INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DESIGN AND MANUFACTURING KURNOOL

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
DEPARTMENT OF CSE

AD352 - Big Data Analytics
Under the guidance of Dr. N Srinivas Naik

Escaping the Big Data Paradigm with Compact Transformers

by Ali Hassani, Steven Walton Date: 7 June 2022

March – April 2024

Shenbagasujan V S – 121AD0011
Madhav Sharma - 121AD0013

Contents

Abstract

Semantic segmentation of large image files, such as .tif, .tiff, or .jpg formats, is a critical task in remote sensing applications, necessitating efficient and accurate solutions. However, the computational demands and cost associated with processing such massive datasets pose significant challenges. In this paper, we present a robust solution that integrates UNet-based architectures with transformer models to achieve high-performance semantic segmentation.

Our approach capitalizes on parallel processing and domain adaptation techniques to enhance segmentation accuracy and efficiency. Specifically, we propose:

1. Parallel Processing: We introduce a novel parallel processing methodology that effectively segments large images by patchifying them and harnessing transformer-based models. This enables efficient utilization of computational resources and accelerates the segmentation process.

2. Domain Adaptation: We devise a domain adaptation strategy to generate masks from input images and reconstruct them to evaluate segmentation accuracy. By adapting to different domains, our approach enhances the model's generalization capability and robustness across diverse datasets.

3. Big Data Techniques: We leverage advanced big data techniques to optimize patchification and bolster scalability. By employing parallel processing frameworks and distributed computing resources, we mitigate computational bottlenecks and expedite the segmentation pipeline.

4. Ensemble Learning: In future endeavors, we advocate for the adoption of ensemble learning techniques, such as Spark-based ensembles, to further enhance segmentation performance. By training multiple models in parallel and combining their predictions, we anticipate achieving superior segmentation results and reducing processing time.

While our approach effectively addresses the challenges posed by large image sizes, it is important to note its limitations. Specifically, our methodology assumes perfect recall and does not explicitly handle scenarios involving incomplete or imperfect information. Moreover, the scalability of our solution may be constrained by available computational resources. Nonetheless, our proposed framework represents a significant step towards efficient and accurate semantic segmentation of large-scale image datasets in remote sensing applications.

Keywords : Semantic segmentation, Large image files, Remote sensing, UNet-based architectures, Transformer models, Parallel processing, Domain adaptation, Segmentation accuracy, Computational efficiency, Big data techniques, Scalability, Ensemble learning, Spark, Patchification, Computational resources, Generalization capability, Robustness, Diverse datasets, Segmentation pipeline, Imperfect information handling

Introduction:

In recent years, the field of deep learning has witnessed remarkable advancements, particularly in the domain of image processing. Convolutional Neural Networks (CNNs) have long been the cornerstone of image classification and segmentation tasks, achieving state-of-the-art results in various benchmarks. However, with the emergence of transformers, originally designed for natural language processing tasks, there has been a paradigm shift in how we approach image understanding.

Transformers, with their ability to capture long-range dependencies and global context, have shown promising results in image classification, object detection, and segmentation. However, their widespread adoption in image processing tasks is hindered by their computational complexity and memory requirements, especially for large-scale datasets. Moreover, existing transformer architectures often rely on extensive parallelism or big data strategies, which may not always be feasible in resource-constrained environments or for smaller datasets.

In this context, our paper proposes a novel methodology for addressing these challenges by leveraging the power of distributed computing through the Spark environment. We introduce a streamlined approach to patchification and preprocessing of input images, tailored specifically for compact transformer architectures. Unlike conventional methods that utilize large-scale parallelism or rely on massive datasets, our approach focuses on maximizing efficiency and scalability while minimizing computational overhead.

The core of our methodology lies in the patchification process, where input images are divided into smaller patches using Spark, enabling efficient parallel processing across multiple nodes. This not only facilitates the handling of large-scale datasets but also ensures optimal resource utilization, thereby accelerating the overall preprocessing pipeline.

Following patchification, the preprocessed images undergo a series of transformation steps aimed at enhancing their suitability for downstream tasks such as image classification or segmentation. These include resizing to a standardized format, conversion to numpy arrays for compatibility with deep learning frameworks, normalization to ensure numerical stability and robustness, and augmentation to augment the dataset and improve generalization.

Moreover, to facilitate model training and evaluation, we employ a train-test-val split strategy, allocating a portion of the data for validation purposes. This ensures that our models are trained on diverse datasets while also providing a reliable mechanism for assessing their performance.

The proposed methodology extends beyond preprocessing and encompasses the design and training of a compact transformer architecture tailored to the specific requirements of image processing tasks. Inspired by both transformer and U-Net architectures, our model incorporates downsampling and upsampling layers to capture both local and global features effectively.

Finally, the efficacy of our approach is evaluated using standard metrics such as categorical cross-entropy and Intersection over Union (IoU) score, providing insights into the model's performance across various image processing tasks.

In summary, this paper presents a comprehensive framework for harnessing the capabilities of compact transformer architectures in image processing, demonstrating the potential of distributed computing environments like Spark to enhance efficiency and scalability. Through extensive experimentation and evaluation, we showcase the effectiveness of our approach and highlight its relevance in real-world applications.

Keywords : Compact transformer architecture, Spark environment, Patchification, Preprocessing, Image resizing, Numpy arrays, Normalization, Data augmentation, Train-test-validation split, Model architecture, Downsampling, Upsampling, Transformer, U-Net, Image classification, Image segmentation, Computational efficiency, Scalability, Model evaluation, Categorical cross-entropy, Intersection over Union (IoU) score

Contributions

1. Parallel Processing Optimization: This paper presents a sophisticated approach to semantic segmentation, primarily focusing on addressing the challenges posed by large image files common in remote sensing applications. A key contribution lies in the introduction of a parallel processing

methodology, meticulously designed to efficiently segment large images. This method intricately combines the patchification of images with the utilization of transformer-based models, ensuring optimal resource utilization and computational efficiency. By breaking down large images into smaller patches and leveraging the parallel processing capabilities of modern computing architectures, our approach significantly accelerates the segmentation process, thereby reducing computational overhead and time complexity.

2. Domain Adaptation Strategy: Another significant contribution of this work is the proposal of a robust domain adaptation strategy tailored specifically for semantic segmentation tasks. By systematically generating masks from input images and subsequently reconstructing them, our methodology effectively adapts to different environmental conditions and diverse datasets. This adaptation mechanism enhances the model's generalization capability and robustness, enabling it to perform effectively across varying domains and real-world scenarios. This contribution is particularly vital in remote sensing applications where the environmental conditions and image characteristics may vary widely.

3. Integration of Big Data Techniques: Furthermore, this paper showcases the integration of advanced big data techniques to optimize the patchification process and enhance scalability. Leveraging parallel processing frameworks and distributed computing resources, our approach addresses the computational challenges associated with processing large-scale image datasets. By efficiently distributing the segmentation workload across multiple computing nodes, we mitigate computational bottlenecks and expedite the segmentation pipeline. This contribution significantly enhances the scalability of our methodology, allowing it to handle massive image datasets with ease.

4. Ensemble Learning Recommendations (Future Work): Additionally, we propose the exploration of ensemble learning techniques, such as Spark-based ensembles, as a direction for future research. While not fully implemented in the current work, the potential of ensemble learning in improving segmentation performance is highlighted. By training multiple segmentation models in parallel and aggregating their predictions, ensemble learning can potentially yield more accurate and robust segmentation results. This recommendation opens avenues for further enhancing the effectiveness and efficiency of semantic segmentation methodologies in remote sensing applications, paving the way for future advancements in the field.

Keywords : Semantic segmentation, Large image files, Remote sensing, UNet-based architectures, Transformer models, Parallel processing, Domain adaptation, Segmentation accuracy, Computational efficiency, Big data techniques, Scalability, Ensemble learning, Spark, Patchification, Computational resources, Generalization capability, Robustness, Diverse datasets, Segmentation pipeline, Imperfect information handling

Literature Survey:

"Attention is All You Need": This groundbreaking paper by Vaswani et al. introduced the Transformer architecture, which revolutionized natural language processing (NLP) tasks. Unlike traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs), Transformers rely solely on self-attention mechanisms to capture global dependencies in sequential data. The architecture comprises multiple self-attention layers and feed-forward networks, enabling parallel processing and efficient modeling of long-range dependencies. This paper paved the way for the widespread adoption of Transformers in various NLP tasks, including machine translation, text generation, and language understanding.

"Image is Worth 16x16 Words": This work extends the Transformer architecture, originally designed for sequential data, to image classification tasks. The authors propose the Vision Transformer (ViT) model, which treats an image as a sequence of fixed-size image patches and processes them using Transformer encoders. By pretraining on large-scale image datasets such as ImageNet, ViT achieves competitive performance compared to traditional convolutional neural networks (CNNs), demonstrating the versatility of the Transformer architecture in handling diverse data modalities.

"Cross-Parallel Transformer: Parallel ViT for Medical Image Segmentation" by Dong Wang: This paper addresses the computational challenges associated with medical image segmentation by proposing a novel parallelization strategy for Vision Transformers (ViTs). The Cross-Parallel Transformer (CPT) leverages cross-parallel processing to accelerate the segmentation process while maintaining high accuracy. By exploiting both spatial and channel parallelism, CPT effectively processes large medical images, demonstrating superior performance compared to conventional CNN-based approaches.

"Unsupervised Domain Adaptation for the Semantic Segmentation of Remote Sensing Images via One-Shot Image-to-Image Translation": This work focuses on domain adaptation techniques for semantic segmentation tasks, particularly in remote sensing applications. The authors propose an unsupervised domain adaptation framework based on one-shot image-to-image translation. By leveraging generative adversarial networks (GANs) and cycle-consistency constraints, the proposed approach adapts segmentation models to new domains without requiring labeled data, leading to significant improvements in segmentation accuracy across diverse environmental conditions.

"Compact Transformers" by Ali Hassan: This paper introduces Compact Transformers, a more computationally efficient variant of the Transformer architecture tailored for vision tasks. By incorporating locality-sensitive hashing and dynamic routing mechanisms, Compact Transformers reduce computational complexity and memory footprint while preserving performance. The authors demonstrate the effectiveness of Compact Transformers in various image classification benchmarks, highlighting their potential for deployment in resource-constrained environments.

"Segmenter: Transformer for Semantic Segmentation" by Robin Strudel, Ricardo Garcia: This paper presents Segmenter, a Transformer-based model specifically designed for semantic segmentation tasks. By leveraging self-attention mechanisms and hierarchical feature representations, Segmenter achieves competitive performance in segmenting complex scenes and objects. The authors provide insights into the architecture of Segmenter and showcase its effectiveness across various segmentation benchmarks.

"Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation" by Hu, Hu Cao, Yueyue Wang: This work proposes Swin-Unet, a variant of the traditional U-Net architecture that replaces convolutional layers with Transformer blocks. By integrating self-attention mechanisms into the U-Net framework, Swin-Unet captures both local and global contextual information, leading to improved segmentation accuracy. The authors demonstrate the superiority of Swin-Unet over conventional U-Net models in medical image segmentation tasks.

Limitations of the Paper:

1. Limited Dataset Diversity : The paper's training strategy relies solely on a single dataset, neglecting the inclusion of more diverse datasets such as Potsdam or other medical imagery datasets for brain tumor detection. This omission restricts the model's ability to generalize across varied scenarios, thus limiting its applicability in real-world settings where diverse data sources are prevalent.

2. Relevance of Computational Resources: While the paper highlights the efficiency of transformers in utilizing minimal computational resources, it fails to address scenarios where ample computational resources are available. In contexts where computational resources are not a constraint, the proposed approach may be considered less relevant, as the focus shifts from resource efficiency to maximizing performance.

3. Lack of Discussion on Domain Adaptation: The paper overlooks the critical aspect of domain adaptation in semantic segmentation tasks. By solely relying on classification approaches and training on specific datasets, the model's adaptability to new domains and variations in image characteristics is compromised. Semantic segmentation inherently demands robust domain adaptation techniques to ensure reliable performance across diverse environments, a facet not adequately addressed in the paper.

Proposed Extensions:

In light of these limitations, our project aims to extend the methodology proposed in the paper by addressing the following:

Incorporating Diverse Datasets: We plan to augment the training dataset with diverse datasets such as Potsdam or other medical imagery datasets for brain tumor detection, thereby enhancing the model's generalization capabilities across varied scenarios.

Exploring Domain Adaptation Techniques : Our project will integrate domain adaptation techniques to enable the model to adapt to new domains and variations in image characteristics. This involves training the model on labeled images and masks, followed by leveraging large amounts of unlabeled data for further refinement.

Enhancing Efficiency through Parallel Patchification: We propose to optimize the patchification process, especially in remote sensing applications, by leveraging parallel processing techniques. This will involve patchifying the data using a parallel approach to improve efficiency and scalability.

By addressing these limitations and incorporating the proposed extensions, our project aims to enhance the robustness and applicability of the proposed methodology in semantic segmentation tasks across diverse domains.

Proposed Methodology

To parallelize image data processing using Apache Spark, first convert images into NumPy arrays and parallelize them into Resilient Distributed Datasets (RDDs). Define a function to split images into smaller patches, considering factors like patch size and overlap. Apply the patch splitting function using Spark's flatMap() transformation to distribute the processing across multiple nodes in the cluster. Collect the resulting patches from all partitions into the driver node for further analysis or processing. Iterate over the collected patches to extract relevant information such as patch coordinates, labels, or features. Utilize Spark's distributed computing capabilities to efficiently handle large-scale image datasets, optimizing processing time and resource utilization. Ensure proper partitioning and resource allocation to maximize parallelism and scalability. Finally, output the patch information for downstream tasks such as machine learning model training or image analysis.

In the Oxford Pets dataset, each image is accompanied by its corresponding mask, representing the ground truth for segmentation tasks. To prepare the dataset for model training, the first step is to normalize the images and resize them to a standard size. This ensures consistency in input dimensions across the dataset. To preserve class labels, a K-Nearest Neighbors (KNN) approach can be employed, matching each image with its corresponding mask based on similarity.

After normalization and label preservation, introducing image augmentation techniques like rotation and flipping enhances the model's robustness and reduces overfitting. By applying random rotations and flips to the images, the model learns to generalize better and becomes less sensitive to outliers or variations in input data. These augmented images, along with their corresponding masks, are then incorporated into the dataset.

Overall, the dataset generation process involves normalizing and resizing images, preserving class labels using KNN, introducing image augmentation for robustness, and pairing each image with its corresponding mask to facilitate supervised learning for segmentation tasks. This comprehensive approach ensures the dataset is well-prepared for training robust and accurate models.

Incorporating domain knowledge into the model architecture, we adopt a transfer learning approach by importing the MobileNetV2 model as the backbone. MobileNetV2 provides a solid foundation for feature extraction due to its efficiency and effectiveness. We then integrate UNet architecture, renowned for semantic segmentation tasks, with Transformer layers serving as upsampling modules.

The model workflow entails downsampling the input samples to capture high-level features and spatial information effectively. Subsequently, upsampling is performed using Transformer layers to recover spatial details and generate masks. This fusion of MobileNetV2, UNet, and Transformer layers leverages both the power of pre-trained feature extraction and the capability of UNet to preserve spatial information.

UNet's popularity in the machine learning domain stems from its ability to handle various segmentation tasks efficiently. By incorporating it into our model alongside Transformer layers, we create a robust architecture capable of capturing intricate details while maintaining computational efficiency.

This hybrid model architecture not only benefits from transfer learning to understand the domain but also combines the strengths of UNet and Transformer layers for effective segmentation tasks. Through this integration, we aim to achieve accurate and detailed segmentation results suitable for diverse applications in image analysis and computer vision.

Conclusion and Future Work

Utilize a high-performance GPU for faster processing. Employ efficient data loading using libraries like tifffile. Implement patchification to divide TIFF images into smaller patches for GPU processing. Apply data augmentation techniques to increase sample diversity. Utilize parallel processing and batch processing on GPU for efficiency. Optimize memory usage and model architecture. Use a custom dataset class to load TIFF image patches. Train the model using DataLoader with appropriate batch size. Use appropriate loss function and optimizer for training. Finally, perform inference on patches similarly to training.

Implement domain adaptation by utilizing techniques like adversarial training or self-supervised learning to reduce the need for manual labeling. Fine-tune the model on a source domain with labeled data, then adapt it to the target domain with unlabeled data. Use methods like domain adversarial training to align feature distributions between domains or self-supervised learning to learn representations from unlabeled data. Employ domain adaptation strategies within the training loop to enhance model generalization across domains. Fine-tune hyperparameters and adjust architecture for optimal performance. Evaluate the adapted model's performance on target domain data to ensure effective transfer learning and label reduction.

Training the model on larger remote sensing datasets involves leveraging techniques like transfer learning and data augmentation. Begin by pre-training the model on a large dataset, such as ImageNet, to learn general features, then fine-tune it on remote sensing data to adapt to domain-specific features. Utilize data augmentation methods like rotation, flipping, and scaling to increase dataset diversity and improve model robustness. Experiment with advanced augmentation techniques tailored to remote sensing data, such as speckle noise simulation or geometric transformations to mimic real-world variations.

For implementing the frontend of the model, prioritize user-friendly design and intuitive interaction. Develop a graphical user interface (GUI) with features for data input, model selection, parameter tuning, and result visualization. Utilize frameworks like Flask or Django for backend development, ensuring scalability and compatibility with various platforms.

To explore additional metrics like dice loss, collaborate with PhD experts to identify relevant evaluation criteria for remote sensing tasks. Integrate dice loss calculation into the model training pipeline to assess segmentation accuracy, complementing traditional metrics like accuracy and IoU. Experiment with different loss functions and evaluation metrics to gain deeper insights into model performance and refine the training process iteratively. Evaluate the model's effectiveness using cross-validation and validation on unseen datasets to ensure robustness and generalization capability.

References

1. Wang, D. (Year). Cross-Parallel Transformer: Parallel ViT for Medical Image Segmentation, v1. *Sensors,* Volume 23(Issue 23), Article Number 9488. [DOI: 10.3390/s23239488](https://doi.org/10.3390/s23239488)

2. Ismael, S. F., Kayabol, K., & Aptoula, E. (2022). Unsupervised Domain Adaptation for the Semantic Segmentation of Remote Sensing Images via One-Shot Image-to-Image Translation. *IEEE.* [arXiv:2212.03826](https://arxiv.org/pdf/2212.03826.pdf)

3. Hassan, A. (Year). Compact Transformers. *Computer Vision and Pattern Recognition.* [arXiv:2104.05704](https://arxiv.org/abs/2104.05704)

4. TensorFlow Documentation. Retrieved from [TensorFlow website](https://www.tensorflow.org/)

5. Kaggle Datasets. Retrieved from [Kaggle website](https://www.kaggle.com/datasets)

6. Strudel, R., & Garcia, R. (2021). Segmenter: Transformer for Semantic Segmentation. *Computer Vision and Pattern Recognition,* Volume 3. [arXiv:2105.05633](https://doi.org/10.48550/arXiv.2105.05633)

7. Hu, C., Cao, H., & Wang, Y. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *Image and Video Processing.* [arXiv:2105.05537](https://doi.org/10.48550/arXiv.2105.05537)

8. Huang, X., & Yang, X. (Year). Test-time bi-directional adaptation between image and model for robust segmentation. *IEEE.* [ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0169260723001438)