Final Report: Cross-Architectural Knowledge Distillation in Medical Imaging: Multi-Scale Geometric Feature Fusion for Point Cloud Object Detection

Mahiyat Nawar Mantaqa
2122455042

MD. Shafat Islam Khan
2121517642

Abstract—**3D medical imaging object detection in point cloud data poses challenges such as high computational demands, sensitivity to noise, and dependency on data quality. We propose a cross-architectural knowledge distillation framework combining a hybrid teacher model—integrating VGG16 for spatial features and Vision Transformer (ViT) for global context—and a lightweight student model based on EfficientNet. The approach addresses issues like noisy and irregular data by leveraging complementary feature extraction and knowledge transfer, while enhancing computational efficiency for deployment in constrained environments. Results on the BraTS 2019 dataset validate the effectiveness of the framework, demonstrating improved accuracy and reduced model complexity.**

I. Introduction

Deep learning has significantly impacted the field of medical imaging, enabling advancements in diagnosis, treatment planning, and patient monitoring. However, these achievements often rely on large, complex models with substantial computational demands. While high-performing models like VGG16 and Vision Transformers (ViT) excel at feature extraction and classification tasks, their use in real-world scenarios, particularly in resource-constrained environments, is challenging.

Medical imaging, especially for 3D point cloud data like those in the BraTS 2019 dataset, presents unique difficulties. Variability in imaging resolutions, irregular anatomical structures, and noise in medical scans demand models capable of robust generalization. Furthermore, dense 3D data often leads to computational inefficiencies. To address these challenges, we explore **Cross-Architectural Knowledge Distillation (KD)**, a method that compresses knowledge from a complex teacher model into a smaller, faster, and equally effective student model.

In our work, we introduce a hybrid teacher model combining **VGG16** and **ViT**, leveraging the strengths of convolutional networks for spatial features and transformers for contextual understanding. This teacher-student setup employs EfficientNet as the student model, chosen for its scalability and computational efficiency. By distilling knowledge, we ensure that the lightweight student model inherits the robust capabilities of the hybrid teacher model, striking a balance between performance and resource utilization.

II. Literature Review

Yufan Liu et al. introduces a novel method [1] for distilling knowledge from Transformer models to Convolutional Neural Networks (CNNs). The authors propose two projectors, the Partially Cross Attention (PCA) projector and the Group-wise Linear (GL) projector, to align student features with teacher features in two projected feature spaces. Additionally, a multi-view robust training scheme is introduced to enhance the robustness and stability of the framework. Extensive experiments on datasets like ImageNet and CIFAR demonstrate that the proposed method outperforms 14 state-of-the-art methods, showcasing its effectiveness in cross-architecture knowledge distillation.

Semantic segmentation of remote sensing images aims [2] to classify each pixel, which is crucial for urban planning and dynamic monitoring. However, existing models struggle with small target pixels and tiny target sizes, leading to poor recognition and segmentation. Additionally, deeper feature extraction modules result in redundant parameters and increased computation time. To address these issues, the authors propose KDMSANet, a lightweight semantic segmentation network that uses knowledge distillation, a multiscale pyramidal pooling module, and an attention mechanism to enhance feature fusion and focus. They trained teacher-student models to create lightweight network models through model pruning and distillation. Experiments on the Vaihingen and Potsdam datasets showed that the proposed network significantly reduces the number of parameters while maintaining accuracy, with the student model's size reduced by 43.6%, training efficiency improved by 22.3%, and accuracy reaching 99.30% of the teacher model.

The document [3] "Feature Extraction from Point Clouds" by Gumhold, Wang, and MacLeod presents a method for detecting feature lines in point clouds without surface reconstruction. It involves assigning penalty weights to points and edges in a neighbor graph to identify feature patterns, followed by recovering feature lines and junctions using wedge and corner fitting. Key steps include graph construction using Delaunay filtering, density estimation, and point classification into surface, crease, or border categories, with penalty functions evaluating the likelihood of feature points. A modified minimum spanning tree extracts feature patterns, while least squares fitting projects noisy data onto feature lines. The approach is robust against noise and has applications in surface meshing, point cloud enhancement, and non-photorealistic rendering, with future work focusing on integrating detection and recovery stages and addressing isolated peaks.

[4] The 3D point cloud (3DPC) has advanced with deep learning (DL), but DL faces challenges like data scarcity and high computational needs. Deep transfer learning (DTL) reduces costs by using knowledge from one task to train another and is effective for aligning point clouds. Domain adaptation (DA), a subset of DTL, improves data quality by addressing noise and missing points. This paper reviews techniques for understanding 3DPC using DTL and DA, covering applications like object detection and segmentation, and discusses the pros and cons of these frameworks, open challenges, and future research directions.

III. Methodology

1. Teacher Model: Hybrid VGG16-ViT

- **VGG16**: Extracts local spatial features through hierarchical convolution layers.
- **ViT**: Captures global context using self-attention.
- **Feature Fusion**: Outputs are concatenated to combine spatial and global features, enhancing representational capacity.

Hybrid Teacher Model (VGG16 + ViT):

The hybrid model leverages VGG16's convolutional layers to extract high-resolution, localized spatial features critical for handling small or irregular structures in medical images. ViT complements this by modeling global context and long-range dependencies, making the system robust to varying resolutions and irregularities.

Impact: The complementary strengths of CNNs and Transformers enable effective feature representation across diverse imaging conditions, reducing dependency on pristine data quality.

2. Student Model: EfficientNet

EfficientNet is designed for scalability and efficiency, making it ideal for resource-constrained applications such as medical imaging. Key reasons for its selection include:

Compound Scaling: EfficientNet balances depth, width, and resolution systematically, achieving better accuracy with fewer parameters.

Lightweight Architecture: Despite being compact, EfficientNet maintains high accuracy, making it suitable for devices with limited computational power.

State-of-the-Art Results: It consistently outperforms other lightweight models in benchmark datasets, proving its robustness in diverse scenarios.

In the context of our study, EfficientNet enables deployment in medical settings where computational resources are limited, such as edge devices or clinical machines.

3. Knowledge Distillation

Loss Function:

$LKD = \alpha \cdot T2 \cdot KL(pteacher, pstudent) + (1 - \alpha) \cdot LCE$ where LCE is cross-entropy loss, KL is Kullback-Leibler divergence, and T is the temperature parameter

- **Training**:
  - Hybrid model trained on BraTS 2019 dataset.
  - Knowledge distillation applied to EfficientNet with weighted loss terms.

IV. Framework

Data Preprocessing: BraTS 2019 dataset preprocessed into voxelized 3D representations.. **Hybrid Model Training**: Training on a full dataset with augmentation. **Student Training**: Knowledge transfer using teacher predictions with temperature scaling.

V. Knowledge Distillation Process

Knowledge Distillation (KD) is a model compression technique wherein a smaller student model learns to replicate the performance of a larger teacher model by mimicking its outputs. The process involves three key steps:

A. Soft Target Generation

The hybrid teacher model generates predictions for the training dataset. These predictions, softened using temperature scaling, provide richer information than hard labels alone.

Temperature scaling is controlled by a parameter T, which softens the probability distribution of the teacher model's

$psfot(yi) = \frac{exp(zi/T)}{\sum jexp(zj/T)}$ where zi is the logit for class *i*, and T¿1.

B. Student Model Training

The student model is trained using a combined loss function: $LKD = \alpha T2 \cdot KL(pteacher, pstudent) + (1 - \alpha) \cdot LCELKD = \alpha \cdot T2 \cdot KL(pteacher, pstudent) + (1 - \alpha) \cdot LCE$ **KL Divergence** (KL) ensures the student mimics the softened outputs of the teacher.

Cross-Entropy Loss (LCE) ensures the student matches the true labels.

The hyperparameter α balances the contributions of the two loss components.

C. Optimization and Iterative Learning

The student model iteratively learns by minimizing *LKD(knowledge distillation loss)* over multiple epochs.

In our implementation:

- The hybrid teacher generates probabilistic outputs for the BraTS 2019 dataset.
- EfficientNet learns both from the teacher's softened logits and the ground-truth labels, resulting in a model that generalizes well while being computationally efficient.

VI. Results

A. Metrics

Teacher Model:

- Training Accuracy: 92.21%
- Validation Accuracy: 94.67%
- Model Size: Hybrid Teacher Model (VGG16 + ViT)

Student Model (EfficientNet):

- Training Accuracy: 90.32%
- Validation Accuracy: 92.17%
- Reduction in Model Size: Significant reduction compared to the teacher model, enabling deployment on resource-constrained systems.

B. Performance

The EfficientNet student model achieved near-parity in accuracy with the hybrid teacher model at a fraction of the computational cost. By leveraging knowledge distillation, the student model efficiently captured the critical features learned by the teacher while optimizing resource usage.

C. Analysis

Teacher Model Performance: The hybrid teacher model demonstrated high accuracy during training and validation, underscoring the effectiveness of combining VGG16 and ViT for feature extraction. However, the large size and computational demands of the hybrid model make it less suitable for real-world deployment.

Student Model Efficiency: The EfficientNet student model successfully distilled the knowledge from the hybrid teacher, achieving comparable accuracy with significantly reduced computational requirements. This validates the efficacy of cross-architectural knowledge distillation in maintaining performance while optimizing resource usage.

D. Key Observations

1) **Training Stability**: Both teacher and student models exhibited stable convergence, with loss decreasing steadily across epochs.
2) **Improved Generalization**: The EfficientNet student achieved higher validation accuracy compared to its training accuracy, indicating better generalization on unseen data.
3) **Reduced Computational Overhead**: The student model's lightweight architecture ensures it is well-suited for deployment on devices with limited computational resources, without sacrificing performance.

These results highlight the potential of our framework to address the computational and accuracy challenges inherent in 3D medical imaging object detection, paving the way for more efficient and accessible solutions in healthcare.

VII. Future Implementation: Multi-Scale Geometric Feature Fusion (MSGFF)

A. Planned Approach for MSGFF Implementation

1) **Objective of MSGFF**: Multi-Scale Geometric Feature Fusion (MSGFF) aims to improve feature representation in 3D point cloud data by effectively handling features at multiple scales. This is particularly crucial in medical imaging, where objects of interest (e.g., tumors or abnormalities) can vary greatly in size, shape, and resolution.
2) **Key Components of MSGFF**:

- **Multi-Scale Feature Extraction**: We will integrate modules to extract features at varying resolutions, capturing both fine-grained local details and coarse global structures. This can be achieved through a combination of:
- **3D Convolutions** for local spatial features.
- **Multi-head Attention** for capturing relationships across different scales.
- **Feature Fusion**: The extracted features from multiple scales will be concatenated or aggregated using techniques like summation, attention-based weighting, or learnable fusion layers.
- **Adaptive Scaling**: Dynamic scaling mechanisms will be implemented to prioritize features most relevant to the target task

B. Integration with Existing Framework

The MSGFF module will be integrated into the hybrid teacher model (VGG16 + ViT). Features extracted from VGG16's convolutional layers and ViT's attention layers will undergo multi-scale processing.

During knowledge distillation, the MSGFF-enhanced teacher model will transfer its multi-scale knowledge to the student model, enabling the lightweight EfficientNet to benefit from the fused features.

VIII. Conclusion

This study presents a cross-architectural knowledge distillation framework for medical imaging, combining the complementary strengths of CNNs and transformers. By distilling knowledge into a lightweight student model, the approach reduces computational demands while maintaining high accuracy, addressing challenges like noisy data and computational overhead. Future work will integrate MSGFF to further enhance 3D point cloud data processing.

References

[1] Y. Liu, J. Cao, B. Li, W. Hu, J. Ding, L. Li, *Cross-Architecture Knowledge Distillation* [Online]. Available: https://arxiv.org/abs/2207.05273

[2] Y. Yang, Y. Wang, J. Dong, B. Yu. ”Diabetic retinopathy: A growing cause of blindness in Bangladesh,” in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, pp. 1-14.

[3] S. Gumhold, X. Wang, R. MacLeod, *Feature Extraction From Point Clouds* [Online]. Available: https://graphics.stanford.edu/courses/cs164-10-spring/Handouts/papers_gumhold.pdf

[4] S. S. Sohail , Y. Himeur, H. Kheddar , A. Amira, F. Fadli, S. Atalla, A. Copiaco, W. Mansoor, *Advancing 3D Point Cloud Understanding through Deep Transfer Learning: A Comprehensive Survey* [Online]. Available: https://arxiv.org/pdf/2407.17877v1

Date of Submission: 25/11/2024 CSE499A Section 23