Applications of Transfer Learning for MRI Segmentation in GI Tract

Flora Diep
flora.diep@queensu.ca
School of Computing
Queen's University
Kingston, Ontario, Canada

Tyler Elliott
tyler.elliott@queensu.ca
School of Computing
Queen's University
Kingston, Ontario, Canada

E Ching (Noon) Kho
noon.kho@queensu.ca
School of Computing
Queen's University
Kingston, Ontario, Canada

ABSTRACT

Radiation therapy (radiotherapy) is a medical treatment that uses high-energy radiation to target and destroy cancerous or abnormal cells within the human body. One challenge that healthcare professionals have with using radiotherapy is how to eradicate tumors while avoiding concurrent harm to the surrounding healthy tissues and organs. As a result, medical imaging is used in radiotherapy to assist with achieving precision and effectiveness in targeting and treating cancer. However, the volume of acquired medical image data is growing far faster than the capacity of available researchers to analyze it due to the global development of medical imaging and the progress of imaging technology. Therefore, deep learning (DL) based auto segmentation methods are needed to support healthcare providers in achieving timely, precise, and accurate imaging-based diagnosis.

KEYWORDS

Deep Learning, Machine Learning, Medical Imaging, Segmentation, Transfer Learning, MRI

ACM Reference Format:
Flora Diep, Tyler Elliott, and E Ching (Noon) Kho. 2023. Applications of Transfer Learning for MRI Segmentation in GI Tract. In *Proceedings of Queen's Conference (Queen's University '23).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

The main research problem in this project is to implement a technique for effectively segmenting the stomach and intestines (small bowel and large bowel) using multiple MRI image scans. In the Kaggle competition for UW-Madison GI Tract Image Segmentation, it was reported that approximately five million people worldwide were diagnosed with gastrointestinal tract (GI) cancers in 2019 [1]. Nearly two million five hundred thousand patients are eligible for radiotherapy, where the treatment is administered for ten to fifteen minutes a day for one to six weeks [1]. Radiation oncologists aim to deliver precise, high doses of radiation using X-ray beams to target tumors without causing damage to healthy tissues and organs.

Advanced technologies like MR-Linancs have allowed oncologists to visualize the tumor and organ positions in real-time. However, a challenge with MR-Linacs is that it is time-consuming for both the oncologist and the patient as it manually outlines each of the organs (stomach and intestines) [2]. By automating the segmentation process through DL, our objective is to improve the time efficiency of treatment in radiotherapy and to reduce anatomical errors.

The significance of this project is to consider the global impact of GI tract cancers. Developing a technique that can segment the organs (stomach and intestines) from MRI scans will improve patient care; treatment plans for cancer will become more effective and time-efficient.

Within this research Kaggle competition, participants were tasked with developing models that can automatically segment the organs (stomach and intestines) in the MRI scans of cancer patients who have been receiving radiation treatment.

Transfer Learning (TL) involves using pre-existing knowledge from pre-trained DL models and applying/adapting that knowledge to a new task [3]. By using TL, new models can be developed that can identify and delineate the organs (stomach and intestines) in the MRI images, thereby creating an effective segmentation algorithm for imaging. Each dataset has one to five MRI scans of different organ positions on multiple days of therapy.

A pre-trained image classification is a model that has undergone training to recognize and categorize patterns, features, or objects within images [3]. For image segmentation of MRI scans, these models will help with achieving our goal of creating faster and more segmentation by fine-tuning the model based on the knowledge that is built from the pre-trained model.

The advantage of using TL is that it helps to accelerate the creation of new models for new tasks. By using the existing knowledge that is embedded in pre-trained models, we can speed up the creation of algorithms capable of accurate and efficient segmentation, which is crucial for improving patient care and developing treatment plans.

Supervised Machine Learning takes training models on labeled datasets [4]. Each data point from the dataset is associated with the correct target. For our project, this is beneficial in teaching and building precise segmentation models of how to identify and outline the organs (stomach and intestines) from MRI images that have been annotated.

Encoders are neural networks that take a variable-length sequence as an input and transform the input data into a different representation. This process of encoding involves capturing the

Queen's University '23, September–December 2023, Kingston, ON

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

important features or patterns from the input data. Decoders transform the abstract representation (obtained from the encoder) back into its original data format. Decoders play a vital role in retrieving the essential information captured by the encoder, which helps to ensure that the reconstructed data retains meaningful features.

U-Net is widely used for semantic segmentation tasks, where the goal is to classify each pixel in an image into pre-defined classes. The U-Net architecture consists of a contracting path (encoder) on the left side and an expansive path (decoder) on the right side, with a bottleneck in the middle. In the context of organ segmentation in medical imaging, where our goal is to accurately delineate and identify specific organs (small bowel, large bowel, and stomach) within the MRI scans, the architecture of U-Net is effective since it can capture both the local and global features while preserving the fine details [5]. When training a U-Net model for organ segmentation, a loss function is chosen. A loss function quantifies the difference between the predicted output of a model and the true target values [5]. A suitable loss function is used to measure the dissimilarity between the predicted segmentation mask and the segmentation truth mask.

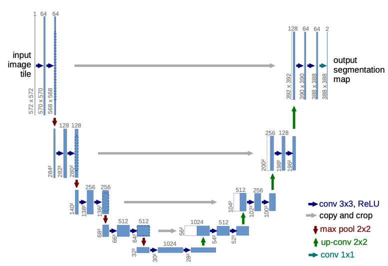

Figure 1: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations [6].

One of the key strategies that our group has decided to employ in this project is to use pre-trained models like MobileNet v2, ResNet18, and EfficientNet B1, for the automated segmentation of the three organs (small bowel, large bowel, and stomach) from the MRI scans. Through TL, we can fine-tune the pre-trained models to adapt them to achieve our specific task of image segmentation, which enhances efficiency and accuracy.

Through automated segmentation, healthcare professionals will be able to obtain clear and detailed visualizations of the segmentations. These visual representations enable radiation oncologists to inspect and analyze the precise positions and contours of the organs (through the coloured pixels on the image). This step would save radiation oncologists time from manually segmenting as they can immediately create a personalized treatment plan for cancer patients based on the results.

Using MobileNet v2, ResNet18, and EfficientNet B1, researchers can use the data extracted from external datasets to draw insights that can be used to enhance the segmentation performed in MRI scans. We can focus on bolstering the accuracy and robustness of our model based on the pre-trained models, by transferring relevant information from the previous dataset to our current dataset. As a result, the performance of our model will improve for the MRI scans of cancer patients. Our evaluation aims to compare the accuracy and efficiency of using TL on different pre-trained models such as MobileNet v2, ResNet18, and EfficientNet B1 to gain valuable insights into the suitability of these models for the task of image segmentation in the treatment planning of cancer.

2 LITERATURE REVIEW

The goal of applying transfer learning to MRI segmentation does not aim to address any gaps in the existing literature, instead it aims to take the most promising results of others and apply it to this new task. In particular, models especially suited to image segmentation were reviewed. Many popular image classification models can't be used directly on medical image segmentation tasks as they are CNN suited for classification. Some popular examples are AlexNet [7] or other architectures trained on the ImageNet dataset. Other deep learning models have been developed to more accurately work on the task of segmentation, instead of classification [8].

The fully convolutional network (FCN) is the original of these segmentation networks [9]. The FCN replaces the fully connected layers of VGG16 with convolutional layers and uses a deconvolution to provide a segmentation result for every pixel in an image. Segnet [10] is an alternative to FCN because it trains faster due to not requiring learning when upsampling.

U-Net [5] is one of the most popular models growing out of the interest in deep learning-based semantic segmentation. It is a deep learning network with an encoder-decoder architecture, and is well-suited for medical image segmentation tasks [5]. U-Net uses upsampling and downsampling as opposed to the fusion step in FCN [11]. It also takes advantage of directly using shallow features called skip connections. U-Net has become so popular that there are now many different models which claim the U-Net title. The original paper [11] called one specific model U-Net, but the modern definition has expanded to mean any encoder-decoder system can be defined as a "U-Net" system.

Given that medical imaging is a field of sparse data, there was interest in research that combined segmentation architectures with classification models that had been trained on generous amounts of image data. Pravitasari et al. have shown that transfer learning is possible in these situations by combining VGG16 with the U-Net architecture as a way of simplifying it [12]. This transfer learning work proved very effective for segmenting brain tumors on MRI, a task similar to the goal of segmenting MRIs of the GI tract. InceptionV3 and EfficientNetB4 were also shown to be compatible as one of U-Nets encoder network architectures. This works through allowing the convolutional neural networks to act as backbones for U-Nets encoders [13]. Due to the feature described above, a library has been created to provide many encoder backbones for the U-Net architecture. Segmentation Models is a python library which provides 9 different model architectures for segmentation, and 124 possible encoders [14]. When this library is combined with the HuggingFace PyTorch Image Models library there are over 500 possible encoders which can be used with U-Net [15].

In particular, the authors focused on MobileNet, ResNet, and EfficientNet as backbone encoders for the U-Net architecture. MobileNet is an architecture which uses depth-wise separable convolutions and two global hyper-parameters to trade-off between latency and accuracy [16]. ResNet is an architecture which enabled significantly deeper networks than those previous, including an eight-times deeper network than VGG nets but with less complexity [17]. These ResNets take advantage by formulating their layers as residual functions with reference to the layer inputs, instead of learning these functions unreferenced. EfficientNet is a technique that has lead to numerous leading architectures through balancing depth, width, and resolution [18]. This is in contrast to most convolutional neural networks, which take a depth-first approach and only scale when additional resources become available [18].

3 METHODOLOGY AND PRE-PROCESSING

To achieve the goal of evaluating the effectiveness of transfer learning for GI tract segmentation, multiple models were selected. Each of these models was an image classifier that had been trained on Imagenet and would be used as the basis for transfer learning. MobileNet, ResNet, and EfficientNet were chosen as candidate pre-trained models [19] to attempt transfer learning on.

Given that these models are classifiers, additional modifications are also needed to suit the task of segmentation. This resulted in the combination of U-Net with these underlying models to provide a transfer-learning environment and segmentation capability. This was seen to be possible through the literature review. Our objective was to accurately segment and identify different regions within the MRI images related to the GI tract. The dataset contained images along with corresponding masks that delineate the regions of interest, which are the large bowel, small bowel, and stomach.

First and foremost, we defined various hyperparameters and settings for the training process in a class called "CFG" on configuration settings. These include parameters such as the model architecture (U-Net with backbones MobileNet v2, ResNet18, and EfficientNet B1), batch size, learning rate, number of epochs, image size, and number of folds.

Secondly, in the data pre-processing pipeline, the dataset is initially loaded from a CSV file, and cleaning operations are performed to handle missing or erroneous information in columns such as 'segmentation' and 'rle len'. The dataset is then organized by grouping entries based on unique image IDs, generating lists of segmentations, and calculating total RLE mask lengths per image. The presence of empty masks is identified through a newly created 'empty' column. Further transformations involve adjusting file paths and creating functions to convert image IDs to corresponding masks, as well as transforming masks between grayscale and RGB formats for visualization. Image and mask loading functions are defined to read and normalize pixel values. The dataset is split using Stratified Group K-Fold cross-validation, with options for reducing the dataset size for debugging.

Data augmentation is introduced through Albumentations transformations, including resizing, flips, rotations, distortions, and dropout. A custom PyTorch dataset class is implemented to facilitate efficient loading and pre-processing, leading to the creation of data loaders for both training and validation. Overall, these pre-processing steps ensure the dataset is appropriately formatted, cleaned, and augmented to train a U-Net model effectively for medical image segmentation tasks. To note, the Sratified Group K-Fold cross-validation is used to split the dataset into training and validation sets, ensuring that each fold maintains the class distribution.

Thirdly, the model architecture is constructed using the U-Net architecture from the segmentation models pytorch library. Specifically, MobileNet v2, Resnet18, and EfficientNet B1 serve as the chosen backbones, offering a lightweight convolutional neural network tailored for mobile and edge devices. Configured for semantic segmentation, the model takes an input image and produces a mask with pixel-wise predictions for each class.

In the implemented U-Net architecture for medical image segmentation, the model follows a classic encoder-decoder structure. The encoder, represented by MobileNet v2, Resnet18, and EfficientNet B1, serves as the initial component responsible for extracting hierarchical features from the input image. Renowned for its lightweight design, MobileNet v2, Resnet18, and EfficientNet B1 efficiently capture features at varying scales. Subsequently, the decoder, implicitly defined within the segmentation models pytorch library, reconstructs the high-resolution segmentation mask by upsampling the encoded features. The decoder incorporates transposed convolutions and skip connections to recover spatial information lost during encoding. Together, this encoder-decoder combination forms the U-Net architecture, enabling the model to effectively leverage both global and local contextual information for precise medical image segmentation. The choice of MobileNet v2, Resnet18, and EfficientNet B1 as the encoder strikes a balance between computational efficiency and feature representation, contributing to the model's overall effectiveness.

For effective model training, a variety of loss functions are defined, encompassing Jaccard Loss, Dice Loss, Binary Cross-Entropy (BCE) Loss, Lovasz Loss, and Tversky Loss. These diverse loss functions collectively capture different aspects of the segmentation task, contributing to a comprehensive training approach.

The training loop, implemented in the run training function, consists of epochs containing both training and validation phases. Within this loop, the model undergoes training on the training dataset, employing the specified loss functions and optimization techniques such as the Adam optimizer with weight decay. Techniques like gradient accumulation and automatic mixed precision (AMP) are applied for enhanced training efficiency, while a learning rate scheduler dynamically adjusts the learning rate during training. Weights and Biases (wandb) are employed for comprehensive metric logging and real-time monitoring of the training process.

Model evaluation occurs during the validation phase using the valid one-epoch function. This evaluation encompasses the calculation of validation loss, Dice coefficient, and Jaccard index. The best-performing model, determined by the highest Dice coefficient on the validation set, is saved to ensure optimal performance on unseen data.

In the final stages of the pipeline, a test dataset is prepared, and the model trained on each of the five folds is loaded to perform evaluations. The best-performing model for each fold is utilized to generate predictions on the test set. The evaluation results, including predictions and corresponding input images, are visualized for a representative batch.

Throughout the entire process, the training progress is closely monitored using Weights and Biases (wandb), ensuring thorough logging of loss values and evaluation metrics. The best model is saved for future use, and key metrics, such as the best Dice coefficient achieved during training, are reported to assess the model's overall effectiveness.

The training process is summarized, and the best Dice coefficient achieved is reported. A visual representation of each of the three pre-trained model's predictions on a sample of test images (MRI scans) is displayed.

4 DATA COLLECTION AND PRE-PROCESSING

The dataset that we are planning to use is from a Kaggle competition: UW- Madison GI Tract Image Segmentation. The dataset is 2.47GB with 38496 files. Each file is an MRI scan of the region of the stomach and intestines. The MRI scans are from actual cancer patients who had 1 to 5 MRI scans on separate days during their radiation treatment. Each scan is in 16-bit grayscale PNG format. The goal is to segment organ cells in images. The true label file is train.csv with id, class, and segmentation columns. Id is the patient id on a specific day with a different slice number. 57% of id do not have labels while the other 43% have true labels. The class has values: `large_bowel`, `small_bowel`, and `stomach`. Segmentation is the true label which is provided as RLE-encoded masks. Every image pixel has a value between 0 and 1 which is already normalized, and there is no missing value.

Initially, for data pre-processing, when we were attempting to use pre-trained models and build a custom pre-trained model to compare results, three steps were performed:

- Convert RLE-encoded masks to binary mask
- Crop every image
- 5-fold cross-validation (based on the patient case)

First and foremost, we initially converted RLE-encoded masks to binary masks so that later on we can visualize the image, and then we crop every image to the same size since some of the images have different sizes. This also applies to their respected binary mask. Lastly, we split the dataset into 5 partitions, which makes it easier to perform 5-fold cross-validation. However, we modified our steps for data pro-processing to better prepare the dataset of MRI scans for training a U-Net model, as outlined in the Methodology section. The data pre-processing steps involve loading and cleaning the dataset, organizing data by image ID, handling empty masks, transforming file paths, implementing Stratified Group K-Fold splitting, introducing data augmentation using Albumentations, defining a custom PyTorch dataset class, and creating data loaders for efficient training and validation. These steps ensure that the dataset is appropriately formatted, cleaned, and augmented for effective training of the U-Net model. The inclusion of Stratified Group K-Fold cross-validation and a diverse set of data augmentation techniques contributes to the robustness and generalization of the trained model.

5 IMPLEMENTATION DETAILS

As a group, we collectively decided that Python would be the best programming language to use for our CISC 867: Deep Learning Project as it is the most popular language used and many libraries supported in machine learning are available for use. Specifically, our group has selected Numpy, Pandas, Tensorflow, Keras, Matplotlib, PyTorch, SKlearn, and Albumentation as the libraries to use.

For our variable naming convention, we are using snake case (e.g. `case123_ day20_slice_0001`). We also enabled virtual environments, which solved the incompatible versions issue. In addition, we found that using Kaggle and Jupyter Notebook was useful to help us with debugging and visualizing the results of each step. GitHub was also used to share code and Microsoft Teams was used as our primary source of communication and to schedule weekly meetings to ensure we were on the right track.

A specific challenge that we had was that the pre-trained models required an $nxnx3$ input, which is RGB channels, and the dataset is all grey-scaled.

To address the issues, two solutions were provided:

(1) Duplicate the grey-scaled channel three times and stack it, feeding it as if it were an RGB channel. This is one of the common ways of solving the problem, based on what we found on the internet.
(2) Using Keras built-in function that will colourize the image into RGB and then fetch it to the pre-trained models.

We decided on the latter solution as all the pre-trained models would come from the Keras libraries which should minimize conflicts when implementing. We were also planning to use the 3D Hausdorff Distance as a metric but we soon realized we did not need to use it, but rather, focus on testing different loss functions like Jaccard Index. However, other issues that we experienced included the splitting of the dataset. We noticed that not every image had the same size and dimensions. Another challenge we experienced was that we found it difficult to find pre-trained models for segmenting, and initially, we believed that we would be doing image classification for tumours. However, we quickly realized that we would be focusing on organ and image segmentation.

First and foremost, to overcome the issue of varying image sizes, a data pre-processing step is implemented. Albumentations library is utilized to resize images to a consistent and specified dimension (CFG.imgsize). This ensures uniformity in input dimensions, mitigating challenges associated with images of different sizes. Secondly, the code corrects an initial misconception regarding the nature of the task. Initially expecting image classification for tumours, it was realized that the focus should be on organ and image segmentation. Consequently, a U-Net architecture, specifically designed for semantic segmentation tasks, is employed. The model, implemented using the segmentation models pytorch library, utilizes MobileNet V2, ResNet 18, and EfficientNet B1 as the backbone for its lightweight design suitable for mobile and edge devices. The U-Net is configured to output pixel-wise masks for each class, making it well-suited for accurate organ and image segmentation. These adjustments in the code effectively address the challenges posed by varying image sizes and align the model architecture with the correct segmentation task.

6 EXPERIMENTAL DESIGN

Two metrics that our group plans to use include the Dice Coefficient and the Jaccard Index to evaluate the performance of the segmentation model. These metrics are essential for assessing the accuracy of the model in delineating and capturing the regions of interest in medical images.

The Dice coefficient is a metric that closely compares the predicted segmentation and ground truth on a pixel-wise agreement [20]. This serves as a quantitative measure of the similarity between the dataset that we are working on. Each case comprises multiple sets of scan slices with various organ positions, separated by the day that the scan occurred. The Dice coefficient is powerful in evaluating the segmentation of images of GI tract delineation and the three organs (small bowel, large bowel, and stomach).

The Jaccard Index, or Intersection over Union - IoU, is another metric assessing the overlap between the predicted and ground truth regions. It is calculated as the intersection of the predicted and true masks divided by the union of their areas. Like the Dice coefficient, the Jaccard index also ranges from 0 to 1, with higher values indicating better segmentation accuracy [21].

These metrics provide quantitative measures of how well the model is capturing the structures of interest in medical images. In the training loop, these metrics are calculated for both training and validation phases, and their values are monitored and logged using Weights and Biases (wandb). The best model is saved based on the highest Dice coefficient, ensuring that the model with superior segmentation performance on the validation set is retained for later use.

7 RESULTS AND ANALYSIS

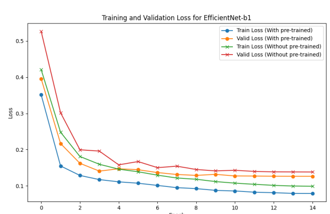

Figure 2: Training and Validation Loss for Pre-Trained Model EfficientNet B1

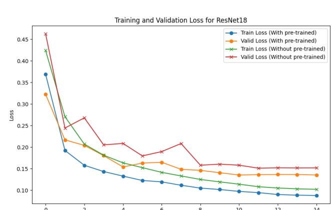

Figure 3: Training and Validation Loss for Pre-Trained Model ResNet 18

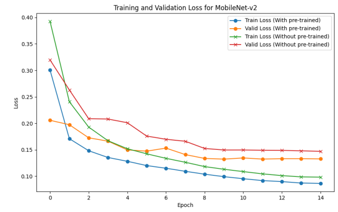

Figure 4: Training and Validation Loss for Pre-Trained Model MobileNet v2

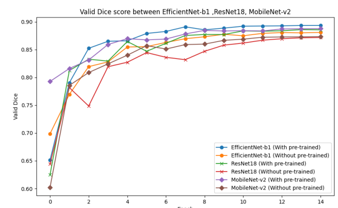

Figure 5: Dice Coefficient Score for Pre-Trained Models EfficientNet B1, ResNet 18, and MobileNet v2

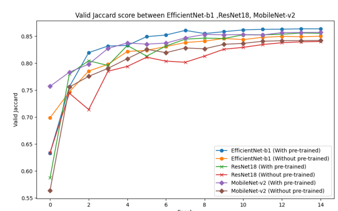

Figure 6: Jaccard Index Score for Pre-Trained Models EfficientNet B1, ResNet 18, and MobileNet v2

Valid Dice	Efficient b1	ResNet18	MobileNet v2
With Pre-trained weights	0.89	0.89	0.89
Without Pre-trained weights	0.88	0.87	0.87

Valid Jaccard	Efficient b1	ResNet18	MobileNet v2
With Pre-trained weights	0.86	0.86	0.86
Without Pre-trained weights	0.85	0.84	0.84

Figure 7: Dice Coefficient and Jaccard Index Score for Pre-Trained Models EfficientNet B1, ResNet 18, and MobileNet v2 with and without weights

Analyzing the loss curves of three pre-trained models (ResNet 18, MobileNet v2, and EfficientNet B1), it becomes evident that models initialized with pre-trained weights exhibit lower starting points and consistently maintain lower losses during training compared to their counterparts without pre-trained weights. This trend indicates that pre-trained weights serve as a beneficial starting point, facilitating more efficient convergence.

In terms of Dice coefficient and Jaccard index scores, both metrics consistently reveal that models with pre-trained weights achieve higher scores across epochs compared to models without pre-trained weights. For instance, the Dice coefficient scores with EfficientNet

B1 are 0.89 with pre-trained weights and 0.88 without. Similarly, ResNet18 and MobileNet v2 show higher scores with pre-trained weights compared to their non-pre-trained counterparts (ResNet18: 0.89 with, 0.87 without; MobileNet v2: 0.89 with, 0.87 without). The Jaccard index scores follow a similar trend, with EfficientNet B1, ResNet18, and MobileNet v2 achieving higher scores with pre-trained weights compared to those without (EfficientNet B1: 0.86 with, 0.85 without; ResNet 18: 0.86 with, 0.84 without; MobileNet v2: 0.86 with, 0.84 without).

Although the performance differences are close, suggesting overall good model performance, the pre-trained models exhibit a slight advantage. In summary, pre-trained weights demonstrate effectiveness across all metrics (loss, Dice, Jaccard), aligning with the expectation that pre-training enables models to begin with learned features, reducing the learning curve and potentially achieving superior final performance.

Regarding model comparison, with pre-trained weights, there is a negligible difference in performance among the different architectures (EfficientNet B1, MobileNet v2, and ResNet18). However, without pre-trained weights, slight variations emerge, but they are not substantial, indicating the robustness of the architectures or the task's limited variation in influencing distinct model performances.

The analysis of segmentation models designed for the identification of three distinct organs (small bowel, large bowel, and stomach) reveals insightful patterns aligning with the research objectives. One noticable observation is the effectiveness of pre-trained weights, as evidenced by consistently lower training and validation losses in models equipped with such weights. This signifies that leveraging pre-existing knowledge provides a valuable starting point for the learning process, contributing to improved convergence and accurate organ segmentation.

Furthermore, the higher Dice coefficient and Jaccard Index scores across epochs for models with pre-trained weights emphasize their superiority in accurately delineating organ boundaries. This achievement directly addresses the research objective of building models capable of precise segmentation. Interestingly, when employing pre-trained weights, there is a negligible difference in performance among different architectures, including EfficientNet B1, MobileNet v2, and ResNet18. This finding underscores the flexibility of using various architectures with comparable effectiveness, enhancing the practical applicability of the segmentation models.

In the absence of pre-trained weights, models exhibit slight variations in performance. However, these differences are not substantial, highlighting the robustness of the chosen architectures for the segmentation task. This aligns with the research objective of testing the efficiency of the models, as even without extensive pre-training, they demonstrate commendable performance, showcasing adaptability to the segmentation task.

Consistency across all evaluated metrics, including loss, Dice coefficient, and Jaccard Index scores, further reinforces the effectiveness of pre-trained weights for organ segmentation. The obtained insights collectively contribute to the development of segmentation models that not only meet the objective of accurate organ identification of the small bowel, large bowel, and stomach but also offer flexibility and efficiency across different architectural choices.

8 DISCUSSION

While we were working on the project, we encountered various challenges and unexpected outcomes throughout its development. The initial misconception about the project's focus, assuming it involved image classification for tumours, underscored the need for clear project objectives and communication. This realization shifted the project's emphasis to organ and image segmentation, reinforcing the significance of a well-defined project scope.

The findings of this project hold significant implications for the field of medical image segmentation, particularly in the context of organ segmentation. Regarding metrics, the initial consideration of 3D Hausdorff Distance was revisited during the project. The decision to focus on evaluating different loss functions, particularly the Jaccard Index, instead, demonstrates the adaptability and responsiveness to evolving project needs. This highlights the iterative nature of model development and the flexibility to adjust strategies based on emerging insights. The consistent superiority of models with pre-trained weights in terms of loss, Dice, and Jaccard scores indicates that leveraging transfer learning can substantially enhance the accuracy and efficiency of segmentation tasks. This suggests that the models equipped with pre-trained weights are better equipped to recognize and delineate organ structures within medical images. Such improved segmentation accuracy is crucial in medical applications where precise organ identification is vital for diagnosis and treatment planning.

One of the notable strengths of the project is the comprehensive evaluation using multiple metrics, including loss curves, Dice coefficient, and Jaccard Index scores. This multi-faceted analysis provides a holistic understanding of the models' performance.

Limitations primarily stem from the availability of pre-trained models tailored for segmentation tasks, suggesting the need for a more extensive and diverse repository. However, our group was able to find three pre-trained models to use from the Keras Library, with which we had great success. The project's focus on diverse architectures, such as EfficientNet B1, MobileNet v2, and ResNet18, enhances its generalizability and applicability to different medical imaging scenarios.

One of the notable strengths of the project is the comprehensive evaluation using multiple metrics, including loss curves, Dice coefficient, and Jaccard Index scores. This multi-faceted analysis provides a holistic understanding of the models' performance. Additionally, the project's focus on diverse architectures, such as EfficientNet B1, MobileNet v2, and ResNet18, enhances its generalizability and applicability to different medical imaging scenarios.

Despite the project's strengths, other certain limitations should be considered. The study primarily focuses on the segmentation of three specific organs (small bowel, large bowel, and stomach). The generalizability of the findings to other organs or medical imaging tasks needs further exploration. The slight variations observed in models without pre-trained weights, although not substantial, warrant further investigation into the robustness of different architectures across a broader range of medical imaging datasets.

The project's success in achieving accurate segmentation of organs in medical images has potential applications in various healthcare domains. Precise organ segmentation is critical for computer-aided diagnosis, surgical planning, and monitoring disease progression. The demonstrated effectiveness of pre-trained models suggests their utility in real-world healthcare settings, where limited annotated data is often a challenge. The project's outcomes provide a foundation for developing practical tools that can assist medical professionals in interpreting and analyzing complex medical images.

One unexpected outcome was the negligible performance differences among different architectures when pre-trained weights were employed. While this may indicate the robustness of the architectures, it could also point to the task's limited variability or complexity. Challenges encountered included dataset-related issues, such as variations in image size and dimensions. The decision to focus on organ segmentation, rather than initial plans for tumor classification, arose from a deeper understanding of the dataset and its specific characteristics.

In summary, the project's findings contribute valuable insights into the application of pre-trained models for medical image segmentation. The developed models showcase the potential for accurate identification of organs in medical images, with implications for diagnostic and treatment planning. The insights gained from this project can contribute to advancements in medical image analysis, enhancing the efficiency of segmentation models. The strengths lie in the project's thorough evaluation, diverse architectural considerations, and potential real-world applications. Limitations highlight the need for further research in broader medical imaging contexts. Unexpected outcomes and challenges underscore the iterative nature of the research process, emphasizing the importance of adaptability and refinement in addressing unanticipated issues.

9 CONTRIBUTIONS

Potential applications of the project can extend to medical imaging tasks, particularly in the field of organ segmentation, and advance the current understanding of transfer learning techniques in medical image segmentation. By systematically comparing models with and without pre-trained weights, the study establishes the superiority of leveraging pre-trained features for accurate organ segmentation. This insight contributes to refining transfer learning strategies for medical imaging tasks, offering a practical guideline for future research in similar domains.

The demonstrated effectiveness of pre-trained weights in improving model performance, as evidenced by lower loss curves and higher Dice coefficient and Jaccard Index scores, contributes to enhancing existing segmentation models. This improvement is crucial for medical applications where precision is paramount. The findings can guide researchers and practitioners in selecting optimal models and strategies for medical image segmentation, thereby elevating the quality of healthcare diagnostics and treatment planning.

The project addresses practical challenges encountered in medical imaging, such as variations in image size and dimensions. The decision to focus on organ segmentation, informed by the dataset's characteristics, highlights the project's adaptability to real-world challenges. This contributes valuable insights into the iterative process of refining project objectives based on the intricacies of the data, paving the way for more effective and targeted applications of AI in healthcare.

The accurate segmentation of organs, facilitated by the project's findings, holds broader implications for the healthcare industry. The project provides a foundation for developing practical tools that can assist medical professionals in interpreting and analyzing complex medical images. These tools have the potential to streamline diagnostic processes, improve surgical planning, and enhance patient outcomes. The focus on pre-trained models aligns with practical considerations in the industry, where limited annotated data is often a challenge.

The project's exploration of different architectures, including EfficientNet B1, MobileNet v2, and ResNet18, contributes to the understanding of robust model selection. The negligible performance differences observed among these architectures with pre-trained weights suggest that practitioners can choose models based on factors such as computational efficiency without sacrificing segmentation accuracy. This insight has implications for resource allocation and scalability in deploying AI models for medical imaging tasks.

10 CONCLUSION

The project aimed to develop accurate models for segmenting three organs (small bowel, large bowel, and stomach) from medical images while evaluating the efficiency and effectiveness of the models. Noteworthy findings emerged from the study. Pre-trained weights significantly improved model performance, evident in lower and consistently maintained loss curves, as well as higher Dice and Jaccard scores. This underscored the importance of leveraging pre-trained features, facilitating better convergence during the learning process.

The exploration of three architectures (EfficientNet B1, MobileNet v2, and ResNet18) revealed minimal performance differences when pre-trained weights were employed. Despite slight variations without pre-trained weights, the architectures exhibited overall robustness. The project's objectives evolved from tumour classification to organ and image segmentation, highlighting the need for adaptability in response to dataset characteristics.

Challenges included dataset variations in size and dimensions, necessitating careful pre-processing, and initial difficulties in finding suitable pre-trained models for segmentation. Despite these challenges, the project can be deemed successful in achieving its goals. The developed models demonstrated proficiency in organ segmentation, providing insights into robust model selection for medical image segmentation.

For future work, refinements in pre-processing techniques could address image variability more effectively. Exploring additional architectures and variations may yield improvements or specialized models for organ segmentation. Integrating 3D metrics, such as the initially considered 3D Hausdorff Distance, could enhance the evaluation of segmentation accuracy. Expanding the dataset and validating models with real-world clinical data could further contribute to the ongoing advancement of AI in medical imaging applications.

In conclusion, this project advances the current understanding of transfer learning in medical image segmentation, improves segmentation model performance, addresses practical challenges in medical imaging, has broader implications for the healthcare industry, and contributes to the selection of robust models. These contributions collectively contribute to the evolution and applicability of AI/ML in medical domains, with the potential to positively impact patient care and diagnostic processes.

REFERENCES

[1] "UW-Madison GI Tract Image Segmentation," Apr. 2022. [Online]. Available: https://kaggle.com/competitions/uw-madison-gi-tract-image-segmentation

[2] I. J. Das, P. Yadav, and B. B. Mittal, "Emergence of MR-Linac in Radiation Oncology: Successes and Challenges of Riding on the MRgRT Bandwagon," *Journal of Clinical Medicine*, vol. 11, no. 17, p. 5136, Jan. 2022, number: 17 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2077-0383/11/17/5136

[3] A. Abu Mallouh, Z. Qawaqneh, and B. D. Barkana, "Utilizing CNNs and transfer learning of pre-trained models for age range classification from unconstrained face images," *Image and Vision Computing*, vol. 88, pp. 41–51, Aug. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885619300629

[4] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, Dec. 2017.

[5] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 1055–1059, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/9053405

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.* Springer, 2015, pp. 234–241.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[8] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine*, vol. 8, no. 11, p. 713, Jun. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327346/

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," 2015, pp. 3431–3440. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7803544/

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, arXiv:1505.04597 [cs]. [Online]. Available: http://arxiv.org/abs/1505.04597

[12] A. A. Pravitasari, N. Iriawan, M. Almuhayar, T. Azmi, I. Irhamah, K. Fithriasari, S. W. Purnami, and W. Ferriastuti, "UNet-VGG16 with transfer learning for MRI-based brain tumor segmentation," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 3, pp. 1310–1318, Jun. 2020, number: 3. [Online]. Available: http://telkomnika.uad.ac.id/index.php/TELKOMNIKA/article/view/14753

[13] A. Abedalla, M. Abdullah, M. Al-Ayyoub, and E. Benkhelifa, "Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures," *PeerJ Computer Science*, vol. 7, p. e607, Jun. 2021, publisher: PeerJ Inc. [Online]. Available: https://peerj.com/articles/cs-607

[14] P. Iakubovskii, "qubvel/segmentation_models.pytorch," Dec. 2023, original-date: 2019-03-01T16:21:21Z. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch

[15] "huggingface/pytorch-image-models," Dec. 2023, original-date: 2019-02-02T05:51:12Z. [Online]. Available: https://github.com/huggingface/pytorch-image-models

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," Apr. 2017, arXiv:1704.04861 [cs]. [Online]. Available: http://arxiv.org/abs/1704.04861

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03385

[18] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sep. 2020, arXiv:1905.11946 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1905.11946

[19] P. Iakubovskii, "Segmentation models pytorch," https://github.com/qubvel/segmentation_models.pytorch, 2019.

[20] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, "Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations," Jun. 2019, arXiv:1906.11031 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1906.11031

[21] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

Received 1 December 2023; revised 8 December 2023