UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio Politécnico

Impacto del Transfer Learning en redes neuronales híbridas para Segmentación de Imágenes Médicas

Xavier Eduardo Casanova Pabón

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero Industrial

Quito, 14 de mayo de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio Politécnico

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Impacto del Transfer Learning en redes neuronales híbridas para Segmentación de Imágenes Médicas

Xavier Eduardo Casanova Pabón

Nombre del profesor, Título académico **Gabriela Baldeón, PhD.**

Quito, 14 de mayo de 2024

© DERECHOS DE AUTOR

Nombres y apellidos: Xavier Eduardo Casanova Pabón

Código: 00339990

Cédula de identidad: 1721623070

Lugar y fecha: Quito, 14 de mayo de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La resonancia magnética (RM) es una técnica de imagen médica crucial que ha revolucionado el diagnóstico y tratamiento de enfermedades. Especialmente, la segmentación de la próstata en RM ayuda a detectar precozmente el cáncer. Las redes neuronales convolucionales (CNN) han sido efectivas en la segmentación, pero los Vision-Transformers (ViT) emergen por su habilidad para mejorar la comprensión contextual, unificando el procesamiento del lenguaje natural (NLP) con tareas visuales. A pesar de su dependencia en grandes datos, el aprendizaje por transferencia permite entrenar modelos en datos específicos. Este estudio investiga si el aprendizaje por transferencia beneficia a modelos pequeños entrenados con imágenes no médicas para la segmentación de la próstata en RM. Se analizó el impacto del aprendizaje por transferencia en el rendimiento de los Transformers, concluyendo que no hay diferencias estadísticas significativas en la precisión de la segmentación, lo que sugiere un potencial para la implementación de IA en medicina sin la necesidad de grandes conjuntos de datos médicos especializados.

Palabras clave: Resonancia Magnética, Próstata, Inteligencia Artificial, Segmentación Semántica, Decatlón de Segmentación Médica, Redes Neuronales Convolucionales, Transformadores de Visión, Aprendizaje por Transferencia.

ABSTRACT

Magnetic resonance imaging (MRI) is a crucial medical imaging technique that has revolutionized the diagnosis and treatment of diseases. In particular, MRI segmentation of the prostate helps to detect cancer early. Convolutional neural networks (CNNs) have been effective at segmentation, but Vision-Transformers (ViT) emerge for their ability to improve contextual understanding by unifying natural language processing (NLP) with visual tasks. Despite its reliance on big data, transfer learning allows models to be trained on specific data. This study investigates whether transfer learning benefits small models trained with non-medical images for prostate segmentation on MRI. The impact of transfer learning on Transformers' performance was analyzed, concluding that there are no statistically significant differences in the accuracy of segmentation, suggesting a potential for the implementation of AI in medicine without the need for large specialized medical datasets.

Key words: Magnetic Resonance, Prostate, Artificial Intelligence, Semantic Segmentation, Medical Segmentation Decathlon, Convolutional Neural Networks, Vision-Transformers, Transfer Learning.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Medical imaging is essential for the accurate diagnosis of diseases, as it offers detailed, non-invasive information about the anatomy and functioning of the organs and systems of the human body. Magnetic Resonance Imaging (MRI) is an advanced medical imaging technology that has transformed the diagnosis and treatment of various medical conditions. It takes advantage of the phenomenon of atomic nuclei resonating in a magnetic field to produce detailed images of the body. Using radio waves and magnetic fields, it generates high-resolution images of anatomical structures and soft tissues, such as the brain, spinal cord, muscles, and internal organs (Mayo Clinic, 2021). This imaging modality allows healthcare professionals to detect abnormalities, track disease progressions, and design the most appropriate treatment strategies, thereby improving care and outcomes for patients.

The prostate is a gland in the male reproductive system that plays a crucial role in semen production, and is located just below the bladder, surrounding the urethra. The prostate is divided into 3 glandular zones and a non-glandular zone called the anterior fibromuscular stroma (Ortiz-Hidalgo & Heredia-Jara, 2022). The glandular areas are the peripheral zone (PZ), central zone (CZ), and transitional zone (TZ) (Ortiz-Hidalgo & Heredia-Jara, 2022). The PZ is the largest region of the prostate and is located at the back of the gland and is where approximately 70-80\% of prostate cancers originate (Lee et al., 2014). The CZ is the second largest region and surrounds the urethra (Ortiz-Hidalgo & Heredia-Jara, 2022)., it is located in the central part of the prostate. The TZ is the smallest region, surrounding the urethra just below the bladder and is where 20 \% of the cancers originate, plus is prone to pathologically grow over time (Lee et al., 2014).

According to the American Cancer Society, on average, approximately 1 in 44 men will succumb to prostate cancer (American Cancer Society, 2024). The SEER database statistics from 2013 to 2019, shows that the 5-year relative survival rates were over 99\% for localized and regional prostate cancer, but drop to 34\% for distant-stage cases (American Cancer Society, 2024). Image segmentation has helped for more precise targeting during treatments such as radiation therapy, ensuring that the maximum dose is delivered to the cancer cells while minimizing damage to healthy tissue(Grégoire et al., 2020) and helping in monitoring the size and shape of the tumor during the treatment to assess how well the treatment is working to make the necessary adjustments (Beaton, Bandula, Gaze, & Sharma, 2019).

However, when segmentation tasks are done manually by health specialists, errors can be made because it is subject to subjectivity and could depend on human factors, as well as being laborious which can adversely affect the diagnosis and treatment of diseases. With the growing research and development of artificial intelligence in areas such as computer vision, the segmentation of medical images has aroused particular interest, because the accuracy of computers can surpass that of humans. For this reason, AI is becoming an essential tool to guide and correct the errors made.

In recent decades, numerous models and techniques have been developed to address this challenge. Early image segmentation approaches were based on classical image processing techniques, such as threshold (Karasulu & Korukoglu, 2011) and edge-based (Pujar, Gurjal, Kunnur et al., 2010) segmentation. Although effective under controlled conditions, these methods had limitations in terms of accuracy and generalizability (Palomino & Concha, 2009). With the advent of deep learning, especially convolutional neural networks (CNN), significant advances were made in the segmentation of medical images, as they demonstrated an

impressive ability to learn hierarchical representations of visual features, which allowed them to overcome challenges associated with variability in the appearance and morphology of anatomical structures.

In 2013, the architecture of R-CNN was published, this was one of the first CNN used for object detection and also for semantic segmentation RCNN, this architecture is known as a two-stage detector, that uses region proposals to localize objects within an image and then classifies those regions using a CNN. R-CNN proved that CNN's have a great potential for complex computer vision tasks. In FCN, Long et al. proposed the first CNN targeted specifically to segmentation tasks, called Fully Convolutional Networks (FCN). (LONG, SHELHAMER, & DARRELL, 2015) was composed entirely of convolutional layers to produce dense predictions over input images of arbitrary size. Ronneberger & Brox (Ronneberger, Fischer, & Brox, 2015) proposed a CNN targeted for biomedical image segmentation called UNet, named for the U-shape of its architecture that consists on a contracting path to capture context, encoder, and an expansive path to enable precise localization, decoder. This architecture has proven to be so effective that is still used and studied today. In 2017 Mask RCNN, Kaiming et al. proposed a CNN architecture based on Faster R-CNN called ``Mask R-CNN'' by adding a parallel branch for predicting segmentation masks alongside object detection, enabling instance segmentation with precise pixel-level delineation, and making clear that CNN's are the best suited architecture for any segmentation task. Later in 2018 UPerNet, Xiao et al., proposed an architecture called UPerNet that incorporates a Unified Perceptual Parsing (UPP) module and a Pyramid Attention (PA) module to enable efficient segmentation of various visual concepts within images through heterogeneous image annotations. The backbone of the network in the original paper was a

fully convolutional feature extractor based on the based on the Feature Pyramid Network (FPN) UPerNet, but this can easily be replaced by another pre-trained backbone.

However, in an effort to achieve a generalized artificial intelligence model that could integrate vision and language tasks, Dosovitskiy proposed the architecture of Vision-Transformers porting the attention mechanism of Transformers used for Natural Language Processing (NLP) to vision tasks. Vision-Transformers (ViT) is a structure first used for image classification, in which an input image is divided into fixed-sized patches, and each patch is represented as a vector, simulating the tokens of a sequence of text characters in a Natural Language Processing (NLP) network. These vectors feed through multiple layers of attention, where the network learns to capture complex features and spatial relationships between the pixels in the image. Finally, the output from the network is fed through a classifier layer to produce the final output (Dosovitskiy et al., 2021).

While segmentation models with Transformers are being extensively researched due to their ability to capture long-range relationships in images and to model the overall context of the scene, they require more training data to reach acceptable accuracy values (Thisanke et al., 2023).

To combat this problem, architectures that combine convolutional layers plus Vision-Transformers in their architecture started to be researched.

One network that mixed the concepts of CNN's and ViT to target a variety of tasks including semantic segmentation was Deep Prediction Transformer (DPT) DPT. DPT consist of a ViT based Encoder and a convolutional Decoder, this network showed state of the art results while not needing as many images as pure ViT. On the other hand on 2021 the architecture of Segformer Segformer came out featuring a hierarchical Transformer encoder

for multi-scale feature extraction paired with a simple and efficient Multilayer Perceptron (MLP) decoder, eliminating the need for positional encodings and making the computations faster. On 2022, Mehta \& Rastegari published an architecture called MobileViT that was intended to be used on mobile devices, making it lightweight and portable. To accomplish this, the authors created a ViT block called MobileViT-Block that used the attention mechanism of ViT and was placed between some convolutional blocks along the network MobileViT. Later the same year, the authors published an improvement on the MobileViT-Block MobileViTV2 by using a new mechanism called Separable-Self attention, that reduces the computational complexity in the transformer from a quadratic to a linear factor, making the network much faster and smaller, they named this improvement as MobileViT-V2. Another approach taken by Russakovsky et al., was to adapt a ViT based architecture as the backbone of a CNN for semantic segmentation. To accomplish this, they used an architecture called Swin and leveraged it as the backbone of UPerNet UperNetSwin, getting state of the art results for semantic segmentation.

All of the above mentioned networks were trained on big datasets for segmentation of natural images. However, due to the cost and complexity of acquiring and labelling medical imaging, MRI datasets are commonly small. Therefore, a method used to train a model in medical images and achieve a good performance is transfer learning. Transfer learning is a technique where a model trained for one task is adapted to perform another related task. This process involves taking a model that has already been trained on a large dataset (the source domain) and fine-tuning it with a smaller dataset for the new task (the target domain). By doing so, it leverages the features and knowledge the model has already learnedTransferLearning. The key idea is that instead of starting training from scratch, the pre-trained model has already learned general characteristics and features that can adapt more easily to specific tasks with

smaller datasets. This allows for faster and more efficient training, especially when training data is limited.

The aim of this work is to statistically analyze the effect transfer learning has on the performance of Transformer architectures on the task of prostate MRI segmentation. We test the Mobile ViT, Mobile ViT V2 and UPerNet with the Swin backbone in the Medical Segmentation Decathlon's Prostate challenge datasetDecathlon using the dice coefficient metric for evaluation. To obtain statistically significant results we perform a paired t-test. Our experiments demonstrate that there is no statistical difference on the mean Dice coefficient per class by using transfer learning, having values closer to 0.676 for the PZ region and 0.706 for the TZ region for the networks trained without transfer learning and

and 0.673 and 0.712 for the ones trained with transfer learning on the PZ and TZ regions respectively, on the validation dataset.

METHODOLOGY

In the next subsections we describe the dataset selected and pre-processing operations applied. We then proceed to explain the architectures tested and training hyper-parameters. Subsequently the evaluation metrics utilized are explained, and finally the statistical test performed is presented.

Dataset and Pre-processing

The Medical Segmentation Decathlon database Decathlon was part of the 2018 Decathlon challenge and has been largely used in various researches, becoming a standard for testing new segmentation models or techniques for medical image segmentation. For this reason, this database was selected in this study.

The prostate dataset from the Medical Segmentation Decathlon databaseDecathlon consists of 32 4D multiparametric magnetic resonance images (mpMRI) of the prostate, the two channels of the 4D volume consist on the transverse T2-weighted and the apparent diffusion coefficient (ADC), along with the ground truth segmentation masks of the peripheral (PZ) and transitional (TZ) regions. The majority of the images are of 320x320 pixels height and width, with a variable number of layers, ranging from 15 to 24 slices. This images were obtained from the Radboud University Medical Center and the Nijmegen Medical Centre in the Netherlands Decathlon.

To train and test the accuracy of the models, the images were divided randomly into 85\% observations for training and 15\% observations for validation, ending up with 28 MRI images for training and 4 for validation.

Each MRI image is a 4D volume with 2 channels containing measurements of the T2 and ADC protocols used in MRI imaging, figure \ref{fig:channels shows the two protocols on a slice of three patients, plus the ground truth labels of the whole prostate, and the PZ and TZ classes. The segmentation networks selected need two-dimensional images, so each slice of the volumes needed to be passed as a single image. To use the information of the two protocols, the T2 znd ADC modality are inserted in the channel dimension. Figure \ref{fig:data shows the images exported for the final dataset.

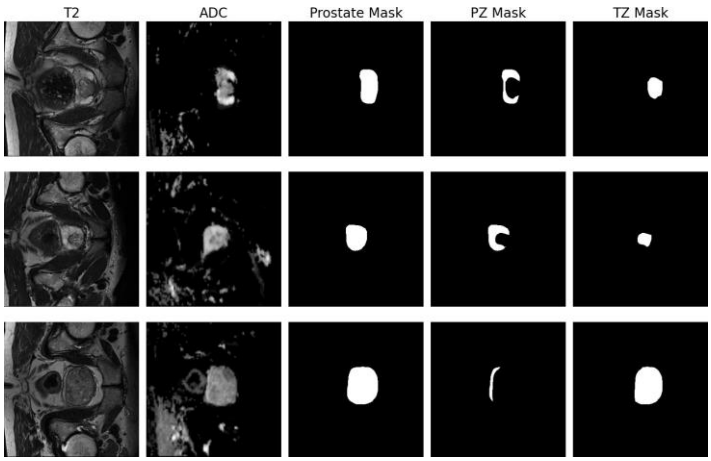

Fig. 1. MRI Protocols of the dataset on a slice of the same patient.

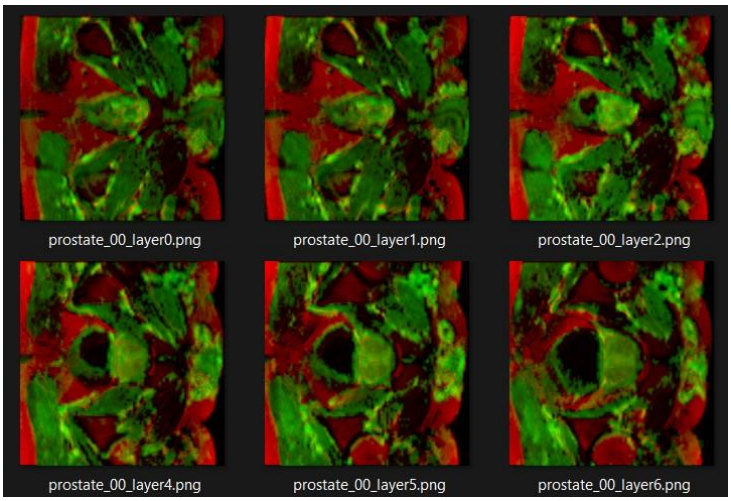

Fig. 2. Final dataset used for training and testing.

After exporting every slice of each volume as RGB images, the total number of png files was of 502 for the training dataset and 100 for the validation dataset. The architecture sizes tested for MobileViT and MobileViT V2 have input layers that accept images of size 512×512, on the other hand, UPerNet with the Swin Backbone requires input images of size 224×224 therefore the training and validation datasets were resized using the bi-linear interpolation method to match it. Furthermore, all the pre-trained backbones were trained on the Imagenet1kimagenet dataset and then fine-tuned to the pascal VOC2012 dataset for the semantic segmentation taskMobileViTMobileViTV2UperNetSwin, therefore every network was resized to the mean and standard deviation of the Imagenet1K dataset withpascal_voc:

$$height = 512, width = 512$$

$$\mu\_I = [0.485, 0.456, 0.406]$$

$$\sigma\_I = [0.229, 0.224, 0.225]$$

During training, data augmentation operations were applied to increase the size and diversity of the images using the albumentations library. The augmentation transformations used were:

- Center Crop: with (height=225, width=225, p=0.5).
- Affine Transformation: with (scale=0.8, translate_percent=0.2, rotate=0, shear=(-10, 10), p=0.5).
- Rotation: between -45 and 45 degrees, with p=0.5.
- Horizontal and Vertical Flips: with (p=0.5).
- Blur: with (blur_limit=7, p=0.5).
- Color Jitter: with (brightness=0.8, contrast=0.5, saturation=0.8, hue=0, p=0.5).

Selected models and Training details

The ViT architectures selected for comparison are the MobileViTMobileViT, MobileViT V2MobileViTV2, and UPerNet with Swin Backbone UperNetSwin. These models were chosen because they are efficient and lightweight, making them suitable for deployment in environments with limited computational resources such as mobile devices or embedded systems commonly used in healthcare settings EmbeddedAI; Despite their smaller size, these models can achieve high enough accuracy to be used in a professional setting; And because they are flexible and can be trained with less data, being well suited for the limited amount of images of the selected dataset. The weights for transfer learning were the available on the hugging face ``transformers" library. A brief description of each architecture is presented below:

- MobileViT: Uses the MobileViT transformer block between some convolutional layers to move the attention mechanism to a lightweight architecture. The model had a total of 6,353,315 trainable parameters.
- MobileViT V2: An enhancement of the original MobileViT architecture, it uses the separable self-service mechanism to reduce the complexity of the MobileViT block, making it faster and less computationally expensive. The model had a total of 9,757,612 trainable parameters.

- UPerNet: Uses the Swin architecture as the foundation of the original fully convolutional UPerNet network to improve accuracy while keeping it small and fast. The model had a total of 59,828,672 trainable parameters, making it the largest model tested.

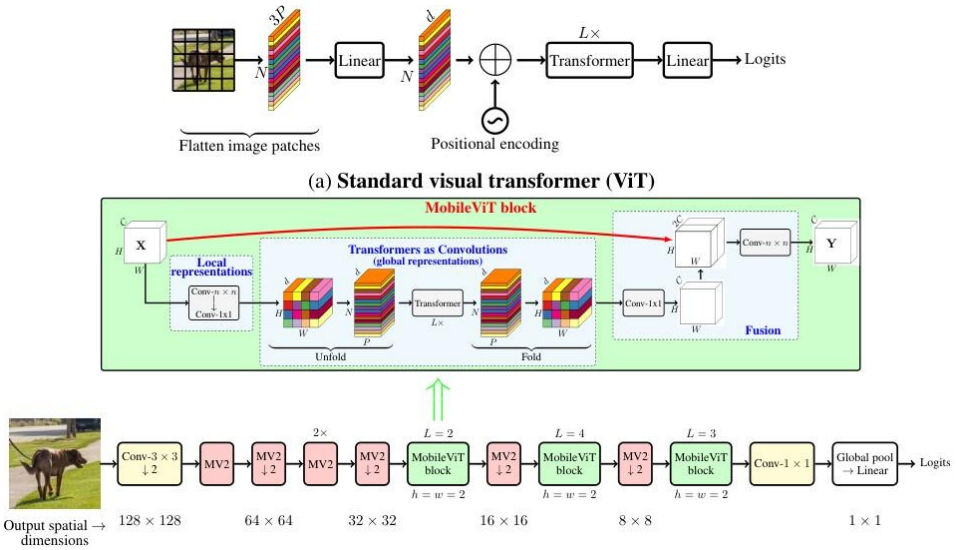

Fig. 3. MobileViT architecture, image from the original paper MobileViT

The training hyper-parameters for all the evaluated architectures are presented in Table \ref{table:hyperparams and were taken based on the ones used on the original papers. All implementations used Python 3.9.13, Pytorch 2.2.1 and run in a Nvidia GPU GTX4070. To get an easy interface and the availability to have pre-trained weights, the hugging face library called ``transformers" was used. From the models available in the library, the following three were selected with the smallest version of each architecture:

- MobileViT: weights "apple/deeplabv3-mobilevit-small"
- MobileViT V2: weights "apple/mobilevitv2-1.5-voc-deeplabv3"
- UPerNet with the Swin Transformer Backbone: "openmmlab/upernet-swin-tiny" weights

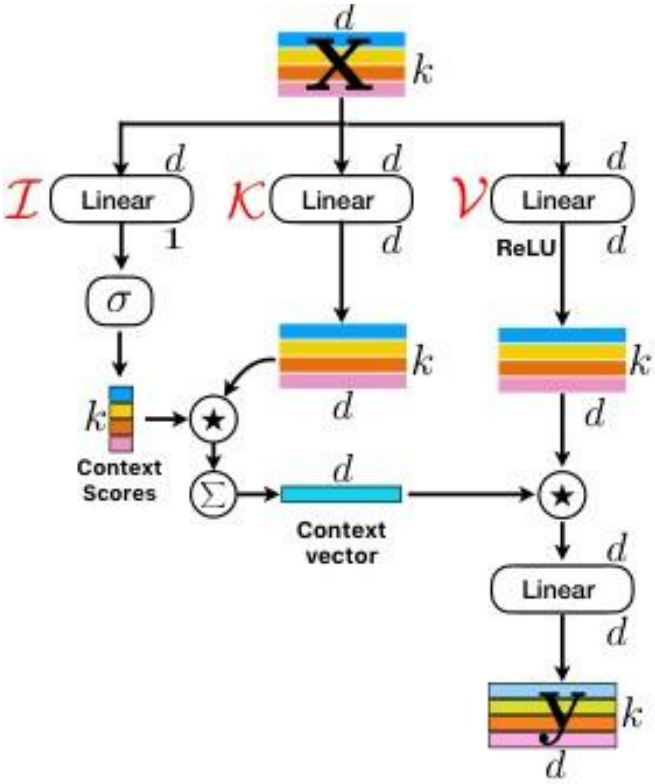

Fig. 4. Separable self-attention mechanism for the MobileViT-V2 architecture, image from the original paper MobileViTV2.

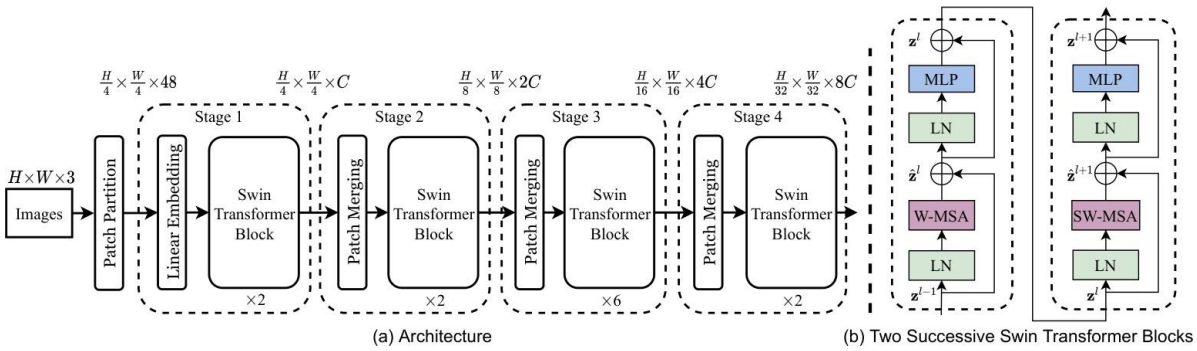

Fig. 5. Swin transformer architecture, image from the original paper UperNetSwin.

MobileViT was pre-trained from scratch on the Imagenet1K dataset for image classificationMobileViT which is a dataset comprised of approximately 1.2 million natural images for training, 50,000 for validation, and 100,000 for testing, with 1000 (1k) classesimagenet. To perform semantic segmentation, the trained model was integrated with DeepLabv3 and fine-tuned on the pascal VOC 2012 datasetMobileViT, that contains 11,540

natural images with 20 different object classes, including vehicles, household items, animals, and others. The training and validation set add to a total 11,530 images for object detection and segmentation tasks fsupervisedSurvey.

Using the same methodology, MobileViT V2 was trained exactly the same but was also tested being implemented with PSPNet and DeepLabv3 and on the ADE20k and PASCAL VOC 2012MobileViTV2, in this study, and to be consistent across models, we only considered the DeepLabv3 with VOC pretrained weights. Finally, UPerNet with the swin transformer architecture was trained on the ADE20k UperNetSwin that consists of 27,574 natural images, divided on 25,574 for training and 2,000 for testing, on 365 different scenes ADE20K.

	MobileViT	**MobileViT-V2**	**UPerNet**
Batch Size	16	16	10
Epochs	300	300	300
Optimizer	AdamW	AdamW	AdamW
Learning rate	8e-5	8e-5	8e-5
Loss Function	CrossEntropy	CrossEntropy	CrossEntropy

Fig 6: Table 1 Hyper-parameters used.

Evaluation Metrics

The Dice coefficient for semantic segmentation tasks can be expressed as:

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|}$$

It is the intersection of both areas divided by the total area of each mask being compared; notice that the sum of both areas is not the same as the joint area, as it's given from the set-theory that:

$$|A \cup B| = |A| + |B| - |A \cap B| \neq |A| + |B|$$

Thus the Dice coefficient needs the normalization factor of 2 in the numerator to be bounded between 0 and 1 with 0 meaning no similarity, and 1 a perfect match.

The Dice coefficient has some advantages like being differentiable, allowing it to be used as a loss function during training, or its correspondence to Precision being equivalent to the F1 measure, which is the harmonic average of precision and sensitivity, and thus providing a balance between precision and retrieval of relevant information LossAndMetrics.

Statistical Test

A paired t-test was conducted to test if the difference on Dice coefficient obtained by training a network with and without transfer learning was statistically significant. The number of samples for this test is the number of images on the validation dataset, therefore $n=100$.

The null hypothesis of the paired t-test states that the mean Dice coefficient of a network trained with Transfer Learning is the same as the mean Dice coefficient of a network trained without Transfer Learning, and can be expressed as:

$$\begin{cases} H_0 : d = 0 \\ H_a : d \neq 0 \end{cases}$$

This is a two tailed test, where:

$$d = t_1 - t_2$$

$t1 \wedge t2$ correspond to the mean dice of the network trained with and without transfer learning respectively. Under the assumption that the null hypothesis is true, the unbiased t statistic for the difference d on this test can be calculated as:

$$t = \frac{\overline{d} - \mu_d}{s_d / \sqrt{n}}$$

RESULTS

The training loop was run for each of the 3 defined networks with and without pre-trained weights, giving a total of 6 training processes. Figure 6 shows that the training progressed correctly on every step of the training loop, by reducing the loss function to values close to 0. Figure 7 shows the inference of every neural network with and without transfer learning at some steps of the training loop for one layer of a patient in the validation dataset to graphically have a sense of the effectiveness of it in the training process.

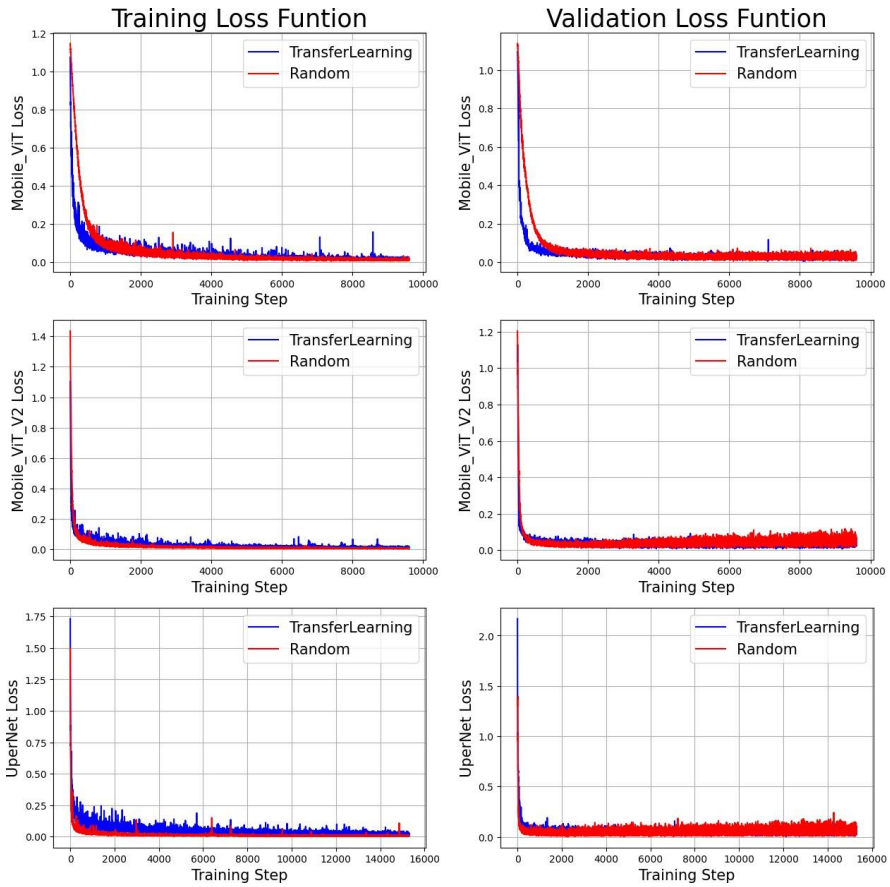

Fig. 7. Loss functions of every experiment conducted.

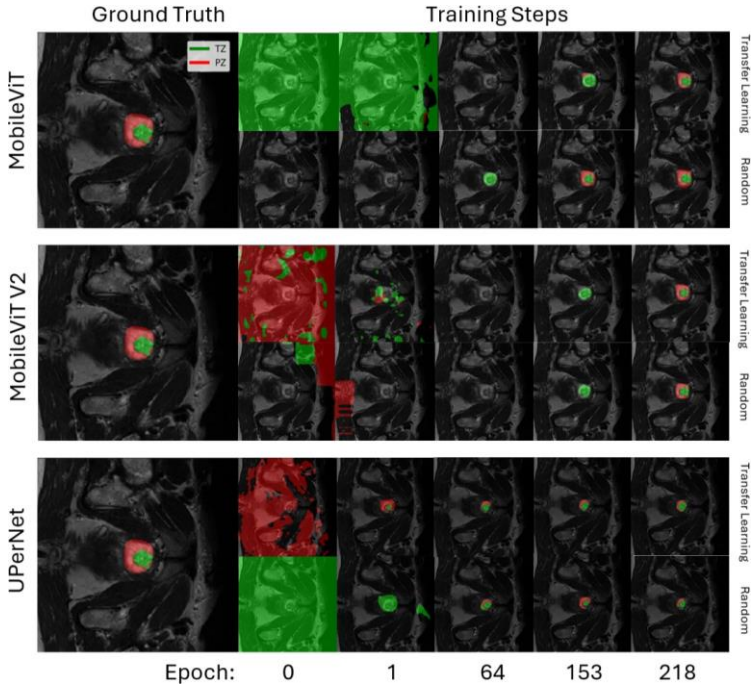

Fig. 8. Inference Results during training.

During training, a validation loop calculated the Dice co- efficient on the validation dataset and the step with the best weights was saved into memory, the results of the validation loops for the best Dice coefficients are reported in table II.

	mean Dice PZ	mean Dice TZ	Epoch
MobileViT (RND)	0.665	0.719	299
MobileViTV2 (RND)	0.684	0.693	172
UPerNet (RND)	0.678	0.705	294
MobileViT (T.L.)	0.648	0.719	203
MobileViTV2 (T.L.)	0.669	0.676	218
UPerNet (T.L.)	0.703	0.742	167

Fig 9: Table 2 Training loop report.

These weights were used to make inference over all the validation dataset and save the dice coefficients per patient and per class. In figure 8 we present examples of segmentation obtained from every network studied with and without transfer learning, we could observe that the segmentation inference for the middle layers was almost perfect, whereas the majority of

inference errors were made for the initial and final layers where the segmentation area was very small or irregular, this is shown in figure 9.

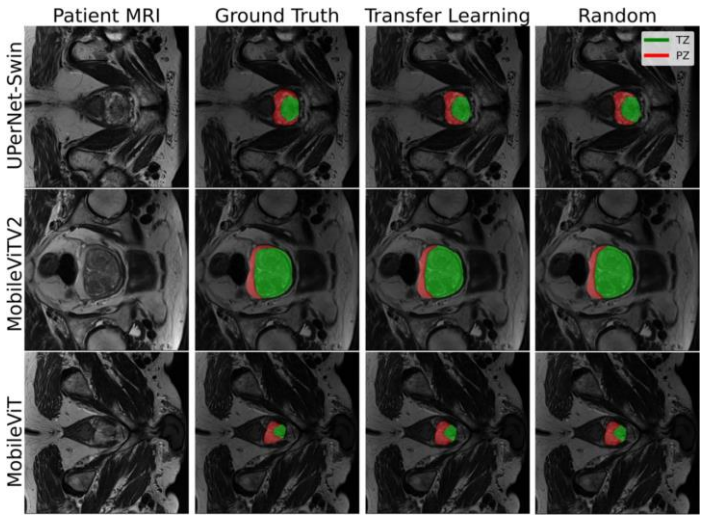

Fig. 10. Good Inference Results on middle layers of the volumes.

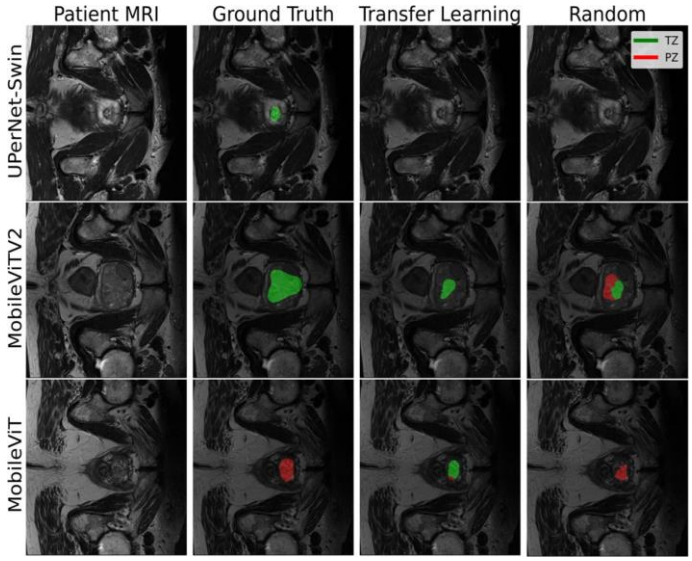

Fig. 11. Error Inference Results on initial layers of the volumes.

Statistical Test results

The results of the Hypothesis tests for the three networks are shown above. Tables \ref{table:PZTests \& \ref{table:TZTests show the p-values for the hypothesis on the dice coefficients for the PZ and TZ classes respectively, whereas table \ref{table:meanTests show the results of the test for the overall mean dice obtained for each network. We appreciate that

the p-values reported in every table is greater than 0.1, therefore, using the standard confidence value of $\alpha = 0.05$, we fail to reject the null hypothesis that transfer learning was significant for MobileViT, MobileViT V2 and UPerNet with Swin Transformers among every class, thus forcing us to conclude that there are no statistical differences in the mean Dice for the tested architectures if transfer learning is used to train them. This is further supported on mean dice coefficients that were obtained when extracting the weights on table \ref{table:ValDice where can be seen that they are indeed very similar to one another, except maybe for UPerNet, where the mean values seem to have increased when transfer learning was used. However, the paired t-test gives more insight as just the mean value.

	PZ t-statistic	PZ p-value	PZ Reject H_0
MobileViT	0.42	0.7037	fail to reject
MobileViT-V2	1.16	0.3313	fail to reject
UPerNet	-0.15	0.8926	fail to reject

Fig 12. Table 3 Hypothesis testing report for the pz class.

	TZ t-statistic	TZ p-value	TZ Reject H_0
MobileViT	0.49	0.6605	fail to reject
MobileViT-V2	-0.79	0.4851	fail to reject
UPerNet	0.62	0.5772	fail to reject

Fig 13. Table 4 Hypothesis testing report for the tz class.

	t-statistic	p-value	Reject H_0
MobileViT	0.95	0.4435	fail to reject
MobileViT-V2	1.94	0.1926	fail to reject
UPerNet	-1.90	0.1975	fail to reject

Fig 14. Table 5 Hypothesis testing report for the mean dice coefficient.

DISCUSSION

The architectures tested in this study correspond to the smallest ones for each of their types, this was done to meet the requirements of time and hardware needed for this study. On the other hand, as was explained, many transformations for data augmentation were used to combat the problem of a small training dataset by adding more diversity to it, therefore, it is possible, that it is more effective to have Data Augmentation in small architectures than to use transfer learning. Similar studies such as that of Ferguson, et al., DefectsMenufacturing on detection and segmentation of defects in manufacturing with CNNs showed that the result of using transfer learning was effective, improving the final precision of the models from 0.874 to 0.957. But RCNN is a fairly big architecture.

Another study by Zhang, et al.BreastMRI showed that in medical MRI imaging with convolutional networks 2 and 3D (U-NET \& nnU-NET) showed that the result of using transfer learning was effective, on segmenting the lungs and heart, muscles and bones, solid tissues with cancer, and on skin and fat by running a Wilcoxon test on the dice similarity coefficient, jaccard coefficient and the Hausdorff distance having p-values <0.05. This may be the case because there are multiple and different segmentation tasks to be tested that have more intrinsic variability, while in this study we only tested for a monotone task of segmentation of prostate regions, and the variability needed for this task may have been entirely addressed by the data augmentation.

CONCLUSION

Magnetic resonance imaging (MRI) is a useful technology used for medical diagnosis and treatment. It is widely used on segmenting prostate regions to help healthcare professionals make better diagnoses and safer procedures. In this context, many artificial intelligence models like convolutional neural networks have proven to be very effective, however, the architecture of Vision Transformers that ported the attention mechanism has allowed them to understand context and let them be able to perform vision alongside with natural language tasks.

Due to the nature of ViT, large datasets are required for their training, however, when large amounts of data are not available, like on medical imaging, Transfer learning can be an effective training strategy, however it may not be very effective for smaller and optimized architectures. To test this, we selected the smallest versions of MobileViT, MobileViT-V2 and UPerNet with the Swin transformer Backbone architectures from the hugging face ``transformer" library and trained them with and without transfer learning, using standard data-augmentation techniques.

The Dice coefficients per prostate class were calculated on the validation dataset with the best performing trained weighs, the mean values are reported on table \ref{table:ValDice. To test if the dice coefficients have improved by using transfer learning, paired t-tests were conducted per model, resulting on p-values of 0.4435, 0.1926 an 0.1975 for MobileViT, MobileViT-V2 and UPerNet respectively, and by using a standard confidence value of 0.05 we failed to reject the null hypothesis, concluding that transfer learning was not effective. This may imply that for small models like those tested, the variability of the training dataset is more important than transfer learning, but further research should be done to test this new hypothesis.

ACKNOWLEGMENT

I first thank God for giving me the opportunity to study, the strength and wisdom to face all adversity and the graces to sustain me in my worst moments. To my mother and father for always being by my side giving me their unconditional support. To my aunt Miriam Pavón who has always been like my second mother. To my mentor Mauricio Barros for giving me his friendship and knowledge. To my sister for always making me see the positive side of things. To my friends Ismael Ramos and Vanessa Castro for their advises and unconditional friendship and always be willing to help me when I needed it. And to all the teachers and colleagues who have pushed me to be better every day.

REFERENCES

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision, 127*(3), 302–321. Springer.

Jing, L., & Tian, Y. (2019). Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. arXiv preprint arXiv:1902.06162.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. Recuperado de http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

Chen, Y., Zheng, B., Zhang, Z., Wang, Q., Shen, C., & Zhang, Q. (2020). Deep Learning on Mobile and Embedded Devices: State-of-the-Art, Challenges and Future Directions. *ACM Computing Surveys, 53*. https://doi.org/10.1145/3398209

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. arXiv preprint arXiv:1409.0575.

Albumentations Team. (2024). Semantic Segmentation on the Pascal VOC Dataset - AutoAlbument. Recuperado el 10 de (Mayo Clinic, 2021) de 2024, de https://albumentations.ai/docs/autoalbument/examples/pascal_voc/

Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. Recuperado de https://arxiv.org/abs/2110.02178

Karasulu, B., & Korukoglu, S. (2011). A simulated annealing-based optimal threshold determining method in edge-based segmentation of grayscale images. *Applied Soft Computing, 11*(2), 2246–2259. Elsevier.

Pujar, J. H., Gurjal, P. S., Kunnur, K. S., et al. (2010). Medical image segmentation based on vigorous smoothing and edge detection ideology. *International Journal of Electrical and Computer Engineering, 4*(8), 1143–1149. Citeseer.

Lee, J. J., Thomas, I. C., Nolley, R., Ferrari, M., et al. (2014). Biologic differences between peripheral and transition zone prostate cancer. *Prostate*. https://doi.org/10.1002/pros.22903. Epub ahead of print. PMID: 25327466.

Beaton, L., Bandula, S., Gaze, M. N., & Sharma, R. A. (2019). How rapid advances in imaging are defining the future of precision radiation oncology. *British Journal of Cancer, 120*(8), 779–790. https://doi.org/10.1038/s41416-019-0412-y. Epub 2019 Mar 26. PMID: 30911090. PMCID: PMC6474267.

Grégoire, V., Guckenberger, M., Haustermans, K., Lagendijk, J. J. W., Ménard, C., Pötter, R., Slotman, B. J., Tanderup, K., Thorwarth, D., van Herk, M., & Zips, D.

(2020). Image guidance in radiation therapy for better cure of cancer. *Molecular Oncology*, *14*(7), 1470–1491. https://doi.org/10.1002/1878-0261.12751. Epub 2020 Jun 29. PMID: 32536001. PMCID: PMC7332209.

American Cancer Society. (2024). Survival Rates for Prostate Cancer. Recuperado de https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/survival-rates.html

American Cancer Society. (2024). Key Statistics for Prostate Cancer. Recuperado de https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html

Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., & Khanal, B. (2023). Exploring Transfer Learning in Medical Image Segmentation using Vision-Language Models. arXiv preprint arXiv:2308.07706.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., van Ginneken, B., … Cardoso, M. J. (2022). The Medical Segmentation Decathlon. *Nature Communications, 13*(1). http://dx.doi.org/10.1038/s41467-022-30695-9

Brainder. (2012). The NIfTI file format. Recuperado de https://brainder.org/2012/09/23/the-nifti-file-format/

Chawla, A. (2023). Transfer Learning vs. Fine-tuning vs. Multitask Learning vs. Federated Learning. *Daily Dose of Data Science*. Recuperado de https://www.blog.dailydoseofds.com/p/transfer-learning-vs-fine-tuning

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

Ferguson, M., Ak, R., Lee, Y.-T. T., & Law, K. H. (2018). Detection and Segmentation of Manufacturing Defects with Convolutional Neural Networks and Transfer Learning. arXiv preprint arXiv:1808.02518.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv preprint arXiv:2103.14030.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524.

Ham, S., Kim, M., Lee, S., et al. (2023). Improvement of semantic segmentation through transfer learning of multi-class regions with convolutional neural networks on supine and prone breast MRI images. *Sci Rep, 13*, 6877. https://doi.org/10.1038/s41598-023-33900-x

La Serna (Palomino & Concha, 2009), N., & Román Concha, U. N. (2009). Técnicas de Segmentación en Procesamiento Digital de Imágenes. *Rev. Investig. Sist. Inform., 6*(2), 9-16. Recuperado el 09 de (Mayo Clinic, 2021) de 2024, de https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/3299

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). Mask R-CNN. arXiv preprint arXiv:1703.06870.

Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., Chen, K., Liu, Z., & Loy, C. C. (2023). Transformer-Based Visual Segmentation: A Survey. arXiv preprint arXiv:2304.09854.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. arXiv preprint arXiv:1411.4038.

Lorenzo Martín, A., et al. (2023). Análisis del estudio de la migraña con resonancia magnética mediante medidas avanzadas y técnicas de inteligencia artificial.

Matallana Camacho, J. P., & Higuera Vázquez, Z. L. (2022-11-16). Pos procesado de imágenes diagnósticas. Recuperado de https://repository.unad.edu.co/handle/10596/58075

(Mayo Clinic, 2021) Clinic. (2021). Demencia vascular: Diagnóstico y tratamiento. Recuperado de https://www.(Mayo Clinic, 2021)clinic.org/es/diseases-conditions/vascular-dementia/diagnosis-treatment/drc-20378798

Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. arXiv preprint arXiv:2110.02178.

Mehta, S., & Rastegari, M. (2022). Separable Self-attention for Mobile Vision Transformers. arXiv preprint arXiv:2206.02680.

(Ortiz-Hidalgo & Heredia-Jara, 2022).-Hidalgo, C., & Heredia-Jara, A. (2022). Histología normal de la próstata con algunas implicaciones clínicas. *Revista Latinoamericana de Patología, 59*.

Pancholi, K., Modi, S., & Chitaliya, G. (2023). A novel mul threshold algorithm for segmentation of the MRI images. *Salud, Ciencia Y Tecnología, 3*, 408. https://doi.org/10.56294/saludcyt2023408

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision Transformers for Dense Prediction. arXiv preprint arXiv:2103.13413.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv preprint arXiv:1505.04597.

Sampath, K., Rajagopal, S., & Chintanpalli, A. (2024). A comparative analysis of CNN-based deep learning architectures for early diagnosis of bone cancer using CT

images. *Sci Rep, 14*(1), 2144. https://doi.org/10.1038/s41598-024-52719-8. PMID: 38273131. PMCID: PMC10811327.

Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., & Chavez-Urbiola, E. A. (2023). Loss Functions and Metrics in Deep Learning. arXiv preprint arXiv:2307.02694.

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are Convolutional Neural Networks or Transformers more like human vision? arXiv preprint arXiv:2105.07197.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., & Herath, D. (2023). Semantic Segmentation using Vision Transformers: A survey. arXiv preprint arXiv:2305.03273.

Universidad Internacional de La Rioja. (2023). Transfer learning: qué es y cómo se aplica en Machine Learning. *Revista UNIR*. Recuperado de https://www.unir.net/ingenieria/revista/transfer-learning/

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv preprint arXiv:2105.15203.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified Perceptual Parsing for Scene Understanding. arXiv preprint arXiv:1807.10221.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A Comprehensive Survey on Transfer Learning. arXiv preprint arXiv:1911.02685.

Zhang, Y., Mehta, S., & Caspi, A. (2023). Rethinking Semantic Segmentation Evaluation for Explainability and Model Selection. arXiv preprint arXiv:2101.08418.