A Multi-Scale Vision Transformer

Jinzhi Yang[1], Sophie Li[1], Albon Wu[1], Jiarui Wan[1], Shiyu Fu[1], Hanzhi Bian[1]

[1]University of Michigan
{jinzhiy, sophieli, albonwu, jrwan, shiyufu, irisbian}@umich.edu

Abstract

We present a novel architecture, the Multi-Scale Vision Transformer (MSViT), that improves the adaptability and performance of Vision Transformers (ViTs) using multi-scale processing capabilities. Our contribution principally addresses the dynamic input resolution problem—the dramatic loss of performance experienced by classical ViTs on out-of-distribution input image resolutions. We achieve this through two primary modifications: (i) the introduction of dynamic patch sizes, and (ii) the integration of Convolutional Neural Networks (CNNs) for enhanced local feature extraction with Spatial Pyramid Pooling (SPP) for robust multi-scale feature representation. These changes introduce desirable properties of CNNs (e.g., spatial scale and distortion invariance) while retaining the merits of Transformers (e.g., long-range attention, global context modeling, and scalability).

The key insight of our contribution is that by introducing CNNs for local feature extraction and SPP for multi-scale feature aggregation to the base ViT architecture, MSViT eliminates ViT's dependency on fixed positional encodings, which are often degraded by variations in image resolution. We account for this theoretically and empirically; our method dynamically adjusts patch sizes while preserving positional consistency, avoiding the distortions introduced by classical ViT interpolation. We corroborate this phenomenon with extensive experimentation on the Modified National Institute of Standards and Technology (MNIST) dataset, training and testing on images of varying scales. Our results indicate significant improvements accuracy and scalability, with MSViT outperforming standard ViTs on scaled images, including resolutions 8x larger than the base MNIST image dimensions. Also, MSViT avoids the runtime bottleneck from interpolation and is more computational efficient than ViT with interpolation at inference time. The hybrid MSViT architecture improves generalization across variable input scales, making it especially promising for real-world CV applications.

Introduction

Vision Transformers (ViTs), which adapt the natural language self-attention mechanism into the image modality [1], have become seminal in computer vision. By modeling long-range dependencies, ViTs have achieved state-of-the-art performance on large-scale datasets such as ImageNet, exceeding traditional convolutional neural networks (CNNs) while minimizing computational cost. Despite their success, however, ViTs rely heavily on fixed positional encodings to capture spatial relationships within images. This reliance is especially problematic when a ViT is presented with images of varying resolutions and spatial scales, since the fixed encodings struggle to generalize to new image sizes. As a result, performance tends to degrade when images deviate from the training data, which can significanlty inhibit the ViT's real-world applicability.

The foundational work by Dosovitskiy et al. (2021) introduced Vision Transformers (ViTs) and demonstrated their superior performance by employing a novel approach of tokenizing images into non-overlapping patches. While effective at the trained image size, the reliance on fixed patch sizes and positional encodings limits the model's adaptability to diverse or higher-resolution images. To address this, Dosovitskiy et al. used interpolation to generalize to larger-scale data encountered during inference, linearly interpolating the learned positional embeddings to generate new ones for larger image sizes. However, this distorts spatial relationships, leading to degraded performance on scaled images. We show this in the Appendix by visualizing the attention heatmap of ViT for images of different scales. This limitation highlights the need for ViT architectures capable of effectively handling multi-scale inputs while maintaining performance.

To address this limitation, we propose a hybrid architecture that enables dynamic patch sizing and seamlessly integrates Convolutional Neural Networks (CNNs) and Spatial Pyramid Pooling (SPP) with the standard ViT model. CNNs enhance local feature extraction and provide spatial invariance, while SPP enables multi-scale feature representation, mitigating issues arising from fixed positional encodings because now our number of patches are the same, so there is no need for any new positional embeddings. Our hypothesis is that this hybrid approach will improve the generalization capabilities ViTs across varying resolutions, making them more applicable in for tasks involving multi-scale and high-resolution imagery.

In this paper, we validate our proposed architecture using the MNIST dataset, which is especially sensitive to spatial distortions due to the inherent nature of handwritten digit recognition. We replicate the baseline ViT model and introduce adaptive patch division and hierarchical embedding to improve generalizability. Having eliminated the standard ViT model's dependency on fixed positional encodings,

our experimental results demonstrate significant improvements in accuracy and scalability on the MNIST dataset. Our work presents a promising hybrid approach to improving the generalizability—and thus real-world applicability–of ViT.

Related Work

Vision Transformer: Foundation and Challenges *Dosovitskiy et al. (2020)* introduced Vision Transformers (ViT), which treat non-overlapping image patches as tokens, similar to words in natural language processing. Despite demonstrating state-of-the-art performance on large datasets, such as JFT-300M, ViT heavily depends on large-scale pre-training and struggles with tasks requiring fine-grained spatial understanding. This foundational work underscores the necessity for architectural improvements, particularly in positional encoding and multi-scale feature extraction. Our proposed hybrid architecture leverages CNNs and spatial pyramid pooling (SPP) to address these issues by enhancing local feature representation and spatial robustness.

Hybrid Vision Models *CvT: Convolutional Vision Transformers [5]* combines the strengths of CNNs and transformers by replacing standard self-attention with convolutional token embeddings, achieving higher efficiency and locality. However, CvT's fixed convolutional design limits its adaptability to scale variance. By incorporating SPP, our approach extends the spatial capabilities of ViTs, providing a dynamic mechanism to aggregate features across varying scales.

Advances in Positional Encoding *LiT: Locality Enhanced Vision Transformers [4]* introduces locality-enhanced attention to address ViT's inability to capture fine-grained features. However, LiT's approach relies heavily on modified attention mechanisms, increasing computational overhead. Our method enhances positional encoding by leveraging CNN and SPP modules, preserving ViT's global context modeling while reducing complexity.

Multi-Scale Feature Learning in Vision Transformers *Swin Transformer V2 [3]* extends hierarchical ViT architectures to handle large-scale datasets by refining window-based attention mechanisms. Despite its scalability, Swin relies on handcrafted hierarchical designs, which may not generalize well to diverse tasks. Our approach leverages SPP for automatic multi-scale feature aggregation, eliminating the need for rigid hierarchical structures.

Contributions While existing works such as CvT, and LiT have advanced hybrid architectures by combining CNNs with transformers, these models often fail to generalize positional encoding across diverse spatial scales. Similarly, positional encoding methods like Swin Transformer V2 focus on only global context.

Our proposed architecture builds on this body of work by:

- **Dynamic multi-scale input encoding:** leveraging CNN and SPP for enhanced robustness to spatial scale variations.
- **Balanced computational efficiency and performance:** achieving generalization for high-resolution and resource-constrained tasks through a hybrid architecture.

By synthesizing the strengths of recent advancements while addressing their limitations, our approach pushes the boundaries of ViT architectures for scalable, fine-grained vision tasks.

Background

Our work builds on the Vision Transformer by utilizing Spatial Pyramid Pooling to process images of different dimensions. In this study, we experiment with the MNIST dataset to evaluate the performance of our model. We use cross-entropy loss in the training process.

Vision Transformer (ViT) The Vision Transformer, introduced by Dosovitskiy et al., is a computer vision application of the transformer architecture for natural language. Classically, ViT partitions an input image into fixed-size patches, analogous to tokens in natural language, from which it computes embeddings that are processed by a standard transformer encoder (pictured). There are numerous advantages to this architecture, including the ability to capture long-range patch relationships and decreased computational cost.

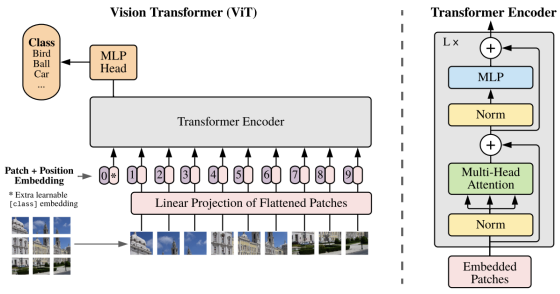


Figure 1: ViT architecture. Reproduced from [1].

Spatial Pyramid Pooling (SPP) Spatial Pyramid Pooling is a mechanism for obtaining fixed-size vector representations of variable-size input images. It divides a feature map into non-overlapping regions of equal size, performs spatial pooling on each region, and concatenates the results. This process is repeated at varying levels of granularity, giving rise to the namesake "pyramid" structure. When the outputs of all levels are concatenated, a fixed-length representation is produced, which may be fed into the fully-connected layers of a CNN.

MNIST The MNIST (Modified National Institute of Standards and Technology) database is a dataset of 28x28 pixel grayscale images of handwritten digits, widely used in computer vision.

Cross-entropy loss Cross-entropy loss measures the deviation of a probability distribution from the target. It is given by

$$\text{Loss} = -\sum_{k=1}^{C} y_k \log(p_k)$$

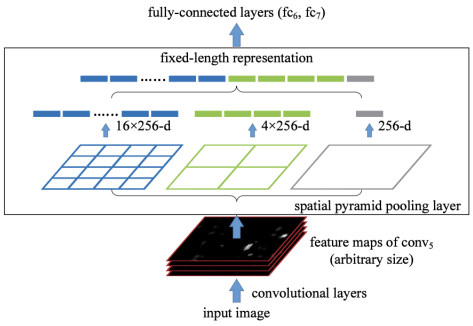


Figure 2: SPP mechanism. Reproduced from [2].

for classes $k \in \{1, \ldots, C\}$, target y_k, and predicted probability p_k.

Methodology, Results, and Discussion

This section describes our model design and training procedures of our replicated Vision Transformer model.

Dataset Selection: For our replication, we selected the MNIST dataset instead of CIFAR-10 or ImageNet, which were used in the original Vision Transformer paper. The primary reason is that MNIST consists of handwritten digit images, which are more sensitive to spatial distortion caused by interpolation. Our tests indicated that images of objects like animals are more robust to spatial distortions, likely because identifying animals is less reliant on precise spatial arrangements. In contrast, classifying handwritten digits requires more accurate positional information, making MNIST a more suitable dataset for evaluating the impact of spatial distortions.

Model design: We adopted the original Vision Transformer architecture with modifications to reduce the number of attention heads and layers. Additionally, we decreased the embedding dimension and MLP hidden size. These changes were made for two main reasons: (i) our dataset, MNIST, is simpler than CIFAR-10 or ImageNet used in the original paper, allowing us to achieve comparable performance with a smaller model; and (ii) a smaller model requires less time and memory for training, making it more suitable given our computational budget.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT (ours)	4	256	512	4	2.2M

Table 1: Comparison of the original Vision Transformer model and our replicated version.

Training procedure: We employed the SGD optimizer with a learning rate of 0.03 and a custom cosine scheduler featuring 500 warm-up steps. The batch size was set to 64, and cross-entropy was used as the loss function. Model evaluation was performed on the validation set every 1,000 steps. We applied an early stopping criterion, halting training if the validation loss did not improve over the past 5 evaluations.

The model was trained for a total of 48,000 steps, achieving a 98.6% accuracy on the MNIST test set. The training curves are shown in Figure 3, and the validation accuracy is illustrated in Figure 4.

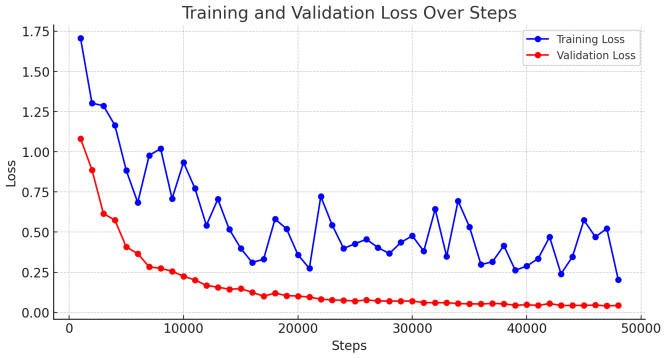


Figure 3: Training curve for training loss and validation loss on replicated ViT

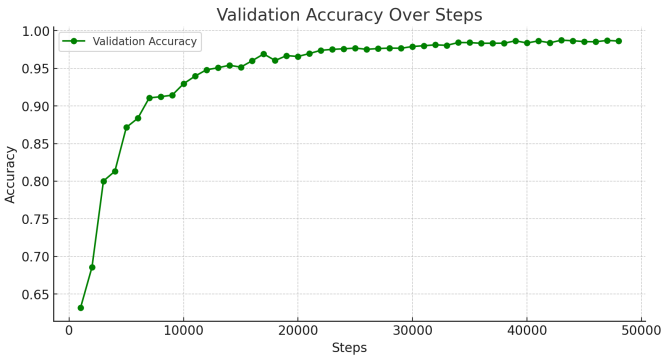


Figure 4: Training curve for validation accuracy on replicated ViT

Extensions

Since the original Vision Transformer requires training data to be fixed sized, this severely limits the model's ability to extrapolate beyond the dimensions that the model has seen during training time. In this section, we explain how we utilize CNN and SPP to allow training data of different sizes. We hypothesize that our new model, MSViT, will achieve significantly better accuracy for higher-resolution images of larger sizes.

Dynamic Patch Size: The original Vision Transformer model uses a fixed patch size. However, when a test image has a larger size or higher resolution than the maximum training image size, it results in unseen positional embeddings due to the creation of additional patches. Suppose the training image size is $N \times N$ and the patch size is $P \times P$. For simplicity, assume P divides N evenly, i.e., $L = N/P \in \mathbb{Z}$. Under this assumption, the image is divided into an $L \times L$ grid, and the model learns L^2 positional embeddings corresponding to these grids.

In our extension, we maintain a fixed grid size of $L \times L$. To achieve this, we introduce a dynamic patch size by first padding the image to an integer multiple k of L. By setting k as the new patch size, we ensure the grid size remains consistent at $L \times L$, regardless of the image's original size or resolution. In this way, the model will not encounter any unseen positional embeddings.

New Tokenization Using CNNs+SPP: In the original Vision Transformer model, fixed-size patches are linearly embedded for tokenization. However, this approach cannot work for dynamic patch sizes, because linear embedding requires all patches are of a fixed size. Inspired by the approach in [**?**], we introduce Spatial Pyramid Pooling (SPP) to address this issue by enabling the extraction of fixed-size embeddings, regardless of the input size.

Another challenge is, when patches are large, linear embedding alone may fail to capture the intricate spatial features within these larger patches. This potentially limits the model's ability to process images effectively. To overcome this, we incorporate one layer of Convolutional Neural Networks (CNNs) to extract robust features before applying SPP.

Therefore, we propose a new tokenization method that combines CNNs and SPP. CNNs can capture local spatial features and are inherently scale-invariance. Following CNN, we use SPP to generate fixed-size embeddings from patches of varying sizes. This ensures a consistent tokenization process that preserves essential spatial and semantic information across different image scales, and enables better generalization to larger scales.

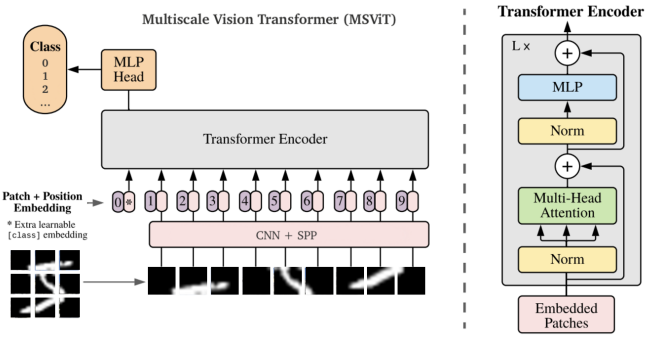


Figure 5: MSViT Architecture. Adapted from [1]

Dataset: To accommodate the multi-scale image capability of our model, we augmented the MNIST dataset by including images scaled by factors of 1, 2, 3, and 4 during the training phase; 1, 2, 4, and 8 during the testing phase. Images are resized with bilinear interpolation, which estimates the value of a pixel based on the four surrounding pixels by calculating the weighted average based on the distance of the target pixel to its nearest neighbors.

Training: The training procedure remained consistent with previous experiments. We trained the model for 40,000 steps, achieving a test accuracy of 98.8%. The training curves are shown in Figure 6, and the validation accuracy is illustrated in Figure 7.

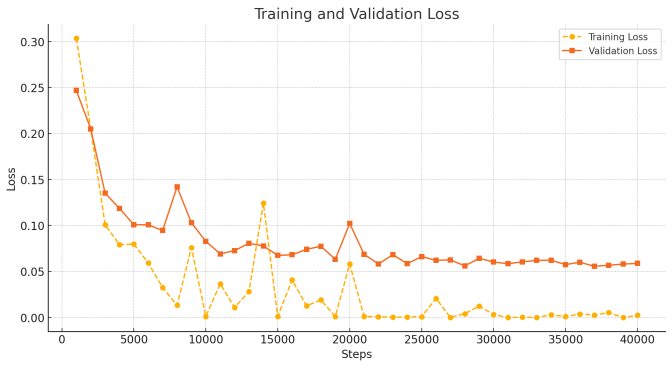


Figure 6: Training curve for training loss and validation loss on our model

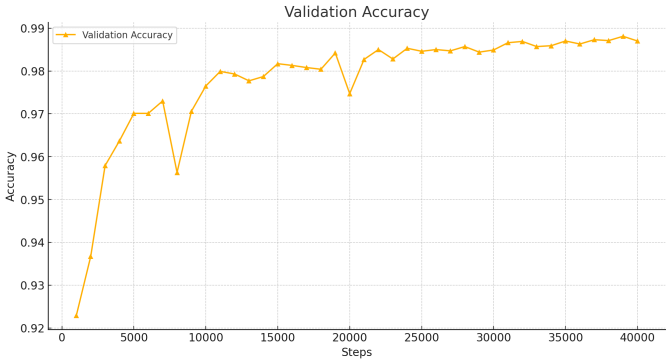


Figure 7: Training curve for validation accuracy on our model

Experiment and Results: We compared our model with the original Vision Transformer model on augmented MNIST test sets, where the image sizes were scaled by factors of 1, 2, 4, and 8, respectively. As shown in Figure 8, the performance of the original Vision Transformer model degrades rapidly as the test image scale increases. In contrast, our model maintains strong performance across all scales. Notably, our model also performed very well on scale of 8, which is unseen during training. This demonstrates that our model can effectively generalize to larger, unseen scales while maintaining good performance.

We also discovered the runtime for interpolation is very slow compared to our model. This indicates our model achieves better computational efficiency on larger scale images.

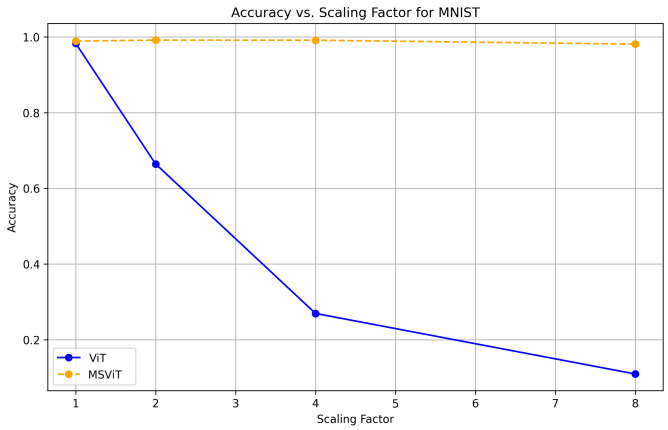


Figure 8: Performance on multi-scale test sets

Conclusions

In this work, we addressed the limitations of ViT in handling multi-scale inputs by integrating CNN and SPP. A key issue with original ViT is their reliance on fixed patch sizes and positional encoding, which are inherently tied to interpolation method which often degrade classification performance by introducing distortions or misalignments in spatial relationships. Through extensive experimentation, our hybrid architecture demonstrated superior performance in maintaining spatial relationships in different image resolutions, achieving a significant improvement in accuracy and scalability on datasets such as MNIST compared to the original ViT model. In addition, our proposed dynamic patch size and new tokenization methods enhanced the generalization ability of ViT, despite having limited resources while tasked with the requirement for high-resolution images. Future work could explore the application of our approach to more extensive datasets like CIFAR-10 or ImageNet, to further validate the effectiveness of our approach.

Societal Impact

The proposed Multi-Scale Vision Transformer (MSViT) has the potential to bring significant societal benefits by enhancing the robustness and scalability of computer vision systems across various application domains. By addressing the limitations of traditional Vision Transformers in handling dynamic input resolutions, MSViT can improve the reliability of vision-based technologies in critical areas such as autonomous vehicles, medical imaging, and disaster response. For instance, its ability to generalize across diverse spatial scales can enhance image recognition systems in medical diagnostics, where high-resolution imaging plays a crucial role in detecting diseases.

References

[1] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

[2] Kaiming He, S. R., Xiangyu Zhang; and Sun, J. 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*.

[3] Liu, Z.; Hu, H.; Lin, Y.; et al. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. *arXiv preprint arXiv:2111.09883*.

[4] Tang, Q.; Liu, Y.; Xiao, J.; et al. 2022. LiT: Locality Enhanced Vision Transformers. *arXiv preprint arXiv:2202.06820*.

[5] Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22–31.

Individual Contributions

Jinzhi Yang: Formulate research question, replicate ViT models, adapt CNN+SPP to ViT model, design experiments and training procedure, analyze results, write paper, review paper, maintain Github repository, manage and supervise project.

Sophie Li: Collect and curate dataset, train both ViT model and our new model, validate and evaluate trained model, maintain Github repository, write paper, review paper.

Albon Wu: Collect and curate dataset, train both ViT model and our new model, validate and evaluate trained model, maintain Github repository, write paper, review paper.

Jiarui Wan: Implement dynamic patch sizing, adapt CNN+SPP to ViT model, write paper, review paper.

Shiyu Fu: Visualize attention map of the model to show interpolation problem, write paper, review paper, design poster.

Hanzhi Bian: Visualize attention map of the model to show interpolation problem, write paper, review paper, design poster.

Appendix A

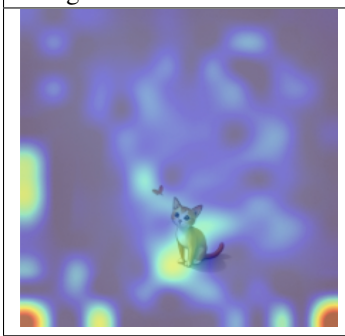
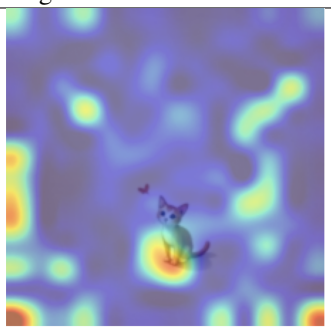
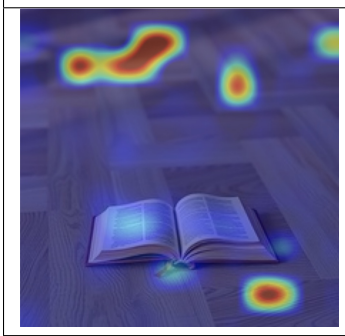
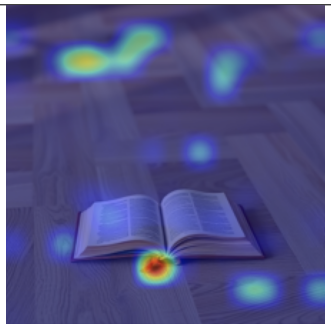
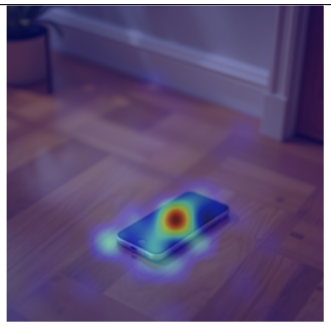
Image resolution of 224x224	Image resolution of 6000x6000
	
	
	

Table 2: Heat Map of Features of Different Image Resolutions

The table of heat map comparison on the features to classify the objects in the images shows the difference of how the model percepts the objects and classify the object. As a result, all images of 224×224 resolution correctly identify the object in the image while the 6000×6000 fails to complete the task. The color on the heat map represents the weights of how the modify use to classify the object. With redder color, the model use more weights on the pixels to classify the object. On the other hand, with darker blue color, the sections would be considered less by the model to classify to object. For 224×224 resolution image, the color covers the whole cat, meaning the model consider the whole cat in order to classify it, while the 6000×6000 resolution image only covers part of the cat and have more focus on the background, which results in misclassification. Similarly, all 6000×6000 images show a distribution shift of the weights compared to 224×224 resolution image. This table shows how the problem of interpolation occurs in the process of classification.

Appendix B

You can find the project repository on GitHub: https://github.com/JinzhiYang03/Multiscale_Vit