the implementation of the vision transformer models.

Additionally, we referred to the codebases from the following GitHub repositories:

- Vision Transformer (ViT) from Google Research: [15]
- Swin Transformer from Microsoft Research: [16]

References

[1] A. D. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[2] A. G. H. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[3] A. P. et al. Pytorch: An imperative style, high-performance deep learning library, 2019. Accessed: 2024-06-17.

[4] A. V. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[5] B. S. et al. Generic and attribute-specific deep representations for maritime vessels. *IPSJ Transactions on Computer Vision and Applications*, 9:1–18, 2017.

[6] C. S. et al. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[7] J. Y. et al. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[8] K. H. et al. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Q. L. et al. Datasets: A community library for natural language processing, 2021. Accessed: 2024-06-17.

[10] T. W. et al. Transformers: State-of-the-art natural language processing, 2020. Accessed: 2024-06-17.

[11] Z. L. et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[12] S. Fayou, H. C. Ngo, and Y. W. Sek. Combining multi-feature regions for fine-grained image recognition. *I.J. Image, Graphics and Signal Processing*, 14(1):15–25, 2022.

[13] M. Milicevic, K. Zubrinic, I. Obradovic, and T. Sjekavica. Data augmentation and transfer learning for limited dataset ship classification. *WSEAS Transactions on Systems and Control*, 13:460–465, 2018.

[14] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[15] G. Research. Vision transformer (vit). `https://github.com/google-research/vision_transformer`, 2021. Accessed: 2024-06-17.

[16] M. Research. Swin transformer. `https://github.com/microsoft/Swin-Transformer`, 2021. Accessed: 2024-06-17.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015.