People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Echahid Hamma Lkhdar University - El oued
Faculty of Exact Sciences
Computer Science Department

Graduation project Report

2nd year academic Master

option : *artificial intelligence systems*

TITLE

Towards transformers application in computer vision

Presented by :
-Anis Hannanou
-Ali Soltani

Supervised by :
-Dr. KHOLLADI Nedjoua Houda

Defended on June 7 ,2023 : In front of the jury composed of :

-Mrs. Sana Sahar Guia	President
-Mr. Faouzi Zaiz	Examiner

Academic year : **2022 – 2023**

Remerciements

It is with great pleasure that we reserve these few lines as a sign of gratitude and deep appreciation to all those who, from near or far, have contributed to the realization and the outcome of this work.

I thank Almighty God who has always been with us and helped us throughout our journeys.

Special thanks to our dear parents

Our dear father Ahmed Hannanou and el Mouldi Soltani

For all the sacrifices made for my education. For all the love you gave me, for your support and your affection that you never stopped lavishing on me. Your unique values have shown me the right path. Your advice, your presence and your generous heart have illuminated my life ; I owe you everything my dear dad. May this modest work be a testimony of my gratitude, my respect, my admiration and above all my love for you

Our dear mother Djamila Berkat and nacira Hammana

For your immense sacrifices, for your limitless patience, for all the love you gave me and for your limitless kindness ; you are the most wonderful mother. Let this work prove that we can do it when we have a mom like you. Thank you for what you are. May God keep you healthy and grant you a long life full of happiness.

I would like to thank all my teachers for their support, their teaching and their advice throughout my school career.

acknowledgements

To our supervisor Dr kholladi-nedjouahouda

We would like to send you our warmest thanks for agreeing to supervise and follow this work, for your valuable advice, for the time you have granted us and for the kindness and patience that you have shown to our forecast.

To the members of the jury

Our deepest thanks go to the members of the jury for the honor they do us by judging this work.

To all professors and teachers of university Echahid Hamma Lkhdar EL oued

We would like to thank all the teachers at Echahid Hamma Lkhdar El Oued University for the training they provided and their encouragement.

To all our friends

For their support and encouragement.

We hope that this project will meet your expectations and that it will enrich the library of our university.

Abstract

Transformers have dominated the field of natural language processing and have recently made an impact in the area of computer vision. The field of medical image analysis has been particularly interested in leveraging the advancements made by Transformers, as opposed to the traditional Convolutional Neural Networks (CNNs). Transformers have proven to be effective in various medical image processing applications, including classification, registration, segmentation, detection, and diagnosis. The purpose of this memoir is to raise awareness about the potential applications of Transformers in medical image processing. we provide firstly an overview of the fundamental concepts of artificial intelligence and its relevance to computer vision, with a specific focus on how Transformers and other essential components contribute to these advancements. Second, we conduct a comprehensive review of different Transformer architectures tailored for medical image applications. We explore their specific applications and discuss the challenges associated with using visual Transformers in this domain. Within this dissertation we delve into the significant differences between CNNs and Transformers, with emphasising the proposed classification model enhancement image (brain MRI) by comparing the results with CNN model.

Key Words

Artificial Intelligence - Computer Vision - Convolutional Neural Networks - Vision Transformers

Résumé

Transformers have dominated the field of natural language processing and have recently made an impact in the area of computer vision. The field of medical image analysis has been particularly interested in leveraging the advancements made by Transformers, as opposed to the traditional Convolutional Neural Networks (CNNs). Transformers have proven to be effective in various medical image processing applications, including classification, registration, segmentation, detection, and diagnosis. The purpose of this memoir is to raise awareness about the potential applications of Transformers in medical image processing. we provide firstly an overview of the fundamental concepts of artificial intelligence and its relevance to computer vision, with a specific focus on how Transformers and other essential components contribute to these advancements. Second, we conduct a comprehensive review of different Transformer architectures tailored for medical image applications. We explore their specific applications and discuss the challenges associated with using visual Transformers in this domain. Within this dissertation we delve into the significant differences between CNNs and Transformers, with emphasising the proposed classification model enhancement image (brain MRI) by comparing the results with CNN model.

Mots Clé

Intelligence Artificielle - Vision ordinateur - Réseaux de Neurones Convolutifs - Transformateurs de Vision

Table des matières

1 Background **1**

1.1 Introduction . . . 1

1.2 Artificial Intelligence . . . 1

1.2.1 The Most Important Trends of Artificial Intelligence . . . 3

1.2.1.1 Machine Learning : . . . 3

1.2.1.2 Types of Machin Learning . . . 3

1.2.2 Recent Advances and Future Perspectives for Artificial Intelligence . . . 4

1.3 Computer Vision . . . 5

1.4 Medical Image Processing . . . 5

1.4.1 Medical Image Processing Aim . . . 6

1.4.1.1 Computer Aided Diagnosis . . . 6

1.4.1.2 Computer-Aided Detection . . . 6

1.5 Dataset . . . 7

1.5.1 Definition . . . 7

1.5.2 Characteristics of Dataset . . . 7

1.5.3 Types of Datasets . . . 7

1.5.4 Medical Image Datasets . . . 8

1.6 Problematic and Motivations . . . 8

1.6.1 Problem of Computer Vision . . . 8

1.7 Conclusion . . . 10

2 Vision Transformers ViT **11**

2.1 Introduction . . . 11

2.2 Visual Transformers ViT Overview . . . 12

2.3 General Architecture of Transformers (Compoments) . . . 12

2.3.1 Encoder and Decoder Stacks . . . 15

2.3.2 Attention . . . 15

2.3.2.1 Scaled Dot-Product Attention . . . 15

2.3.2.2 Multi-Head Attention . . . 16

2.3.2.3 Applications of Attention in our Model . . . 17

2.3.2.4 Position-wise Feed-Forward Networks . . . 17

2.3.2.5 Embeddings and Softmax . . . 17

2.3.2.6 Positional Encoding . . . 17

2.4 Transformer application . . . 18
2.4.1 Image Classification . . . 18
2.4.1.1 Image Classification Methode . . . 18
2.4.1.2 Image Classification Aim . . . 19
2.4.2 Anomaly Detection . . . 19
2.4.3 Object Detection . . . 19
2.4.3.1 Transformers Application for Object Detection . . . 19
2.4.3.2 Object Detection Aim . . . 19
2.4.4 Image Segmentation . . . 19
2.4.4.1 Image Segmentation Methode . . . 20
2.4.4.2 Image Segmentation Aim . . . 20
2.4.5 Image Compresion . . . 20
2.4.6 Videos Deepfake Detection . . . 20
2.4.6.1 Problem of Videos Deepfake Detection . . . 20
2.4.6.2 Solution . . . 20
2.4.7 Clusters Analysis . . . 21
2.4.8 Cluster Analysis Applications . . . 21
2.5 Challenges . . . 21
2.6 Visual Transformers Models . . . 22
2.6.1 General Visual Transformer Model Pipeline . . . 22
2.6.2 Transformer Backbone . . . 23
2.6.3 Transformer Neck . . . 26
2.7 Visual Transformers Models . . . 27
2.8 Conclusion . . . 29

3 Related Works **30**
3.1 Introduction . . . 30
3.2 Binary Classification of MRI Brain Tumors Using CNN Models . . . 31
3.3 Transformers for Medical Image Segmentation Tasks . . . 32
3.4 Lesion Classification of Brain MRI Tumors Using Transformers (ViT) . . . 33
3.5 Conclusion . . . 35

4 Proposed Methodology **36**
4.1 Introduction . . . 36
4.2 Proposed Pipeline . . . 37
4.3 Dataset . . . 37
4.3.1 Dataset Processing and Splitting . . . 37
4.4 Evaluation Metrics . . . 38
4.5 Brain Tumor Detection Using CNN and ViT . . . 38
4.5.1 Architecture CNN . . . 38
4.5.1.1 General Overview of a CNN Architecture for Image Classification . . . 40

4.5.1.2 Visualization of CNN Model . . . 40
4.5.2 ViT Architecture . . . 41
4.5.3 Visualization of ViT Model . . . 42
4.5.4 Models . . . 43
4.5.5 Fine-tuning Details . . . 44
4.6 Traind Model . . . 44
4.6.1 CNN Model Classification . . . 46
4.6.2 Visual Transformer Model Classification . . . 46
4.7 Results (Validation Model) . . . 48
4.7.1 Convolutional Neural Network (CNN) Model . . . 48
4.7.2 Visual Transformer (ViT) Model . . . 48
4.7.3 Models Evaluation . . . 48
4.7.3.1 CNN Model . . . 48
4.7.3.2 ViT Model . . . 49
4.8 Discussions . . . 50
4.9 The Beginning of our Search . . . 50
4.10 Conclusion . . . 51

List of Figures

1.1 Machine Learning Deep Learning[1] 2
1.2 Machine Learning in the context of AI[1] 3

2.1 The development of transformers in medical image analysis. Selected methods are displayed relating to classification, detection, segmentation, and synthesis applications[2]. 12
2.2 The Transformer - model architecture[3]. 14
2.3 (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel[3]. 16
2.4 Applications of transformers in medical image analysis[2]. 18
2.5 Taxonomy of Visual Transformer[4] 23

4.1 Proposed pipeline. 37
4.2 Example of brain tumor images with different classifications. 38
4.3 An simple CNN architecture 39
4.4 Illustration of the treat of the images in a CNN model 40
4.5 Overview of the vision Transformer used in classification 41
4.6 Illustration ot the to treat of the images in a ViT model 42
4.7 CNN model classification 45
4.8 CNN model classification 47
4.9 CNN model performance evaluation schema 49
4.10 ViT model performance evaluation schema 49

List of Tables

3.1 The main studies working on Brain MRI classification using CNN models 31
3.2 The main studies working using ViT models : 32
3.3 Transformers used in medical image classification tasks 34

4.1 Results of CNN models 48
4.2 Results of CNN models 48

General Introduction

Transformers have had a significant impact on computer vision, particularly in medical image processing Recent works have shown that these Transformer modules can fully replace CNN in deep neural networks by operating on a sequence of image patches, giving rise to Vision Transformers (ViTs). Since their inception, ViT models have been shown to push the state-of-the-art in numerous vision tasks, including image classification, object detection,Anomaly detection, Image Segmentation, image colorization Furthermore, recent research indicates that the prediction errors of ViTs are more consistent with those of humans than CNNs. These desirable properties of ViTs have sparked great interest in the medical community to adapt them for medical imaging applications, thereby mitigating the inherent inductive biases of CNNs.

Vision Transformers (ViTs) have revolutionized computer vision, especially in the field of medical image processing. Groundbreaking studies have demonstrated that Transformer modules can completely replace Convolutional Neural Networks (CNNs) in deep neural networks by processing a sequence of image patches. This breakthrough has given birth to ViTs, which have showcased remarkable advancements in various vision tasks, including image classification, object detection, anomaly detection, image segmentation, and image colorization. Moreover, recent research suggests that ViTs exhibit prediction errors that align more closely with human perception compared to CNNs. These favorable attributes of ViTs have generated substantial interest within the medical community, as they offer a promising solution to address the inherent biases of CNNs when applied to medical imaging applications.

Recently, medical imaging community has witnessed an exponential growth in the number of Transformer based techniques, especially after the inception of ViTs. in this dissertation we employed with pre-trained CNN networks for the classification task, specifically using brain tumor datasets. We then compared the results obtained from the CNN-based architecture with those from the ViT-based architecture in which We perform a tow-classes classification task of detection tumor (with tumor and no tumor). We use datasets to 256 images medical for MRI. All experiments are done on the fixed training and testing sets from the dataset for comparison purposes between images who Contains tumors or not, then we highlight CNN & ViT models by calculaiting accuracy loss function for (training and validation) task to the tested models. Furthemore, after setting both models (CNN and ViT) and training them on the same dataset, we obtain ViT results performing ont much better than CNN on various metric standards (ACC, LOSS).

This dissertation is devided on four chapter :

Chapter one : an overview of the fundamental concepts of artificial intelligence and its relevance to computer vision, with a specific focus on how Transformers and other essential components contribute to these advancements.

Chapter two : we conduct a comprehensive review of different Transformer architectures tailored for medical image applications. We explore their specific applications and discuss the challenges associated with using visual

Transformers in this domain.

Chapter three : We discussed the various works related to our dissertation and the various applications of converters to many fields in image processing.

Chapter four : we set two models (CNN and VIT) and train them on the same set of dataset then we provide the results of each model and we compare both models in terms of accuracy and loss.

Chapitre 1

Background

1.1 Introduction

Transformers have revolutionized Natural Language Understanding (NLU), a branch of Natural Language Processing (NLP) that has become a cornerstone of artificial intelligence within the worldwide digital economy.

Transformer models mark the beginning of a new era in artificial intelligence in general and especially in computer vision. Language under- standing has become the pillar of language modeling, chatbots, personal assistants, question answering, text summarizing, speech-to-text, sentiment analysis, machine translation, and more

The significant advancements achieved by Transformer networks in the realm of Natural Language Processing (NLP) have generated considerable interest among computer vision experts to apply these models to vision and multi-modal learning tasks. However, visual data adheres to a specific structure (e.g., spatial and temporal coherence), thereby necessitating the development of innovative network architectures and training methods. Consequently, Transformer models and their variations have been effectively utilized for a variety of applications, such as image recognition, object detection, segmentation, image super-resolution, video comprehension, image generation, text-image synthesis, visual question answering, and many others.

The architecture of Transformers is based on a self-attention mechanism that can capture the relationships between the elements of a sequence. Unlike recurrent networks that process sequence elements recursively and can only focus on short-term context, Transformers can attend to the complete sequence, enabling them to learn long-range relationships. While attention models have been widely used in both feed-forward and recurrent networks, Transformers rely entirely on the attention mechanism and have a unique implementation (multi-head attention) that is optimized for parallelization. These models have the crucial feature of being scalable to high-complexity models and large-scale datasets. In contrast to some of the other alternatives such as hard attention, which is stochastic in nature, Transformers do not rely on prior knowledge of the problem's structure. Instead, they are usually pre-trained on large-scale (unlabeled) datasets using pretext tasks. This pre-training enhances their generalization ability when fine-tuning on downstream tasks.

1.2 Artificial Intelligence

The impact of Artificial Intelligence (AI) on computer vision has been significant, enabling computers to perceive, understand, and interpret visual data with unprecedented efficiency. Computer vision is a subfield of AI that focuses on equipping machines with the ability to process, analyze, and interpret visual information from their surroundings.

AI techniques in computer vision involve the use of deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to facilitate computers in learning and recognizing patterns and features in visual data. These algorithms are trained on large sets of labeled images, allowing them to learn and identify objects, people, and other visual elements with high precision.

The integration of AI and machine learning and deep learning technologies plays a pivotal role in driving computer vision forward. AI enables computer vision systems to comprehend, identify, and analyze various forms of visual data. AI models, algorithms, and logic can quickly process, assimilate, and learn from vast amounts of labeled and unlabeled visual data. This capability empowers computer vision-enabled machines to recognize diverse features, patterns, and relationships in images, videos, graphics, and even infographics.

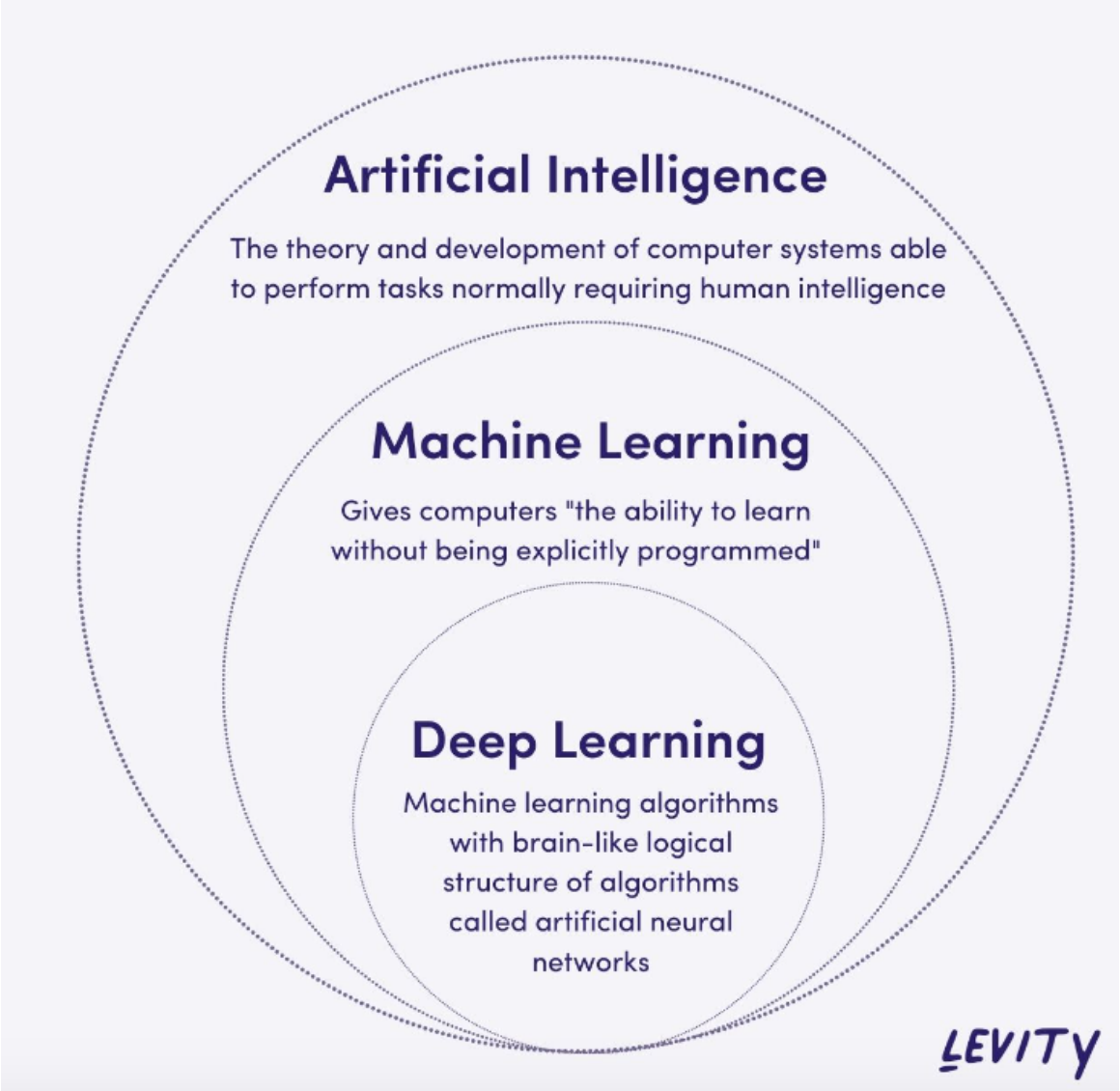

FIGURE 1.1 – Machine Learning Deep Learning[1]

1.2.1 The Most Important Trends of Artificial Intelligence

1.2.1.1 Machine Learning :

Machine Learning, a subfield of Artificial Intelligence (AI), is instrumental in enabling computer vision. It uses algorithms and training data to automatically identify patterns with minimal human intervention. AI, on the other hand, is a method for teaching computers to perform tasks that typically require human-like intelligence. Deep Learning, a subset of Machine Learning, draws inspiration from the structure and function of the human brain, which is represented symbolically by an artificial neural network. Although deep learning was originally proposed in the 1980s, it has gained significant traction in recent years due to two primary reasons : The availability of vast amounts of training data required for building highly complex models. For instance, developing autonomous vehicles requires the collection of a large number of images and videos to train the machine learning models.

The increase in computational power, which allows the training of deep neural networks with a large number of layers, thereby significantly improving their accuracy in recognizing and understanding visual data[5].

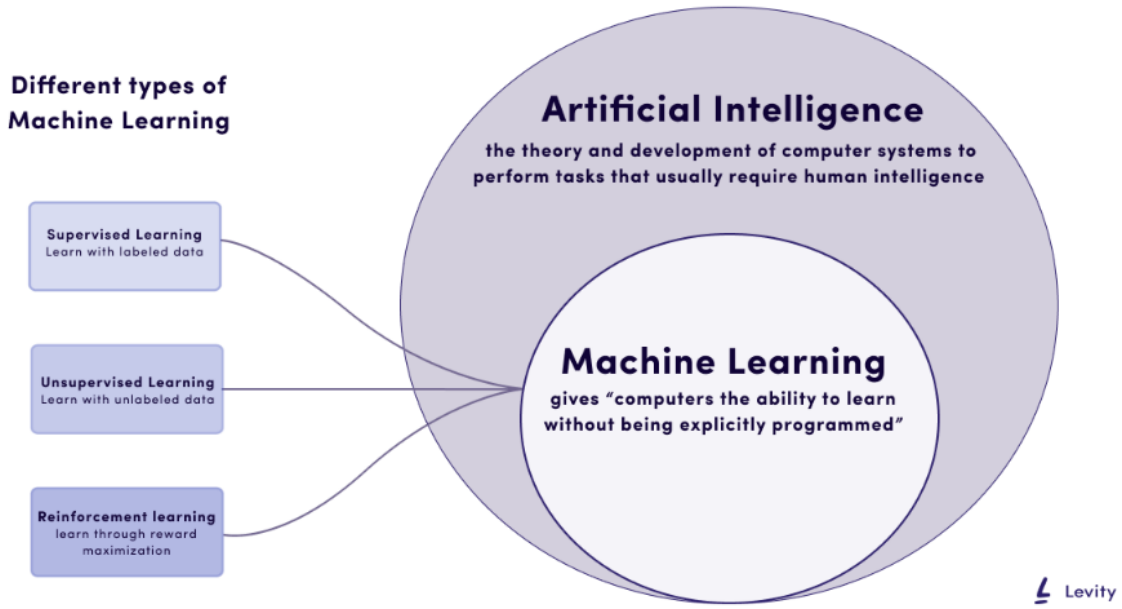

FIGURE 1.2 – Machine Learning in the context of AI[1]

1.2.1.2 Types of Machin Learning

— Supervised Learning

Supervised learning, a subfield of both machine learning and artificial intelligence, involves using labeled datasets to train algorithms to accurately classify data or predict outcomes. In other words, the key characteristic of supervised learning is the use of labeled data for training[6].

— Unsupervised Learning

Unsupervised learning, a branch of machine learning and artificial intelligence, employs algorithms to analyze and cluster unlabeled datasets, discovering latent patterns or data groupings without any human intervention. In contrast to supervised learning, unsupervised learning does not require labeled data for

training[6] .

— Reinforcement Learning

Reinforcement learning is a type of machine learning in which an algorithm, also known as an agent, learns by trial-and-error using feedback from its actions. In this approach, rewards and punishments serve as signals that reinforce desired and discourage undesired behavior[6].

Deep Learning

Deep Learning refers to a class of algorithms that mimic the way humans draw conclusions by analyzing data with a logical structure. This can be achieved through both supervised and unsupervised learning methods. Deep Learning applications typically use a layered structure of algorithms known as an artificial neural network (ANN) to accomplish this. The design of ANNs is inspired by the biological neural network of the human brain, resulting in a learning process that is much more powerful than that of traditional machine learning models[5].

— Fildes of Deep Learning

Deep Learning has a broad range of applications across many fields. In the domain of automated driving, for instance, Deep Learning is used to identify and recognize objects such as STOP signs and pedestrians. The military leverages Deep Learning to detect objects from satellite imagery, allowing them to identify safe or unsafe zones for their troops. Additionally, Deep Learning is ubiquitous in the consumer electronics industry[5]

1.2.2 Recent Advances and Future Perspectives for Artificial Intelligence

The use of artificial intelligence (AI) in image recognition has made significant advancements in recent years, reaching a level of performance that is comparable to that of humans. If AI-based image recognition technology can be effectively applied in various fields, particularly in the medical field, it has the potential to revolutionize the way medical procedures are conducted, such as intraoperative decision-making support, image-guided surgery, and automated surgery. These advancements could lead to a more precise and efficient medical practice, ultimately improving patient outcomes[7].

Also, Artificial intelligence and deep learning have become increasingly important in assisting clinicians with more accurate diagnoses. In this regard, convolutional neural networks (CNNs) and visual transformers (ViTs) are applied to medical images to aid healthcare professionals in tasks such as disease classification, lesion detection, anatomical structure segmentation, automated report generation, image denoising, medical image registration, and more [8].

1.3 Computer Vision

Computer vision relies on optical sensors and algorithms to simulate human vision and extract useful information from visual data. Compared to traditional methods, which are time-consuming and require sophisticated laboratory setups, computer vision has emerged as a subfield of image processing and artificial intelligence.

Computer vision, a field of artificial intelligence (AI), empowers computers and systems to extract meaningful information from digital images, videos, and other visual inputs, enabling them to take actions or make recommendations based on this information. While AI enables computers to think, computer vision empowers them to "see," observe, and understand visual data.

Similar to human vision, computer vision involves training machines to perform functions such as object recognition, depth perception, motion detection, and anomaly detection. However, humans have a head start, as they have a lifetime of context and experience to distinguish objects, estimate distances, detect motion, and identify anomalies in images.

In contrast, computer vision achieves these tasks in a shorter timeframe, utilizing cameras, data, and algorithms instead of human retinas, optic nerves, and visual cortex. The advantage of computer vision is that a system trained to inspect products or monitor production assets can analyze thousands of products or processes per minute, quickly detecting imperceptible defects or issues that may surpass human capabilities[9].

1.4 Medical Image Processing

The most common computer vision tasks in applications that utilise medical images including image classification, medical image segmentation, and shape/object recognition. DL-based methods have demonstrated superiority over traditional methods in various medical applications, such as cancer detection, magnetic resonance imaging (MRI) segmentation, X-ray analysis, and more. These advancements have led to more accurate diagnoses, improved treatment planning, and ultimately better patient outcomes.

Medical image processing techniques leverage computer vision algorithms to analyze and manipulate various types of medical images, such as X-rays, Computerized Tomography (CT) scans, Magnetic resonance imaging (MRI) scans, and ultrasound images. These algorithms enable medical professionals to enhance images, segment regions of interest, classify abnormal regions, and perform measurements and quantitative analysis.

One important application of medical image processing is image segmentation, which involves separating an image into different regions based on their characteristics. This technique can be used to identify tumors, blood vessels, or other structures in medical images. Image segmentation can also be used to detect abnormalities and assist in the planning of surgical procedures[10].

Another application is image registration, which involves aligning multiple images of the same patient taken at different times or from different angles. This technique can be used to track the progression of a disease or to plan surgical procedures.

Overall, medical image processing using computer vision techniques has the potential to significantly improve the accuracy and efficiency of medical diagnoses and treatments, leading to better patient outcomes. With continued advancements in technology, we can expect to see even more powerful and innovative applications of medical image processing in the future[10].

1.4.1 Medical Image Processing Aim

Computer-aided detection (CADe) and computer-aided diagnosis (CADx) are computer-based systems utilized in the medical imaging field to facilitate prompt decision-making by doctors. Given the time-sensitive nature of medical imaging, doctors must quickly evaluate and analyze the information in the images to identify abnormalities. While imaging techniques such as MRI, X-ray, endoscopy, and ultrasound are essential modalities for early disease diagnosis, high-energy imaging can be harmful to the human body. To minimize the risk, images are typically acquired using low energy, resulting in low-quality and low-contrast images. CAD systems are employed to enhance image quality, aiding in the accurate interpretation of medical images by highlighting conspicuous features[11].

1.4.1.1 Computer Aided Diagnosis

— Definition

computer-aided Diagnosis (CAD) technology is a multifaceted system that encompasses artificial intelligence (AI), computer vision, and medical image processing. Its primary purpose is to identify abnormalities within the human body.

— Objectifes

CAD aims to improve disease detection by minimizing false negatives resulting from human observation oversights. Unlike involving a second human observer, CAD's computer-based approach doesn't add to the demands on the radiologist or trained observer pool. Regardless of the approach, the ultimate goal is to enhance disease detection without significantly increasing recall and work-up rates. Furthermore, in certain scenarios, CAD's associated automated software tools offer potential workflow efficiencies, though this aspect lies outside the scope of this overview.

CAD algorithms are developed to search for the same features that a radiologist looks for during case review[12].

— Applications of CAD System

CAD is applied to diagnose various medical conditions such as breast cancer, lung cancer, colon cancer, prostate cancer, bone metastases, coronary artery disease, congenital heart defects, pathological brain detection, Alzheimer's disease, and diabetic retinopathy[10]..

1.4.1.2 Computer-Aided Detection

— Definition CADx systems perform the characterization of the lesions, for example, the distinction between benign and malignant tumors.

— Objectifes

CADx systems extract the characteristics of the images and use a classifier to measure the malignancy[13]. The objectives of CADx (computer-aided diagnosis) include improving the accuracy and consistency of medical diagnoses, reducing the risk of errors, aiding in early disease detection, and enhancing the efficiency of the diagnostic process. Additionally, CADx aims to minimize the subjectivity in interpretation by providing quantitative and objective analysis of medical images

1.5 Dataset

1.5.1 Definition

In the field of computer vision, a dataset is a collection of visual data, such as images or videos, that are used to train and test machine learning models and algorithms. The dataset can include various types of visual data, such as natural scenes, objects, faces, medical images, etc[14].

Datasets are vital for the development and evaluation of machine learning models in computer vision. They provide a standardized and diverse set of visual data to assess the accuracy and robustness of different models or algorithms[14].

1.5.2 Characteristics of Dataset

Size : A large dataset is necessary to train machine learning models that can generalize well to new data. Diversity : A diverse dataset should contain images or videos that represent various types of objects, scenes, or situations. Quality : A high-quality dataset should have accurate and consistent annotations or labels, free of errors and biases. Balance : A balanced dataset should have approximately equal numbers of images or videos for each class or category. There are various publicly available datasets in computer vision, such as ImageNet, COCO, CIFAR, and PASCAL VOC, which are commonly used by researchers and practitioners to develop machine learning models for different applications, including object detection, image segmentation, facial recognition, and autonomous driving.

1.5.3 Types of Datasets

There are various categories of datasets accessible for varying types of information. These include

— Numerical data sets

A dataset that comprises data expressed in numerical form instead of language is known as a numerical dataset. Quantitative data is another term for numerical data. The complete collection of quantitative/numerical data is referred to as the numerical dataset. Numerical data is always represented in numeric values, allowing us to perform mathematical operations on it[14].

— Bivariate Datasets

When a dataset involves two distinct variables and explores the relationship between them, it is known as a bivariate dataset[14].

— Multivariate Datasets

A dataset that involves multiple variables is referred to as a multivariate dataset. This means that the dataset includes three or more distinct data types (variables), and each measurement is obtained as a function of these variables. Put differently, a multivariate dataset is composed of individual data points that are determined by three or more variables[14].

— Categorical Datasets

Categorical datasets illustrate the attributes or traits of an individual or object. Such datasets are composed of a categorical variable, also known as a qualitative variable, which can assume only two values. Therefore, it is referred to as a dichotomous variable. If a categorical variable can take on more than two possible values, it is termed a polytomous variable. By default, qualitative/categorical variables are typically assumed to be polytomous unless explicitly stated otherwise[14].

Example

A person's gender (male or female)

Marital status (married/unmarried)

— Correlation Datasets

Correlation datasets are collections of values that display some form of interrelationship or association with each other. Such datasets typically consist of values that are mutually dependent.

Statistically speaking, correlation refers to the relationship between two variables/entities. In some cases, it may be necessary to predict the correlation between different items, making it important to have a solid understanding of how correlation operates. Correlation is commonly classified into three distinct btypesv [14].

1.5.4 Medical Image Datasets

Almost 90 percent of all healthcare data is comprised of images, creating numerous opportunities to develop computer vision algorithms for healthcare applications. It's important to mention that medical image data is primarily produced in radiology departments using imaging technologies such as X-Ray, CT, and MRI scans. The standard format for storing and sharing diagnostic images in the healthcare industry is DICOM (Digital Imaging and Communication in Medicine).

1.6 Problematic and Motivations

The emergence of deep learning has revolutionized the fields of computer vision and natural language processing, and has become a crucial component of artificial intelligence (AI). This has also opened up new possibilities for developing intelligent healthcare systems. However, while the healthcare industry is rapidly accumulating vast amounts of imaging data, deep learning algorithms for automated medical image analysis still face significant challenges.

One major obstacle is the need for large amounts of labeled data to train deep learning models, which is expensive and requires specialized expertise. Additionally, medical images differ from natural images in various ways, such as being in 3D formats and having multiple phases or modalities. Moreover, the incorporation of prior medical knowledge is often overlooked in the development of deep learning algorithms. In healthcare, the consequences of errors can be costly, and thus, there is a need to establish robust AI systems for medical image analysis[15].

Despite these challenges, deep learning has already shown great potential in various aspects of the clinical workflow, including disease screening, malignancy diagnosis, prognosis prediction, and pathology. With continued advancements in technology and further research, we can expect to see even more promising applications of deep learning in healthcare.

1.6.1 Problem of Computer Vision

computer vision, like any other technology, has its own set of problems. Some of the common challenges in computer vision are :

— Ambiguity

Visual data can be highly ambiguous, and different objects can appear similar to each other.

— Variability

  Visual data can vary greatly in terms of lighting, color, texture, and other factors, making it challenging for computer vision systems to accurately recognize objects in different contexts.

— Occlusion

  Objects in real-world scenarios can be partially or fully occluded by other objects or obstacles, making it difficult for computer vision systems to detect and recognize them.

— Scale

  Objects in visual data can vary in size, and it can be challenging for computer vision systems to recognize objects at different scales.

— Limited training data

  Deep learning models used in computer vision require large amounts of labeled training data, and it can be challenging to obtain sufficient amounts of high-quality labeled data.

Recently, a new technology called visual Transformer (ViT) has emerged as a promising alternative to Convolutional Neural Networks (CNNs) in image recognition computer vision tasks. ViT models have shown to outperform the current state-of-the-art CNNs by almost four times in terms of computational efficiency and accuracy.

Unlike CNNs, which use pixel arrays as input, ViT splits the input image into visual tokens. It divides the image into fixed-size patches, correctly embeds each patch, and includes positional embedding as an input to the transformer encoder. The self-attention layer in ViT allows the model to embed information globally across the entire image, and to learn the relative location of the image patches to reconstruct the structure of the image.

While transformers have been widely used in Natural Language Processing (NLP) and have achieved high success rates, ViT models are now being applied to image recognition tasks with promising results. With its superior computational efficiency and accuracy, ViT has the potential to improve the performance of automated medical image analysis and other computer vision applications[16].

1.7 Conclusion

Despite the ground-breaking progress in artificial intelligent, how to endow deep learning and machine learning models with reasoning ability remains a formidable challenge for modern computer vision systems. In this regard, and this chapter We touched on artificial intelligence and the latest advances in it, which enabled the development of artificial intelligence tasks in computer vision so that The use of Transformers is revolutionizing computer vision, particularly in the field of medical image analysis where research in this area is rapidly expanding. However, current Transformer-based approaches are typically applied to medical imaging problems without significant modifications.

Chapitre 2

Vision Transformers ViT

2.1 Introduction

Deep neural networks (DNNs) have become a crucial element of modern AI systems, with different types of networks tailored to specific tasks. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been extensively used for processing images and sequential data, respectively. However, transformer, a new type of neural network, has emerged as a promising alternative that mainly utilizes the self-attention mechanism to extract intrinsic features, making it suitable for various AI applications. Transformer first achieved significant improvements in natural language processing (NLP) tasks, where it exhibited strong performance on downstream tasks without requiring fine-tuning. This breakthrough led researchers to apply transformer to computer vision (CV) tasks, which traditionally relied on CNNs. Vision transformers, such as ViT, apply pure transformers to image patch sequences and have shown state-of-the-art performance on multiple image recognition benchmarks. Besides image classification, transformer has also been utilized to solve other Computer Vision problems, such as object detection, semantic segmentation, image processing, and video understanding. As a result of its exceptional performance, transformer-based models are increasingly being proposed to enhance various visual tasks[17].

2.2 Visual Transformers ViT Overview

Visual Transformers (ViT), is a novel model used for image processing that utilizes a Transformer-based architecture on image patches. This process involves dividing an image into patches of equal sizes, embedding each patch linearly, adding position embeddings, and then feeding the sequence of resulting vectors to a standard Transformer encoder. To classify the image, a learnable "classification token" is appended to the sequence following the standard approach.

In recent times, transformers have been applied to medical image analysis for tasks such as disease diagnosis and other clinical purposes. A study used transformers to distinguish COVID-19 from other types of pneumonia using CT or X-ray images and other, providing a quick and effective solution to treat COVID-19 patients. Transformers have also achieved state-of-the-art results in medical image segmentation, detection, and synthesis, as depicted in Figure 2.1 These findings demonstrate the potential of transformers in advancing medical image analysis[2].

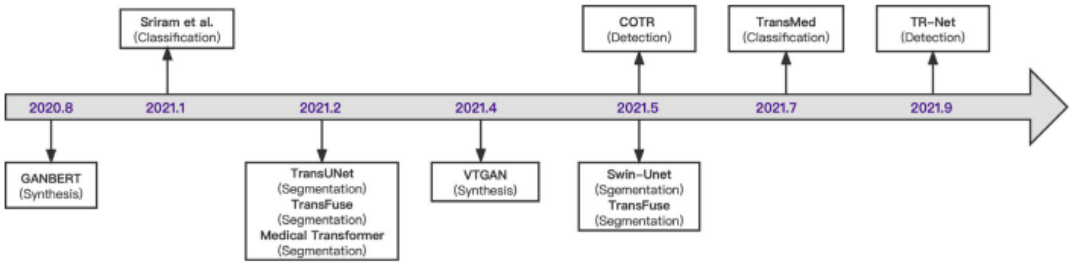

FIGURE 2.1 – The development of transformers in medical image analysis. Selected methods are displayed relating to classification, detection, segmentation, and synthesis applications[2].

And Compared to Convolutional Neural Networks (CNN), Vision Transformer (ViT) has demonstrated impressive results while using significantly fewer computational resources for pre-training. However, due to its generally weaker inductive bias, ViT tends to rely more heavily on regularization or data augmentation when training on smaller datasets than CNN. Originally designed for text-based tasks, ViT is a visual model based on transformer architecture. It represents an input image as a series of image patches, similar to word embeddings used for text in transformers, and directly predicts class labels for the image. With sufficient training data, ViT has shown exceptional performance, outperforming a comparable state-of-the-art CNN with only one-fourth of the computational resources.

These transformers have high success rates when it comes to NLP models and are now also applied to images for image recognition tasks. CNN use pixel arrays, whereas ViT splits the input images into visual tokens. The visual transformer divides an image into fixed-size patches, correctly embeds each of them, and includes positional embedding as an input to the transformer encoder. Moreover, ViT models outperform CNNs by almost four times when it comes to computational efficiency and accuracy[18].

2.3 General Architecture of Transformers (Compoments)

Typically, neural models for transducing sequences have an encoder-decoder structure, where the input sequence of symbols (x1, ..., xn) is first mapped by the encoder to a sequence of continuous representations (z1, ..., zn). Then, given z, the decoder generates the output sequence (y1, ..., ym) one symbol at a time, with the model being auto-regressive, consuming the previously generated symbols as additional input for generating the

next. The Transformer architecture follows this overall structure, with stacked self-attention and point-wise, fully connected layers used for both the encoder and decoder, illustrated in the left and right halves of Figure 2.2, respectively [3].

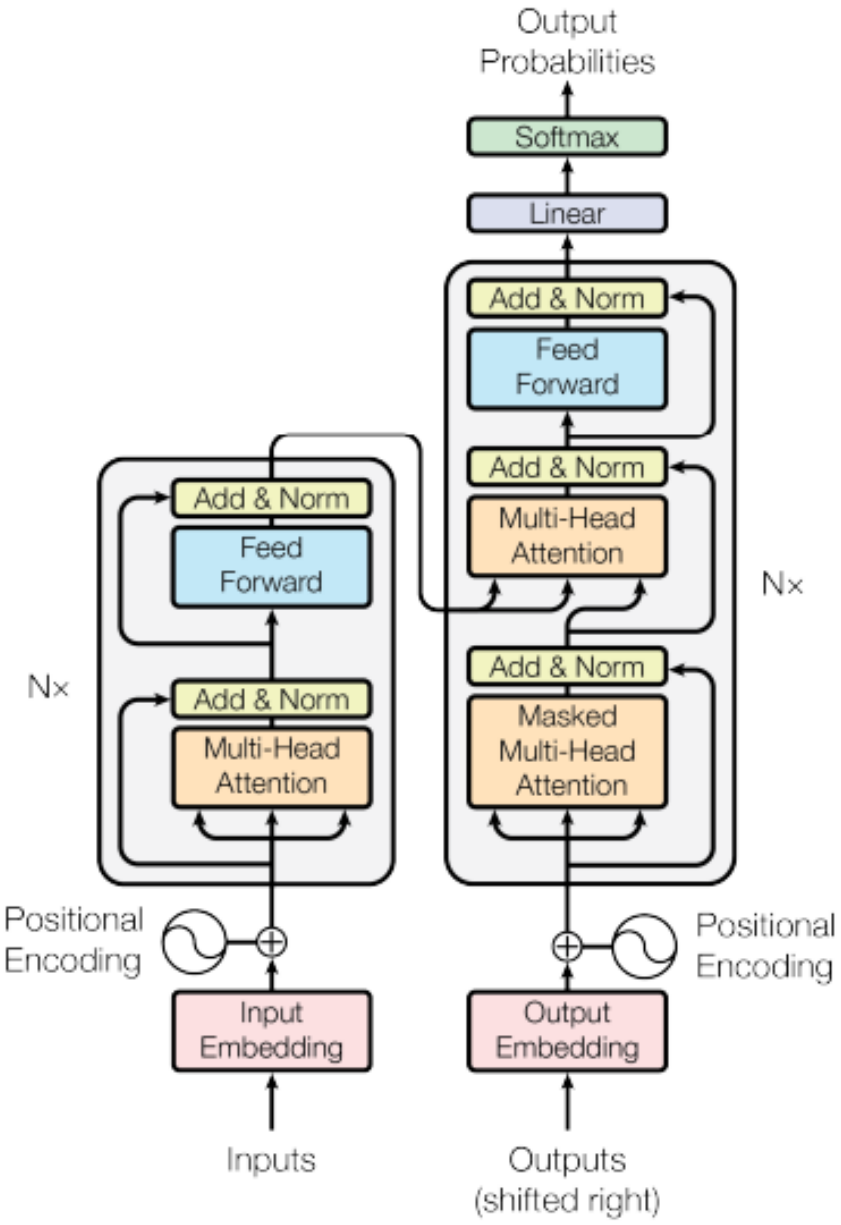

FIGURE 2.2 – The Transformer - model architecture[3].

2.3.1 Encoder and Decoder Stacks

Encoder

The architecture of the ViT model consists of a stack of N = 6 identical layers in the encoder. Each layer is composed of two sub-layers : a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. To ensure the flow of information through the layers, residual connections are employed around each of the two sub-layers, followed by layer normalization.

The output of each sub-layer is computed as LayerNorm(x + Sublayer(x)), where Sublayer(x) is the function implemented by the sub-layer itself. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension dmodel = 512.

Overall, the ViT model uses a transformer-based architecture with multi-head self-attention to embed information globally across the image and learn the relative position of image patches for image recognition tasks. This architecture has shown to be more computationally efficient and accurate than the current state-of-the-art CNNs for image recognition tasks [3].

Decoder

The decoder in the Transformer model also consists of N = 6 identical layers, just like the encoder. However, it has three sub-layers in each layer. The first sub-layer is a masked multi-head self-attention mechanism that prevents positions from attending to subsequent positions. The second sub-layer is a multi-head attention mechanism that attends over the output of the encoder stack. The third sub-layer is a simple, positionwise fully connected feed-forward network.

Similar to the encoder, the decoder also uses residual connections and layer normalization. The output embeddings in the decoder are offset by one position, and the masking ensures that predictions for a position can depend only on the known outputs at positions less than that position. This architecture enables the decoder to generate output sequences based on the input sequence and the encoded information from the encoder stack[3].

2.3.2 Attention

In the case of self-attention, the input to the attention mechanism is split into three parts : queries, keys, and values, which are all derived from the same input sequence. The compatibility function used to compute the weight assigned to each value is called the dot-product attention, and it computes the dot product between the query and each key, followed by a softmax function to obtain a probability distribution over the values. The output of the attention mechanism is then a weighted sum of the values, where the weights are given by the softmax probabilities[3].

2.3.2.1 Scaled Dot-Product Attention

We call our particular attention "Scaled Dot-Product Attention" (Figure)(left). The input consists of queries and keys of dimension d_k, and values of dimension d_v. We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values.

In practice, we compute the attention function on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V . We compute the matrix of outputs as :

$$Attention(Q, K, V) = softmax(\frac{1}{\sqrt{d_k}})V \tag{2.1}$$

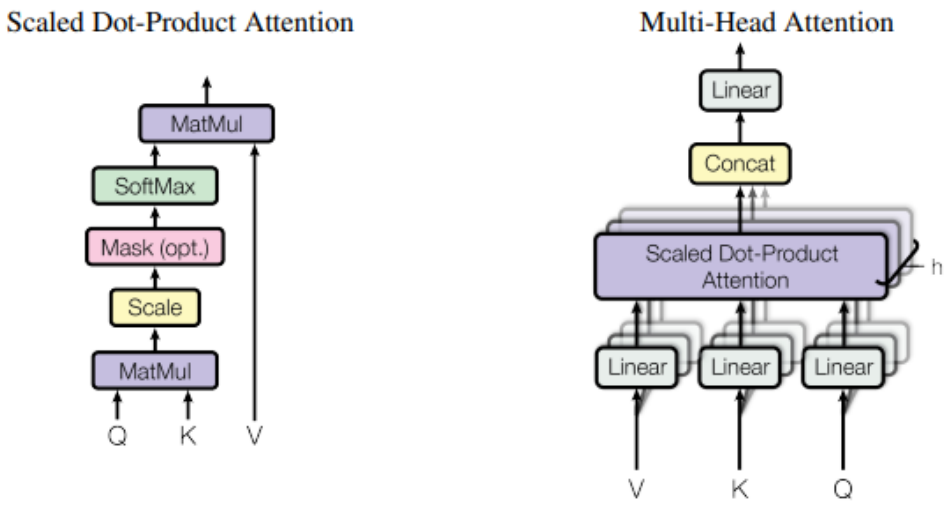

FIGURE 2.3 – (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel[3].

The attention mechanism has two commonly used functions : additive attention and dot-product attention. Additive attention uses a feed-forward network with a single hidden layer to compute the compatibility function. While both functions have similar theoretical complexity, dot-product attention is faster and more space-efficient in practice due to its implementation using highly optimized matrix multiplication code.

While for small values of d_k the two mechanisms perform similarly, additive attention outperforms dot product attention without scaling for larger values of d_k . We suspect that for large values of d_k, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. To counteract this effect, we scale the dot products by $\frac{1}{\sqrt{d_k}}$

2.3.2.2 Multi-Head Attention

Rather than using a single attention function with keys, values, and queries of dimension d_{model}, it is advantageous to linearly project them h times using distinct, trained linear projections to d_k, d_k, and d_V dimensions, respectively. The attention function is then applied independently to each of these projected queries, keys, and values, resulting in dv-dimensional output values. These values are concatenated and projected once more to generate the final values, as shown in Figure 2.3 (right) .

Multi-head attention allows the model to jointly attend to information from different representation sub-spaces at different positions. With a single attention head, averaging inhibits this.

$$MultiHaed(Q, K, V) = Concat(head_1, ..., head_h)^0 \tag{2.2}$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.3}$$

Where the projections are parameter matrices

$$W_i^Q \in R^{d_{model} * d_k}, W_i^k \in R^{d_{model} * d_k}, W_i^V \in R^{d_{model} * d_V} and W^0 \in R^{hd_v * d_{model}}$$

2.3.2.3 Applications of Attention in our Model

The Transformer uses multi-head attention in three different ways :

— In the "encoder-decoder attention" layers of the Transformer model, the queries are generated from the output of the previous decoder layer, while the memory keys and values are obtained from the output of the encoder. This setup allows each position in the decoder to attend to all positions in the input sequence, which is similar to the encoder-decoder attention mechanisms commonly used in sequence-to-sequence models.

— The encoder in the transformer model contains self-attention layers, where all the queries, keys, and values come from the same source, which is the output of the previous layer in the encoder. In other words, each position in the encoder can attend to all positions in the previous layer of the encoder.

— Similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position. We need to prevent leftward information flow in the decoder to preserve the auto-regressive property. We implement this inside of scaled dot-product attention by masking out (setting to $-\infty$) all values in the input of the softmax which correspond to illegal connections[3]. See Figure 2.3.

2.3.2.4 Position-wise Feed-Forward Networks

In both the encoder and decoder, each layer also contains a feed-forward sub-layer, which is applied to each position independently and identically. This sub-layer is composed of two linear transformations with a ReLU activation function applied in between them[3].

$$FFN(x) = max(0, xW_1 + b_1)W_2b_2 \tag{2.4}$$

While the linear transformations are the same across different positions, they use different parameters from layer to laye.

2.3.2.5 Embeddings and Softmax

To convert input tokens and output tokens to d_{model} dimensional vectors, we use learned embeddings. We apply a learned linear transformation and softmax function to convert the decoder output to predicted next-token probabilities. To share weights and reduce the number of parameters, we use the same weight matrix between the two embedding layers and the pre-softmax linear transformation, We scale the weight matrix by $\sqrt{d_{model}}$ in the embedding layers[3].

2.3.2.6 Positional Encoding

The Transformer/Attention architecture processes input embeddings in parallel, meaning it neglects the sequential order of the input sequence. To incorporate the sequence position information, a common approach is to append a positional vector to the input embeddings, which is referred to as "positional encoding". There are various options for positional encoding, with one common choice being the use of cosine functions with different frequencies[3].

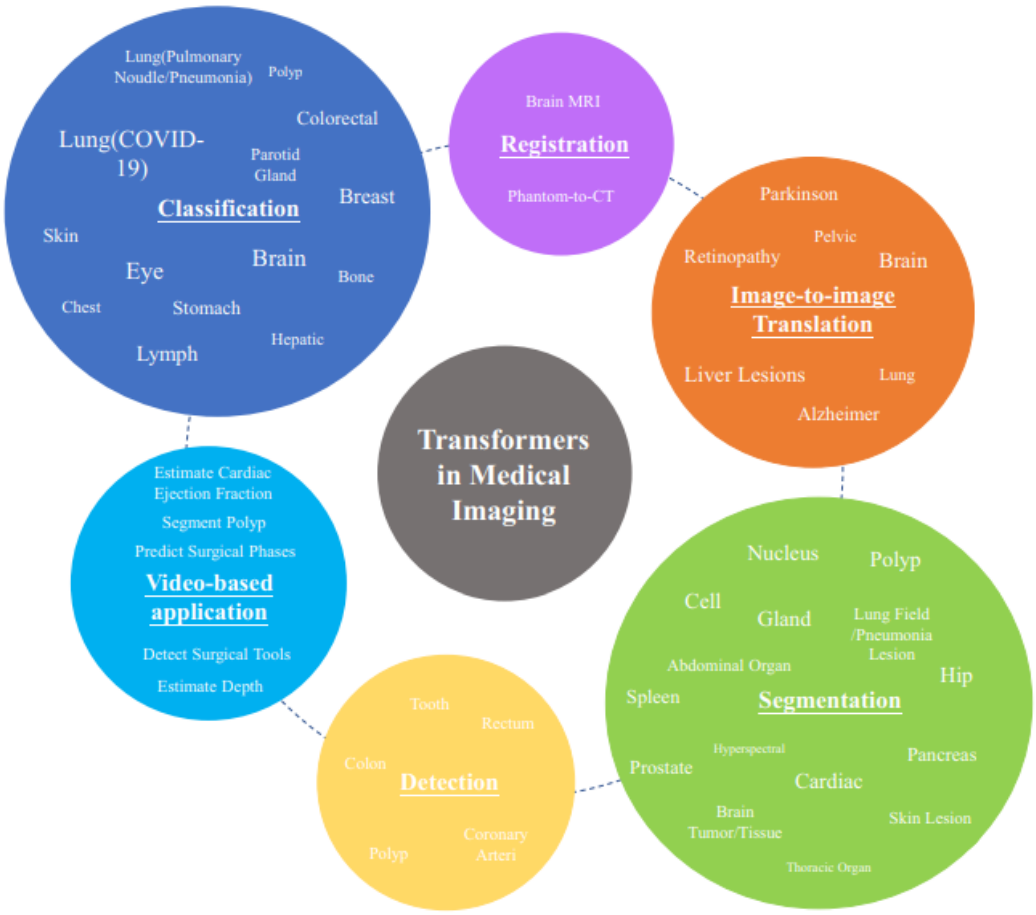

FIGURE 2.4 – Applications of transformers in medical image analysis[2].

2.4 Transformer application

Transformers have been widely used in medical image analysis. In this section, we first introduce transformer-based medical image analysis applications, including classification, segmentation, Anomaly Detection, Objects Detection, Image Compression, Videos Deepfake Detection and Clusters Analysis. We categorize these applications according to their learning tasks as illustrated in Figure 2.4 .

2.4.1 Image Classification

Image classification with transformers is a relatively new approach in the field of computer vision. Traditionally, convolutional neural networks (CNNs) have been the go-to architecture for image classification tasks, but transformers have shown promising results in recent years[19].

2.4.1.1 Image Classification Methode

To apply transformers to image classification, the input image is first divided into a set of patches, which are then processed through a sequence of transformer layers. The transformer layers compute the self-attention mechanism between all patches, allowing the network to learn long-range dependencies in the image. The final output of the transformer layers is then passed through a classifier, which outputs the predicted class label.

2.4.1.2 Image Classification Aim

One of the key benefits of using transformers for image classification is that they can handle variable-sized inputs, allowing them to process images of different resolutions without the need for cropping or resizing. Additionally, transformers are capable of modeling global dependencies in the image, which can be useful for recognizing patterns that span the entire image.

2.4.2 Anomaly Detection

Anomaly detection with transformers refers to the use of transformer-based models for detecting anomalies or outliers in a given dataset. Transformers are a type of deep neural network architecture that has achieved state-of-the-art results in natural language processing and other sequence-to-sequence tasks. Recently, transformers have also been applied to the problem of anomaly detection, with promising results[20].

2.4.3 Object Detection

Object detection is a computer vision task that involves detecting objects within an image and identifying their location with a bounding box. Transformers, which are a type of deep learning model commonly used in natural language processing (NLP), can also be adapted for object detection[21].

2.4.3.1 Transformers Application for Object Detection

In ViT, the image is divided into a set of patches, and each patch is treated as a separate token. The tokens are then fed into the transformer model, which processes them in a sequence. The final output of the transformer model is a set of embeddings, one for each patch.

To adapt ViT for object detection, a set of learnable parameters are added to the model to predict the bounding boxes and labels of objects within the image. This is typically done by adding a multi-head self-attention mechanism to the model, which allows it to attend to different regions of the image when predicting object locations.

Another approach for using transformers for object detection is to use the DETR (DEtection TRansformer) model. DETR is a transformer-based architecture that directly predicts object detections without using any anchor boxes or region proposal networks.

2.4.3.2 Object Detection Aim

Overall, transformers have shown promise for object detection tasks, and they offer several advantages over traditional object detection models, including the ability to process variable-sized inputs and the ability to leverage self-attention mechanisms to capture complex spatial relationships between objects in an image.

2.4.4 Image Segmentation

Image segmentation is a computer vision task that involves dividing an image into multiple segments or regions, each of which corresponds to a different object or part of the image. Transformers are a type of neural network architecture that have been successfully applied to various natural language processing tasks. However, they have also been explored for image segmentation tasks.

2.4.4.1 Image Segmentation Methode

One approach for using transformers for image segmentation is to first convert the input image into a sequence of smaller patches or tokens. Each patch is then processed by the transformer network, which uses self-attention to capture long-range dependencies between the patches. Finally, the output of the transformer is decoded to produce a segmentation mask that assigns a label to each pixel in the image.

2.4.4.2 Image Segmentation Aim

There are several advantages of using transformers for image segmentation, including the ability to handle variable-sized inputs and the ability to capture global context information. However, there are also some challenges, such as the need to balance the trade-off between local and global information and the high computational cost of processing large images.

2.4.5 Image Compresion

Image compression is the process of reducing the size of an image while maintaining its visual quality. One approach to image compression is using transformers, which are a type of neural network commonly used in natural language processing tasks, but can also be applied to image processing tasks.

Transformers can be used for image compression by treating an image as a sequence of pixels and applying the transformer architecture to this sequence. The pixels are flattened into a 1D sequence and fed into the transformer, which learns to compress the image by extracting the most important features and discarding the rest.

One common approach for using transformers for image compression is the use of the Vision Transformer (ViT) architecture, which was originally proposed for image classification. In this approach, the image is divided into non-overlapping patches, which are then flattened into a sequence and fed into the transformer. The output of the transformer is then decoded back into an image[22].

2.4.6 Videos Deepfake Detection

Deepfake detection is a crucial task in the field of artificial intelligence and computer vision, especially with the increasing use of deep learning techniques to create highly realistic synthetic media. Transformers are a popular deep learning architecture used for various natural language processing tasks, and they can also be adapted for image and video analysis[23].

2.4.6.1 Problem of Videos Deepfake Detection

Deepfake detection is a complex problem because deepfake videos are created using advanced techniques that can produce highly realistic and convincing content. Traditional approaches to detecting deepfakes, such as examining the video's metadata or analyzing visual artifacts, are often ineffective against sophisticated deepfakes.

2.4.6.2 Solution

One way that transformers can be used for deepfake detection is by analyzing the spatial and temporal features of videos. Spatial features refer to the visual information in each frame of the video, such as facial

expressions and body movements. Temporal features refer to how these spatial features change over time, which can be important for detecting inconsistencies in the video's content.

Transformers can be trained to extract spatial and temporal features from videos and then use this information to distinguish between real and fake videos. They do this by processing the video frames sequentially, much like how they process words in a sentence during natural language processing tasks.

To achieve high accuracy in deepfake detection, transformers are often combined with other techniques, such as contrastive learning, which involves training the model to distinguish between similar and dissimilar pairs of videos. This helps the model to learn more robust representations of real and fake videos, which can improve its ability to detect deepfakes.

Overall, using transformers for deepfake detection in videos is a promising area of research that has the potential to make significant strides in addressing the growing problem of deepfake content online.

2.4.7 Clusters Analysis

Cluster analysis in transformers is a technique that is used to group together similar words or embeddings within the transformer's hidden layers. It is a way to gain insights into how the transformer is processing and representing information.

2.4.8 Cluster Analysis Applications

There are many different applications of cluster analysis in transformers, including :

— Understanding the relationships between different concepts in the input

— Identifying common patterns or themes in the input

— Improving the interpretability of the transformer's representations

— Identifying potential errors or biases in the model's representations

2.5 Challenges

Computer vision involves capturing visual information in the form of arrays of pixels, which are then processed using convolutions - a deep learning technique commonly used in computer vision. Despite the success of this approach, there are significant challenges that need to be addressed :[24].

Firstly, not all pixels in an image are equally important. For example, in image classification, foreground objects are more important than the background, and in segmentation, Brain MRI (s) lessions should be prioritized over the whole region of interested. However, convolutions treat all image patches equally, which results in spatial inefficiency in both computation and representation.

Secondly, not all images contain the same concepts. Low-level features likeBrain MRI in all images, making low-level convolutional filters appropriate. However, high-level features like Brain MRI are specific to certain images, making the application of high-level filters to all images computationally inefficient. Rarely-used filters end up consuming significant amounts of computing power, as they are not applicable to all images.

Finally, convolutions struggle to relate spatially-distant concepts. While approaches like increasing kernel sizes, model depth, and using operations like dilated convolutions, global pooling, and non-local attention layers have been used to address this issue, they add complexity to the model and computational requirements. Overall,

while convolutions have been successful in computer vision, addressing these challenges requires moving beyond the current pixel-convolution paradigm.

2.6 Visual Transformers Models

Visual Transformers are a class of deep learning models that use a self-attention mechanism to process image data. They were introduced in the paper "An Image Is Worth 16x16 Words : Transformers for Image Recognition at Scale" by Dosovitskiy et al. in 2021[18].

Visual Transformers are similar in architecture to the popular language model called "Transformer," which was originally developed for natural language processing. Instead of processing sequences of words like the Transformer does, Visual Transformers process image patches as sequences of vectors using self-attention.

2.6.1 General Visual Transformer Model Pipeline

The Vision Transformer model consists of the 7 steps

— Split an image into patches (fixed sizes)

— Flatten the image patches

— Create lower-dimensional linear embeddings from these flattened image patches

— Include positional embeddings

— Feed the sequence as an input to a state-of-the-art transformer encoder

— Pre-train the ViT model with image labels, which is then fully supervised on a big dataset

— Fine-tune the downstream dataset for image classification

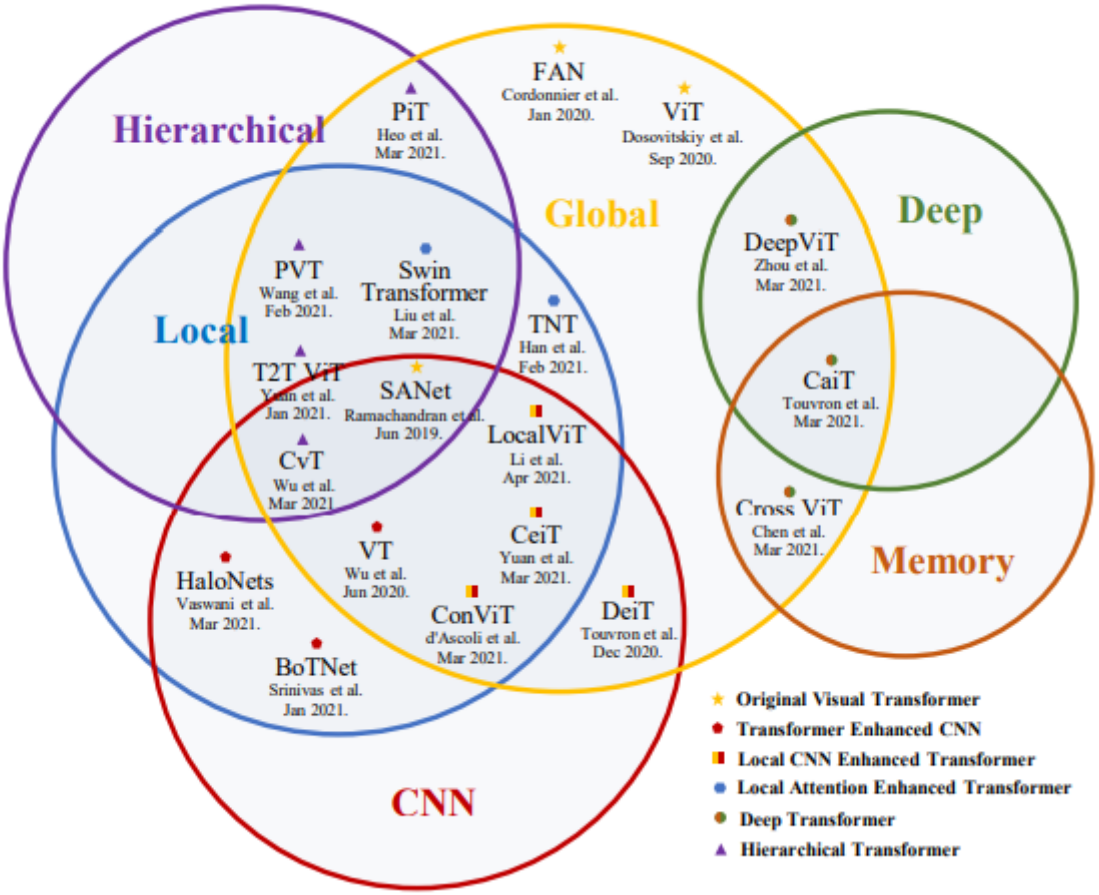

FIGURE 2.5 – Taxonomy of Visual Transformer[4]

2.6.2 Transformer Backbone

— Original Visual Transformer

Drawing inspiration from the remarkable accomplishments of Transformers in the field of Natural Language Processing (NLP), several recent technology trends in the area of computer vision tasks have integrated attention mechanisms with convolution models to enhance the models' receptive field and global dependency. While hybrid models of this nature have been explored, Ramachandran et al. have explored the possibility of completely replacing convolution with attention. They introduced the Stand-Alone self-attention network (SANet), which has demonstrated superior performance on vision tasks when compared to the original baseline. In their approach, the authors replaced the spatial convolution layer (3×3 kernel) in each bottleneck block of a ResNet architecture with a locally spatial self-attention layer, while keeping other structures unchanged from the original ResNet setting. Various ablations have indicated that the efficacy of the network can be further enhanced by adding positional encodings and a convolutional stem.

In contrast to the focus on light-scale models, the Vision Transformer (ViT) has explored the effectiveness of large-scale pre-trained learning using vanilla Transformers. This pioneering work has had a significant impact on the community. Since vanilla Transformers only accept sequential inputs, the input image in ViT is split into non-overlapping patches, which are then projected into patch embeddings. A 1D

learnable positional encoding is added to the patch embeddings to retain spatial information, and the joint embeddings are then fed into the encoder, as shown in Figure 5. Similar to BERT, a learned [class] token is attached to the patch embeddings to aggregate global representation, which serves as the final output for classification. Additionally, a 2D interpolation complements the pre-trained positional encoding to maintain patch order when the input images have arbitrary resolutions. By pre-training with a large-scale private dataset (JFT-300M), ViT has achieved similar or even superior results on multiple image recognition benchmarks (ImageNet and CIFAR-100) when compared to prevailing CNN methods. However, the model's generalization capability tends to deteriorate with limited training data[4].

— Transformer-enhanced CNNs

The Transformer architecture is comprised of two main components : Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN). While the convolutional layer and MHSA are known to be approximately equivalent, the Transformer can further improve the effectiveness of MHSA by incorporating skip connections and FFN. Recently, there have been attempts to integrate the Transformer into Convolutional Neural Networks (CNNs) to enhance representation learning. One such approach is the Vision Transformer (VT), which decouples semantic concepts for an input image into different channels and relates them densely through the VT-block of the encoder. This VT-block replaces the last convolution stage in a CNN model, enhancing its ability for semantic modelling[4].

Unlike previous approaches that directly replace the convolutional layer with an attention structure, a new conceptual framework has been proposed, which defines successive bottleneck blocks with MHSA as Bottleneck Transformers (BoTNet) blocks. This approach adopts relative position encoding to mimic the original Transformer and has demonstrated superior performance compared to most CNN models with similar parameter settings on the ImageNet benchmark.

— CNN-enhanced Transformer

Inductive bias refers to a set of assumptions about data distribution and solution space that are incorporated into machine learning models. Within convolutional neural networks (CNNs), the bias takes the form of locality and translation invariance, which allow CNNs to process images effectively by taking advantage of the large covariance within local neighborhoods. However, these strong biases can limit the upper bound of CNNs when there is sufficient data available. Recent efforts have focused on leveraging the appropriate CNN bias to enhance the Transformer architecture[4].

One example of such an effort is the Data-efficient image Transformer (DeiT), which aims to reduce the ViT's dependence on large datasets. In addition to existing strategies for data augmentation and regularization, DeiT employs a teacher-student distillation strategy for auxiliary representation learning. In this approach, the ViT serves as the student model, with a distilled token attached to the patch embeddings and supervised by pseudo-labels from the teacher model. Surprisingly, extensive experiments have shown that CNNs serve as better teacher models than Transformers, with the distilled student Transformer even outperforming its teacher CNN model. This suggests that the teacher CNN transfers its inductive bias in a soft way to the student Transformer through knowledge distillation. DeiT-B, based on ViT's architecture, has achieved a top-1 accuracy of 85.2percent without external data[4].

Other approaches that impose inductive biases on the Transformer architecture include ConViT, which appends a parallel convolution branch to the vanilla Transformer to impose biases softly. The convolution branch has a learnable embedding that initially approximates the locality as closely as possible to the

convolution. The gating parameter is then adjusted to give each attention head the freedom to escape the locality. CeiT and LocalViT extract the locality by directly adding a depth-wise convolution in the feed-forward network (FFN)[4].

— Local Attention Enhanced Transformer

ViT's coarse patchification process overlooks local image information, prompting researchers to propose various local attention mechanisms to enhance the model's local feature extraction capabilities. One notable approach is the Shifted Windows (Swin) Transformer, which utilizes a shifted window to capture global and boundary features. Swin employs two consecutive window-wise attention layers to promote cross-window interactions, similar to how CNNs expand their receptive field. This method reduces computational complexity while achieving 84.2percent accuracy on ImageNet and the latest State of the Art (SoTA) on multiple dense prediction benchmarks[4].

To capture both patch- and pixel-level information, researchers proposed the Transformer-iNTransformer (TNT) model, which has two blocks per layer. The inner block models pixel-wise interactions within each patch, while the outer block captures global information. TNT uses a spatially separable self-attention mechanism, similar to depth-wise convolution or window-wise TNT, to extract local-global representations. ViL is another approach that replaces the single class token with a set of local embeddings (global memory) that interact with their corresponding 2D spatial neighbors using inner attention[4].

— Hierarchical Transformer

ViT's fixed-resolution columnar structure results in the loss of fine-grained features and significant computational costs. To address this issue, Tokens-to-Token ViT (T2T-ViT) introduces a hierarchical Transformer paradigm and utilizes overlapping unfold operations for down-sampling, which incurs high memory and computation costs. In contrast, Pyramid Vision Transformer (PVT) adopts non-overlapping patch partitions to reduce feature size and employs a spatial-reduction attention (SRA) layer to further decrease computational costs by learning low-resolution key-value pairs. Empirically, PVT performs well on dense prediction benchmarks that require large inputs and fine-grained features with computational efficiency. Additionally, PiT and CvT utilize pooling and convolution, respectively, for token downsampling. CvT improves upon PVT's SRA by replacing the linear layer with a convolutional projection, and thanks to the convolutional bias, CvT can handle inputs of arbitrary sizes without requiring positional encodings[4].

— Deep Transformer

Recent studies have shown that increasing the depth of a Transformer model can enhance its learning capacity. To investigate the scalability of deep Transformer, recent works have employed a deep architecture and conducted extensive experiments to analyze the cross-patch and cross-layer similarities, as well as the contribution of residual blocks. However, in deep Transformers, the features from deeper layers tend to be less representative, and patches are often mapped into indistinguishable latent representations, leading to attention collapse and patch over-smoothing. To overcome these limitations, current methods propose solutions from two perspectives[4]. Efficient Class-attention in image Transformers (CaiT) is a two-stage model that addresses the limitations of the class token in image Transformers. The first stage consists of multiple self-attention layers without a class token, where a learned diagonal matrix is used to update channel weights dynamically. This allows for flexibility in adjusting channel weights. In the second stage, a class token is inserted in the last few class-attention layers with frozen patch embeddings to model

global representations. This separation is based on the assumption that the class token is invalid for the gradient of patch embeddings in the forward pass. With a distillation training strategy, CaiT achieves a new state-of-the-art (SoTA) on imagenet1k with 86.5percent top-1 accuracy without external data[4].
Despite the attention collapse and oversmoothing problems of deep Transformers, the diversity of the attention map between different heads is largely preserved. Building on this observation, Deep Vision Transformer (DeepViT) aggregates different head attention maps and re-generates a new one by using a linear layer to increase cross-layer feature diversity. Refiner expands the dimension of attention maps using a linear layer, indirectly increasing the head number, and employs Distributed Local Attention (DLA) to better model both local and global features. DLA is implemented using a head-wise convolution effecting on the attention map.

— Transformers with Self-Supervised Learning

Recently, there has been an increasing interest in designing self-supervised learning schemes for visual Transformers in both generative and discriminative ways, following the success of self-supervised learning in NLP. Generative models like iGPT directly resize and flatten the image into lower resolution sequences, which are then fed into a GPT-2 model for auto-regressive pixel prediction. BEiT, on the other hand, uses dVAE to vectorize the image into discrete visual tokens, which serve as pseudo-labels for pre-training ViT. For discriminative models, MoCo v3 and DINO are proposed, which extend MoCo and teacher-student recipe to self-supervised learning, respectively. DINO uses a momentum encoder as a teacher model and an online encoder as a student model to fit the teacher's output, connected with a standard cross-entropy loss. Self-supervised ViT can learn features that are not attainable by supervised models, particularly for segmentation tasks[4].

2.6.3 Transformer Neck

— Original Transformer

To achieve direct set predictions in detection, two crucial factors are necessary : (1) a loss function that ensures a one-to-one correspondence between predicted and actual bounding boxes, and (2) an architecture capable of predicting a group of objects and their relationships in a single iteration.
The DETR model makes N predictions of a fixed size during a single decoding pass, where N is intentionally set to be considerably greater than the usual number of objects present in an image. One of the major training challenges is evaluating the predicted objects' (class, location, size) accuracy compared to the ground truth. Our loss function achieves an ideal bi-partite matching between predicted and actual objects and optimizes object-specific (bounding box) losses[25].

— Sparser Attention

When using Transformer attention on image feature maps, the primary challenge is that attention must be paid to every possible spatial location, which can be computationally intensive. To solve this issue, we propose a deformable attention module inspired by deformable convolution (introduced by Dai et al., 2017 and Zhu et al., 2019b). This attention module focuses solely on a limited set of key sampling points surrounding a reference point, regardless of the feature maps' spatial size. By assigning a small, fixed number of keys for each query, convergence and feature spatial resolution problems can be alleviated[26].

— Spatial Prior

We introduce a highly effective yet straightforward object detection framework called Efficient DETR.

It consists of only one decoder layer and three encoder layers, without the cascade structure present in the decoder. Efficient DETR is composed of two parts : dense and sparse. The dense section makes predictions on dense features from the encoder. To select a proposal set from the dense predictions, we use a top-k selection method. The 4-d proposals and their corresponding 256-d features from the decoder are utilized as the initialization for reference points and object queries. In the sparse section, object containers, reference points, and object queries initialized with dense priors are given to a one-layer decoder for further refinement by interacting with the encoder features. The final results are predicted from the refined object containers. The detection head is shared by both parts. Deformable attention modules are utilized in all encoder and decoder layers[27].

— Structural Redesign

When it comes to model design, YOLOS largely mirrors the original ViT architecture and is tailored for object detection in a similar fashion. Moreover, YOLOS can be easily adapted to numerous canonical Transformer architectures that are available in both NLP and computer vision. This simple configuration is not designed to improve detection performance, but rather to expose the Transformer family's characteristics in object detection as impartially as possible[28].

— Pre-trained Model

UP-DETR consists of two procedures : pre-training and fine-tuning. In the pre-training stage, the transformers are trained in an unsupervised manner on a large-scale dataset without any human annotations. In the fine-tuning stage, the entire model is fine-tuned with labeled data that is identical to the original DETR on the downstream tasks[29].

— Matching Optimizer

We utilized the DAB-DETR architecture as the foundation for our training approach. Our model follows the same structure as DAB-DETR, whereby we explicitly represent the decoder queries as box coordinates. The only variance between our architecture and theirs is the decoder embedding, which is implemented as a class label embedding to support label denoising. Our primary contribution is the training method, as illustrated in Fig. 6. Our model comprises a Transformer encoder and a Transformer decoder, similar to DETR. The encoder utilizes a CNN backbone to extract image features, which are then processed through the Transformer encoder with positional encodings to obtain refined image features. On the decoder side, queries are inputted to the decoder to search for objects via cross-attention[30].

2.7 Visual Transformers Models

Visual Transformers, or ViTs, can be seen as an extension of the Transformer architecture that was originally developed for natural language processing. However, ViTs have been adapted for the image domain with some modifications to handle image data.

In ViTs, an input image is first divided into a set of image patches or visual tokens, which are then embedded into a set of encoded vectors with fixed dimensions. The position of each patch in the image is also embedded along with the encoded vector.

The ViT model then uses a Transformer encoder network to process these embedded visual tokens and learn representations that capture the relationships between them. This is similar to the Transformer's processing of text input, but with modifications to accommodate the visual nature of the input data.

Overall, while ViTs are an extension of the Transformer architecture, they are specifically designed to handle image data and utilize modified tokenization and embedding methods to do so. The core architecture of the Transformer remains largely intact, but with adjustments to suit the unique characteristics of visual input data.

2.8 Conclusion

Visual Transformers, represented by the ViT model, have proven to be highly effective in computer vision tasks and have challenged the dominant position of convolutional neural networks (CNNs) in this field. In this chapter, we have conducted a comprehensive review of over one visual Transformer models that have been applied to various computer vision tasks, including classification, detection, anomaly detection, and segmentation.

In addition to our review of visual Transformer models, we also discussed the underlying principles of how these models work in computer vision tasks. Specifically, the ViT model uses multi-head self-attention to process images without relying on image-specific biases.

The ViT model divides an input image into a series of positional embedding patches, which are then fed into the transformer encoder. This allows the model to capture both local and global features of the image and learn representations that are effective for various computer vision tasks.

Furthermore, the ViT model has been shown to achieve higher precision rates on large datasets with reduced training time compared to traditional convolutional neural networks. Overall, the ViT model represents an exciting advancement in the field of computer vision and has the potential to significantly improve the performance of various computer vision tasks.

Chapitre 3

Related Works

3.1 Introduction

Artificial intelligence has become one of the fastest-growing fields in computer science, with a wide range of applications in various domains, including healthcare and image processing As such, the field has attracted significant attention from researchers, and a considerable amount of research has been conducted to improve the performance and efficiency of machine learning models.

In this chapter, we provide an overview of the existing literature related to machine learning. We then review recent research that has been done on improving the performance and efficiency of machine learning models in computer vision

Moreover, we explore the current state-of-the-art techniques used in medical image processing, such as neural networks, transformer.

Overall, this section aims to provide a comprehensive overview of the existing literature on transformer, its applications in medical image processing, and the state-of-the-art techniques and approaches, dataset to used and and its degree of success.

3.2 Binary Classification of MRI Brain Tumors Using CNN Models

Conventionally, identification of objects is mostly based on visual observation or physical sensors [31] ; the final conclusions are also inaccurate since human visual observation is subjective and infrequent, and observation times and locations can be limited by objective conditions. The development of a brain tumor occurs in tissues surrounding the brain or the skull, and it has a significant impact on a person's life There are two categories of brain cancer growth benign or malignant tumors. Tumors that develop from outside the brain are known as brain metastasis tumors. Primary cancers begin within the brain, whereas secondary cancers are caused by tumors that have spread there from somewhere else [32]. Among the most commonly occurring primary brain tumors are Gliomas, Meningiomas, and pituitary adenomas. It is common for these tumors to grow unevenly in the brain, placing pressure on existing tissue [33] . The development of malignant tumors is uneven as compared to benign tumors, triggering damage to the surrounding tissues [34]. Due to its noninterfering characteristic, MRI is preferred over other imaging techniques like computed tomography (CT), positron emission tomography (PMT), and Xrays [35] [36]. Approximately 72,360 people aged 40 and over in the United States will be diagnosed with a primary brain tumor by 2022. The number of Americans with primary brain tumors is estimated at 700,000. Benign tumors account for 71% of all cases while malignant tumors account for 29% [37]. Medical methods such as MRI are used to detect tumors, and biopsies are taken for further examination [38] [37]. By literature, Convolutional neural network (CNN) is extensively known to classify brain tumors with high performance. Since CNN is widely used for collecting features randomly without perceiving the local and global features which causes the overfitting problems, In the current study, we used a convolutional neural network (CNN) to compare classification from images, initially the pretraining step that used several neuronal network models applied on several images that belongs the same brain dataset, then tuning the trained model parameters to the new CNN model and lastly fine-tuning the CNN model to achieve the classification of two types (binary classification), whether images infected or not (positive or negative). The main experimental results that have showed the efficiency of the CNN models, which were interpreted by reaching excellent classification performance and achieving better classification results, in the following table demonstrates the important research papers using binary classification using brain MRI images based on CNN models.

The main studies working on Brain MRI classification using CNN models table3.1 :

TABLE 3.1 – The main studies working on Brain MRI classification using CNN models

Research paper	Dataset	Task	Proposed Model	Performance	Highlight
[JIANG, Yun., 2018][39]	MICCAI BRATS2015	seg	CNN	86.30%	threshold filter
[LIU, Dongnan., 2018][40]	MICCAI BRATS2017	seg	DCNN	86.50%	multi-dimensional
[S. Yannick,. 2019][41]	Brats2018	Binary	DenseNet+CNN+SVM	72.20%	3D CNN
[IRSHEIDAT, Suhib., 2020][38]	Kaggle	Binary	ACNN	88.25%	threshold filter
[Y. Bhanothu,. 2020][42]	MR dataset consist 3 class	multi class	R-NN	97.60%	Data Augmentation
[RAHMAN, Takowa., 2023][37]	Kaggle	Binary	CNN	97.60	

3.3 Transformers for Medical Image Segmentation Tasks

In this section, it exposes a brief review of the main research papers for lesions segmentation from brain MRI images 3.2. Due to the successful performance of Transformer-based approaches in medical image analysis for computer aided for detection and computer aided for diagnosis, scientists have focalized on suggesting diverse ViT-based models for recognizing automatically the presence of the anomalies(such as infections, lesions, cancers, . . .etc.) [43].

Table 3.2 – The main studies working using ViT models :

References	Task	Dataset	Performance	Highlight
|trnsUnet, Chen et al.|[44]	ACT_MOS	Synapse multi-organ CT	77.48	Claw Unet
|TrancClaw, yao et al.|[45]	ACT_MOS	Synapse multi-organ CT	78.09	Claw Unet
|LeVit-Unet, Xu et al.|[46]	ACT_MOS	Synapse multi-organ CT	78.53	Claw Unet
|trnsUnet, Chen et al.|[44]	Cardiac segmentation	ACDC	89.71	Claw Unet
|LeVit-Unet, Xu et al.|[46]	Cardiac segmentation	ACDC	90.32	Claw Unet
|TransBTSV2, Fu et al.|[47]	Brain tumor segmentation	BraTS 2019 dataset	85.18	-
|CA-GANformer, You et al.|[48]	Liver Tumor segmentation	liTS dataset	73.82	-
|TransFuse, Zhang et al.|[49]	Prostat segmentation	MSD dataset	76.4	-

3.4 Lesion Classification of Brain MRI Tumors Using Transformers (ViT)

Computer-aided diagnostic (CAD) systems are often used by radiologists in order to provide another opinion and to handle a large number of patients. It has been demonstrated that such emerging software systems are beneficial diagnostic tools that are capable of detecting and classifying even confusing brain tumors classes. Additionally, a CAD system could also analyze a large quantity of patients rapidly and efficiently without requiring any labor concentration or effort [50]. It is possible to use the CAD system alongside brain MRI images to determine lesion density and shape as well as detect anomalies such as masses and calcifications, which will result in a more positive prognosis and a higher survival rate. Aside from that, MRI modalities that provide 2D and 3D brain imaging are available ; however, radiologists struggle to differentiate normal from abnormal tissue in the image scans. In order to minimize the likelihood of false positives and recall rates, CAD systems must be developed for accurate classification of MRI scans. With the use of artificial intelligence technology, creating CAD systems can become simpler, more reliable, and more reliable. AI is capable of generating a million or more deep high-level features at once without requiring any user input.

TABLE 3.3 – Transformers used in medical image classification tasks

modalities	Research paper	Malady	Organ	Datasets	focus	Accuracy
X-ray	|Park et al.|[51]	COVID-19	Lung CT	-	Pretrained backbone on CXRs	-
	|Tanzi et al.|[52]	Femur fracture	Bone	-	Unsupervised learning, compare CNNs with ViTs	77.0
	|Van et al. |[53]	Mammography Chest X-ray	Breast Lung	CBIS-DDSM CheXpert	Trans combine multi-view info	-
CT	|Costa et al|[54]	COVID-19	Lung	COVIDx	ViT with performer	91.0
	|COVID-VIT|[55]	COVID-19	Lung	COV19-CT-DB	use sub-volumes for 3D images	96.0
	|MIA-COV19D |[56]	COVID-19	Lung	COV19-CT-DB	use sub-volumes for 3D images	76.6
MRI	|He et al.|[57]	Brain Ag	Brain	BGSP,OASIS-3,NIH-PD,IXI	image-level and patch-level	-
	|Kim et al.|[58]	Gender Classification	Brain	HCP-Rest	lspatio-temporal attention for brain graph representation	88.2
	|3DMeT|[59]	Knee Cartilage Defect	Knee	image-level and patch-level	Generalize Trans	66.4

3.5 Conclusion

In this chapter, we have explored the existing literature and related work in the field of medical image proccessing. The aim was to understand the progress made in utilizing Convolutional Neural Networks (CNN) and Vision Transformers (ViT) for this task and to identify the gaps and opportunities for further research.

Our review revealed that CNN-based approaches have been widely adopted for image proccessing and have shown promising results. CNNs have proven effective in capturing local spatial features and have been successfully applied to various medical image analysis tasks. They have been utilized in different stages of image of illnesses such brain tumor, including segmentation, classification, and localization.

Additionally, the emerging Vision Transformer (ViT) architecture has gained attention in the computer vision community. Although relatively new, ViTs have shown impressive performance on image classification tasks. However, their application in the context of tumors detection is still limited, and more research is needed to explore their full potential.

We also observed that most studies focused on individual models, either CNNs or ViTs, and their performance on specific datasets. There is a lack of comparative studies that directly compare the performance of CNNs and ViTs for brain tumor detection. Such comparative analyses would provide valuable insights into the strengths and weaknesses of each approach and could potentially lead to the development of hybrid models that combine the advantages of both architectures.

In conclusion, the literature review has highlighted the progress and potential of CNNs and ViTs in tumors detection. While CNNs have been extensively studied and applied, ViTs present an emerging and promising avenue for exploration. The comparative analysis of these models and the utilization of large-scale datasets are key areas for future research.

Chapitre 4

Proposed Methodology

4.1 Introduction

Brain tumors are abnormal growths of cells in the brain that can cause severe health issues and even be life-threatening. Early detection and accurate diagnosis of brain tumors are crucial for effective treatment and improved patient outcomes. In recent years, advancements in deep learning techniques have shown promising results in medical image analysis, particularly in the field of brain tumor detection.

Convolutional Neural Networks (CNN) and visual Transformers (ViT) are two powerful deep learning architectures that have revolutionized computer vision tasks. By leveraging their ability to learn complex patterns and features from images, CNNs and ViTs have been successfully applied to various medical image analysis tasks, including brain tumor detection and classify them.

CNNs are well-known for their effectiveness in capturing local spatial features within an image. They consist of multiple layers of interconnected neurons that perform convolution operations, pooling, and non-linear activations. CNNs excel in extracting hierarchical representations of images, allowing them to identify distinctive features associated with brain tumors.

On the other hand, visual Transformers (ViT) have emerged as a new paradigm for image classification tasks. Unlike CNNs, ViTs leverage the self-attention mechanism to capture global image patterns and long-range dependencies. They divide the input image into patches and process them using transformers, which are originally designed for natural language processing tasks. ViTs have demonstrated impressive performance on various computer vision benchmarks and are gaining popularity in medical imaging applications.

The combination of CNN and ViT architectures holds great potential for accurate and efficient brain tumor detection. By combining the strengths of both models, we can leverage CNNs' ability to capture local features and ViTs' capability to capture global context, leading to improved detection accuracy and robustness.

In this chapter, we set two model the first model is CNN and second model is ViT. We will leverage a dataset of brain images, including both tumor and non-tumor samples, to train and evaluate the model.

The remainder of this chapter is organized as follows : Firstly we provides a detailed overview of the methodology, including the data collection, model architecture, training process, and evaluation metrics. secondly presents the experimental results and analysis. Finally, we compare the two models and prove the superiority of one of the models over the other, and determine this from the results obtained through the implementation of the two models.

4.2 Proposed Pipeline

Brain tumors are abnormal cell masses that can take different forms and may be either noncancerous (benign) or cancerous (malignant). These tumors can originate within the brain (primary brain tumors) or spread to it from other parts of the body (secondary or metastatic brain tumors). The speed of tumor growth and its location can significantly impact the nervous system's functionality. The type, size, and location of the brain tumor determine the available treatment options.

Our objective is to create a deep learning model that utilizes both vision transformers and Convolutional Neural Networks (CNNs), and evaluate their performance using various metrics For this we Proposed pipeline the following :

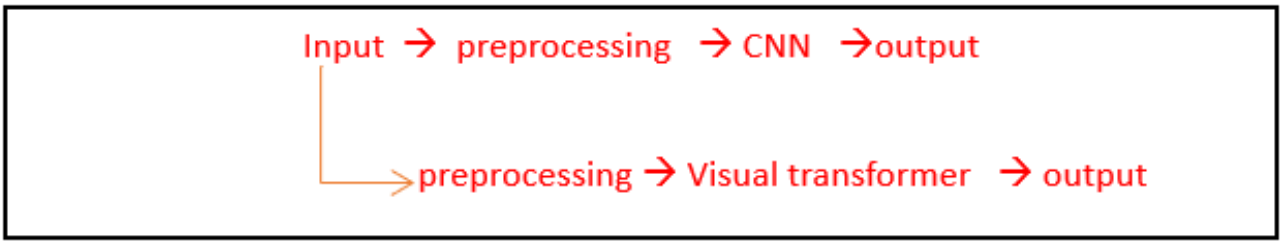

FIGURE 4.1 – Proposed pipeline.

4.3 Dataset

In this study, we used Brain Tumor Detection dataset, which is considered one of the most comprehensive and widely used datasets in the literature on Brain cancers diagnosis. The choice of dataset in any machine learning project is crucial as it has a direct impact on the performance of the models trained. The dataset was obtained available on Kaggle, which we carefully curated and preprocessed to ensure that they contained only high-quality of brain MRI images that could provide valuable information about several type of brain tumors. This was done to ensure that the dataset was focused on images that could help in the detection and diagnosis of brain cancers. After preprocessing, the dataset consisted of a total of 253 images,with an average image size of 240 x 240 pixels. which were divided into two categories : consists of 98 normal images (have not tumor), 155 Containing tumor . This Brain Tumor large and diverse dataset provided a rich and representative sample of brain images, which could help to train and evaluate VIT and CNN models for brain cancer diagnosis. Examples of datasets Brain MRI Images are shown in Figure 4.2.

4.3.1 Dataset Processing and Splitting

To prepare the dataset for machine learning, we transformed the data into a format that can be easily used for model training. The images in our dataset underwent the following transformations :

— The first step, We use OpenCV's imread function to read the image file. It builds the path to the image by concatenating the path_folder, the /yes or no/ subfolder, and the image_name. The imread function reads the image as a NumPy array.

— Next, We convert the NumPy matrix representation of the image into a PIL Image object using Image from array function. It defines the color mode as "RGB", indicating that the image has three color channels : red, green and blue.

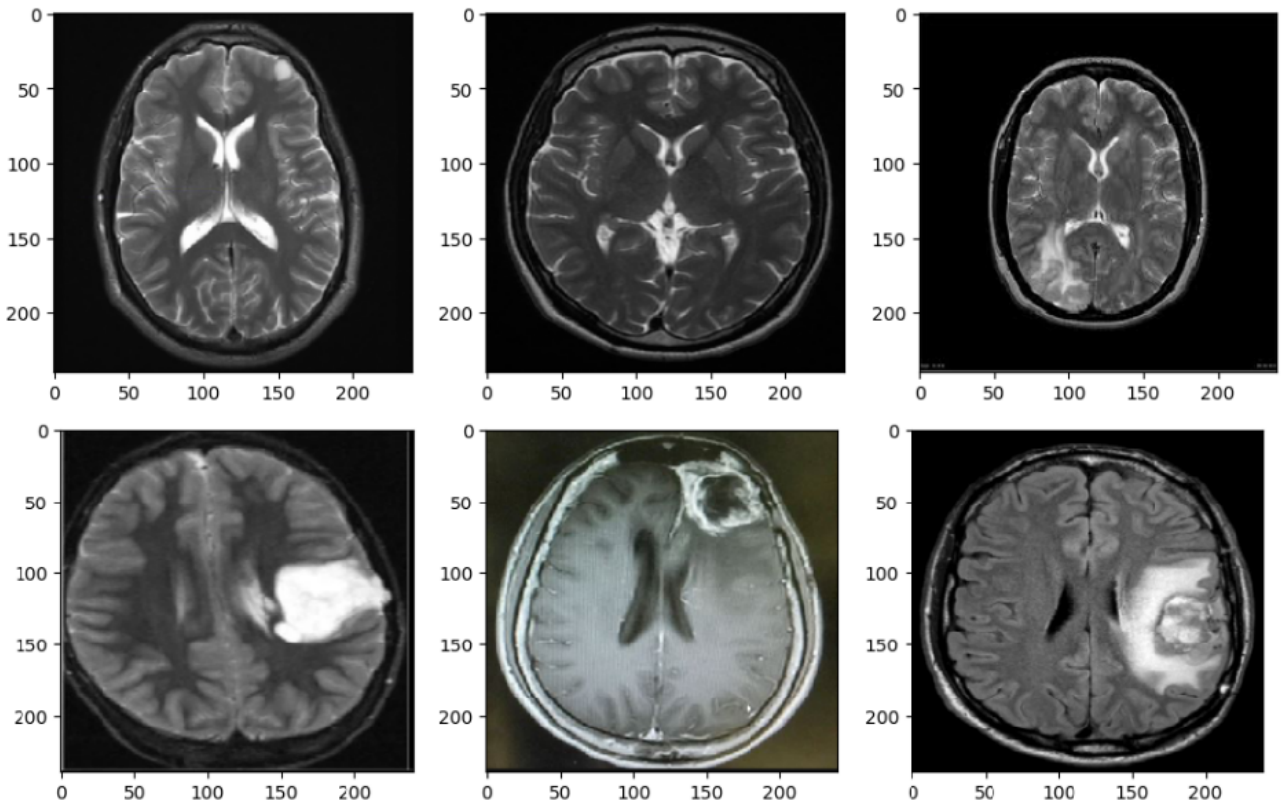

FIGURE 4.2 – Example of brain tumor images with different classifications.

— We resize the image to a target size of (240, 240) pixels using the Image Object Resize method. The image is scaled to fit the specified dimensions while maintaining its aspect ratio.

— After resizing the image, it converts the Image object back to a NumPy array using np.array(image). The resulting array represents the resized image and is appended to the dataset list.

— Finally, the label is appended to the lab list, indicating that the image does or does not contain a brain tumor.

In summary, an image with a brain tumor is read using OpenCV, converts it into a PIL Image object, resizes it to a specific dimension, converts it back into a NumPy array, and appends both the processed image and its corresponding label to the dataset and lab lists, respectively.

4.4 Evaluation Metrics

The performance metrics we use for evaluation purposes are common metrics employed in medical image classification studies, including classification accuracy (Training Accuracy and Validation Accuracy), Loss (Training Loss and Validation Loss).

4.5 Brain Tumor Detection Using CNN and ViT

4.5.1 Architecture CNN

Convolutional Neural Networks (CNNs) are designed specifically for processing and analyzing images as input data. The architecture of a CNN is tailored to effectively handle the unique characteristics and features present in images. By leveraging specialized layers and operations, CNNs are optimized to extract and learn meaningful

patterns, structures, and spatial relationships from image data. This specialized focus on image processing allows CNNs to excel in image classification tasks and achieve state-of-the-art performance in various computer vision applications.

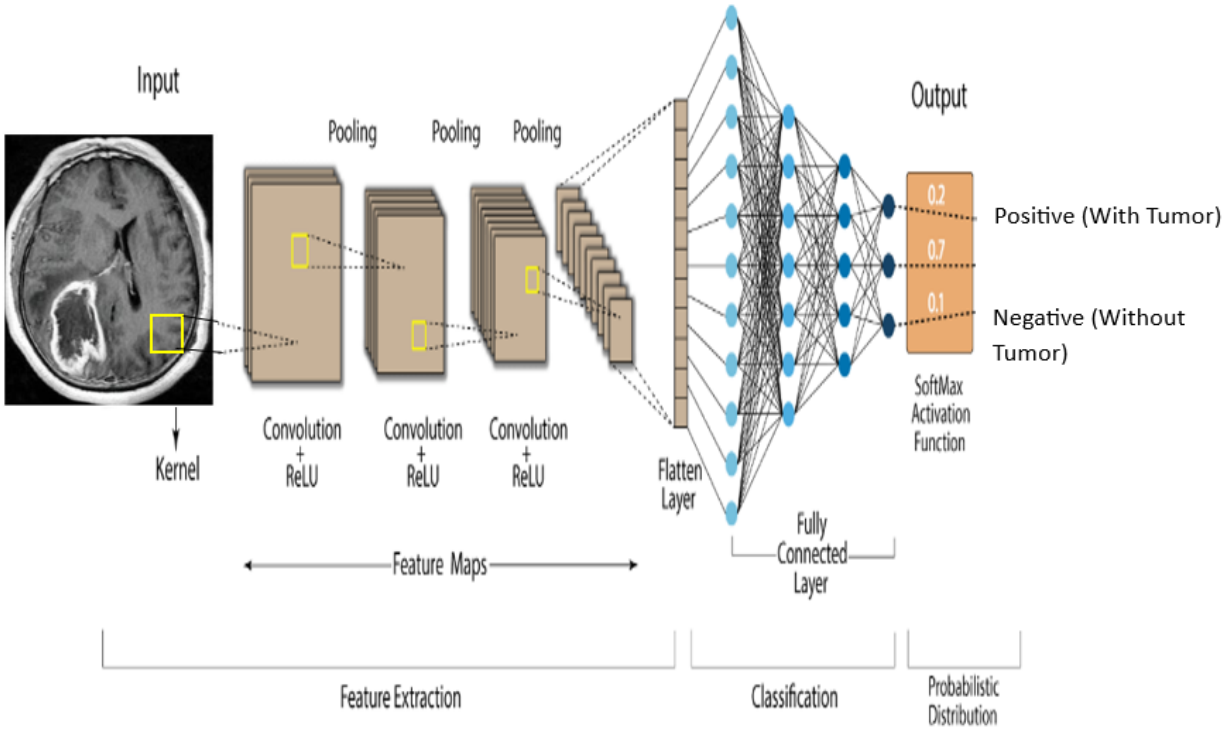

FIGURE 4.3 – An simple CNN architecture

4.5.1.1 General Overview of a CNN Architecture for Image Classification

The pre-processing of the input images is done, where a set of convolutions using the kernel , in addition to the latter, the relu process takes place in which every negative number is replaced by zero in order to remove the complexity of the features, and these multidimensional outputs are considered inputs for the flatten layer that converts Multiple dimensions into a one dimension (1D) that can be processed by fully connected layers that are commonly used for classification tasks Figure 4.3.

4.5.1.2 Visualization of CNN Model

In a convolutional neural network (CNN), there are several main layers used to build the network structure, the figure 4.4 represents a visualization of the CNN model and the changes that occur on the data where the outputs of each layer are inputs for the next layer where we can explain each of these steps as follows :

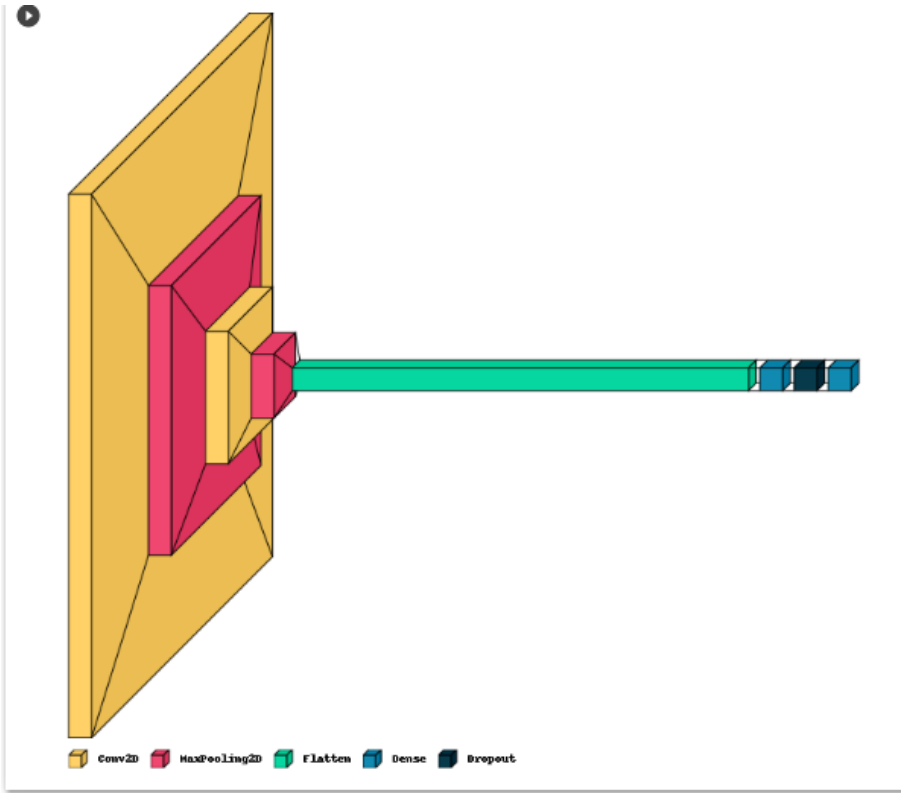

FIGURE 4.4 – Illustration of the treat of the images in a CNN model

— Conv2D :

Conv2D stands for Convolutional 2D. It is the fundamental building block of a CNN. This layer performs the convolution operation, which involves applying a set of filters to the input data. These filters detect patterns and features in the input images. The Conv2D layer applies these filters across the spatial dimensions of the input image, typically in a sliding window fashion, and produces a feature map as output.

— MaxPooling2D :

MaxPooling2D is a pooling operation commonly used in CNNs to reduce the spatial dimensions (width and height) of the input feature maps while retaining the most important information. MaxPooling2D divides the input feature map into non-overlapping rectangular regions and takes the maximum value

within each region, thereby downsampling the feature map. It helps in reducing the computational complexity and making the model more robust to variations in the input.

— Flatten :

The Flatten layer is used to convert the multi-dimensional feature maps into a one-dimensional vector. It essentially flattens the input data, preserving only the spatial structure. This is typically done before passing the features to a fully connected layer (Dense layer) for classification or further processing.

— Dense :

The Dense layer, also known as a fully connected layer, is a standard layer in neural networks. It connects every neuron from the previous layer to every neuron in the current layer. In other words, all the inputs from the previous layer are connected to each neuron in the Dense layer. The Dense layer performs a linear operation followed by an activation function, allowing the network to learn complex patterns and make predictions based on the input features.

— Dropout :

Dropout is a regularization technique used to prevent overfitting in neural networks. It randomly sets a fraction of input units to 0 during training, which helps to reduce the interdependencies between neurons and forces the network to learn more robust and generalized representations. By dropping out some neurons during training, Dropout helps in improving the network's generalization ability and reduces the risk of overfitting.

These layers are commonly used in CNN architectures to extract and learn features from input data, reduce spatial dimensions, and perform classification or regression tasks. The specific arrangement and combination of these layers depend on the problem domain and the desired architecture of the CNN.

4.5.2 ViT Architecture

Figure 4.5 presents an overview of the design architecture of ViT.

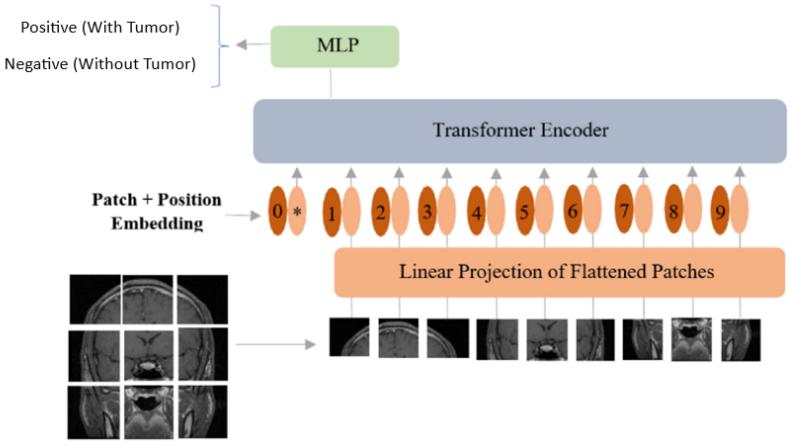

FIGURE 4.5 – Overview of the vision Transformer used in classification

The ViT model divides the input image into patches. These patches are then converted into a sequence of 1D patch embeddings. This sequence is passed through the Transformer encoder, where self-attention modules are employed to calculate a weighted sum of the outputs from each hidden layer, taking into account the relationships between different patches. By using this approach, Transformers are able to learn global dependencies in the input images.

4.5.3 Visualization of ViT Model

The Vision Transformer (ViT) module consists of several steps that are involved in processing an input image and extracting meaningful features. Here are the key steps involved in a typical ViT module As shown in the figure 4.6 :

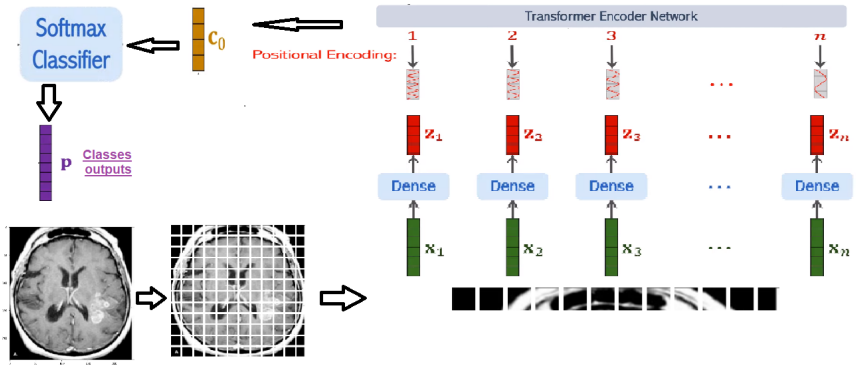

FIGURE 4.6 – Illustration ot the to treat of the images in a ViT model

— Input Image and Patching

The input image is divided into smaller, non-overlapping patches. Each patch represents a local region of the image. The size of the patches is typically predetermined.

The patches are flattened and treated as a sequence of vectors, which can be seen as the input tokens for the subsequent Transformer layers.

— Token Embedding

Each input patch is transformed into a token embedding using an initial linear projection. This projection maps the patch's spatial information into a latent vector space.

— Position Embedding : Position embeddings are added to the token embeddings. Position embeddings capture the relative spatial position information of the patches. They help the model understand the spatial relationships between the patches.

— Transformer Encoder

  — The Transformer encoder consists of multiple layers. Each layer is composed of self-attention mechanisms and feed-forward neural networks.

  — Self-Attention : Self-attention allows the model to capture relationships between different patches by attending to all other patches. It assigns weights to the patches based on their relevance to each other.

  — Feed-Forward Networks : Feed-forward neural networks process the outputs of the self-attention layers. They apply non-linear transformations to the token embeddings to capture more complex patterns and higher-level features.

— Classification Head

  — The output of the last Transformer layer is typically used for classification. It is passed through a classification head, which consists of one or more fully connected layers and an output layer with the desired number of classes.

  — The classification head transforms the encoded features into class probabilities using activation functions such as softmax.

During training, the parameters of the ViT module are updated iteratively to minimize the loss and improve the model's performance on the training data. Fine-tuning involves training the ViT module on a larger dataset or continuing the training on a specific task to improve its performance.

In summary , By going through these steps, the ViT module can process an input image, extract meaningful features using self-attention and feed-forward networks, and perform classification tasks by mapping the features to class probabilities using the classification head.

4.5.4 Models

We employed with pre-trained CNN networks for the classification task, specifically using brain tumor datasets. We then compared the results obtained from the CNN-based architecture with those from the ViT-

based architecture .

4.5.5 Fine-tuning Details

We perform a tow-classes classification task of detection tumor (with tumor and no tumor). We use datasets to 256 images medical for MRI. All experiments is done on the fixed training and testing dataset for comparison purposes between images Contains tumor and images does not contain tumor by using CNN model and ViT model and calculates accuracy and loss for (training and validation) task for the models.

4.6 Traind Model

The algorithm figure 4.7 describes the stages of image processing within a CNN model :

Algorithem 1: CNN algorithem

1. **Input:** Input image: X with dimensions (batch_size, input_height, input_width, input_channels)

   Input height and width: **pic_size**

   Number of input channels: 3 (RGB color channels)

   Epoch_numer

2. **output:** probability_value_binary;

repeat epoch ← 1

 # Convolution operation

conv1 = convolution(X, K1) + B1

 # ReLU activation

output1 = relu(conv1)

output2 = max_pooling(output1, pool_size=(2, 2))

 # Convolution operation

conv2 = convolution(output2, K2) + B2

 # ReLU activation

output3 = relu(conv2)

output4 = max_pooling(output3, pool_size=(2, 2))

flattened = flatten(output4)

 # Linear transformation

linear = dot(flattened, W) + b

 # ReLU activation

output5 = relu(linear)

output6 = dropout(output5, dropout_rate=0.5)

 # Linear transformation

linear_final = dot(output6, W_final) + b_final

 # Sigmoid activation

output_final = sigmoid(linear_final)

show(probability_value_binary)

until epoch <= epochs_length ;

FIGURE 4.7 – CNN model classification

4.6.1 CNN Model Classification

— The Training Phase

These equations represent a forward pass of the CNN model, with the output of each layer serving as the input for the subsequent layer. Specific activation functions, aggregation processes, and organizing techniques are applied. We apply this model in the training phase to all training data for a specified number of times, as it calculates the probability that the input belongs to the positive group by returning a value between 0 and 1 , to summarize, the output of this model is a probability score indicating the likelihood of each input sample belonging to the positive or negative class in a binary classification task.

— The Validation Phase

During training, the model will adjust its internal parameters to reduce the difference between the expected output and the actual labels. The optimization is performed using techniques such as backdiffusion and gradient ratios. The model will continue to train for a set number of periods, and evaluate its performance on the validation data set after each epoch.

4.6.2 Visual Transformer Model Classification

The algorithm figure4.8 describes the stages of image processing within a ViT model :

— The Training Phase

During the training phase of the ViT algorithm, the model is trained on a labeled dataset to learn the patterns and features in the data. The process generally involves the following steps :

Training loop : The training loop consists of iterating through the dataset in batches. For each batch, the model performs a forward pass through the model, computes the loss between the expected output and the ground truth labels, and updates the model's parameters using backpropagation. This is usually done using the model.train_on_batch() function. Progress Monitor : Optionally, the model prints training progress after each batch or epoch. This may include display of the current age, batch number, loss and accuracy.

Iteration : The model repeats the training loop for a set number of periods, allowing the model to learn from the data repeatedly. And we adjust hyperparameters, such as learning rate and batch size, as needed.

— The Validation Phase

Once the model has been trained, the validation phase is performed to evaluate the model's performance on unseen data. This helps assess how well the model generalizes to new examples. The validation phase typically involves the following steps :

Evaluation : We use the trained model to make predictions on the validation dataset. We calculate loss and accuracy metrics by comparing the expected output with the ground truth labels. This is usually done using the model.evaluate() function.

Results presentation : We print or log validation metrics, such as validation loss and accuracy, to evaluate the performance of the model on the unseen data.

By following this training and validation process, we can iteratively train the ViT model on the training dataset and evaluate its performance on the validation dataset. This helps to optimize the model parameters, adjust the hyperparameters, and ensure that the model generalizes well to new, unseen examples.

Algorithem 2: VIT algorithem

Input : Image X ;patch_size ; num_epochs ; num_patches ;image_size ;

output : X_final ; X_final_dimension;

image_size (X)← (B ‹H ‹W ‹C) ;

 # where B is the batch size, H and W are the height and width of the image, and C is the number of channels.

num_patches(X) ← P x P;

image_size (X)← (B ‹N ‹P ‹P ‹C) ;

image_size (X)← (B * N ‹P ‹P ‹C) ; **# Reshape X**

X_reshaped ← (B*N, P, P, C) ;

num_layer ← 1 ;

Repeat

num_layer ← num_layer + 1 ;

Layer Normalization

X_norm = LayerNormalization(X_in) ; **#X_in be the input tensor to the current layer.**

 # Multi-Head Self-Attention (linear projections)

Q = X_norm @ W_q ;

K = X_norm @ W_k;

V = X_norm @ W_v ;

attention = softmax((Q @ K^T) / sqrt(d_k)) @ V; **# scaled dot-product attention**

attention = Dropout(attention); **# optional dropout regularization**

X_attention = LayerNormalization(X_in + Attention) **# MLP (Feed-Forward)**

FF = X_attention @ W_1 + b_1 **# (first linear transformation)**

FF = activation(FF) **#(e.g., ReLU)**

FF = Dropout(FF) **#(optional dropout regularization)**

X_out = LayerNormalization(X_attention + FF) ; **# (residual connection)**

until num_layer <= vit_encoder_layers ;

X_final_dimension← (B*N, F); **# where F is the flattened feature dimension.**

X_final = Dropout(X_final);

Y = X_final @ W + b; **# where W is the weight matrix and b is the bias vector**

Logits = softmax(Y) ; **# Classification**

FIGURE 4.8 – CNN model classification

4.7 Results (Validation Model)

We teste the CNNs and ViT models for models to classify brain tumor images into tow categories : no tumor and with tumor and measure the proportion of the most accurate model in the classification.

These ratios were calculated after entering the training data for the CNN model and the Vit model and specifying the number of epoch equal to 110, meaning that the entered training data will be traversed 110 times, and each time the accuracy and loss of the model are calculated on the training data by means of validation data and labels to calculate validation measures To monitor the progress of the model, and upon completion of the training, the model is evaluated on the test data set, and we obtained these results

4.7.1 Convolutional Neural Network (CNN) Model

The classification accuracy (Acc) and loss results for different CNN model on the brain tumor dataset are reported. based on the CNN model has the results following in Table 4.1

Evaluation	**Model (CNN)**
Acc	70.59 %
Loss	2.23 %

TABLE 4.1 – Results of CNN models

4.7.2 Visual Transformer (ViT) Model

Table 4.2 shows the classification results for ViT model

Evaluation	**Model (VIT)**
Acc	80.39 %
Loss	0.53 %

TABLE 4.2 – Results of CNN models

*Note : As for these results, they are not fixed, so they can improve more or less according to the extent of learning and training this model.

4.7.3 Models Evaluation

4.7.3.1 CNN Model

plan The graph shows the results and accuracy and loss measures of a CNN model progression 4.9 :

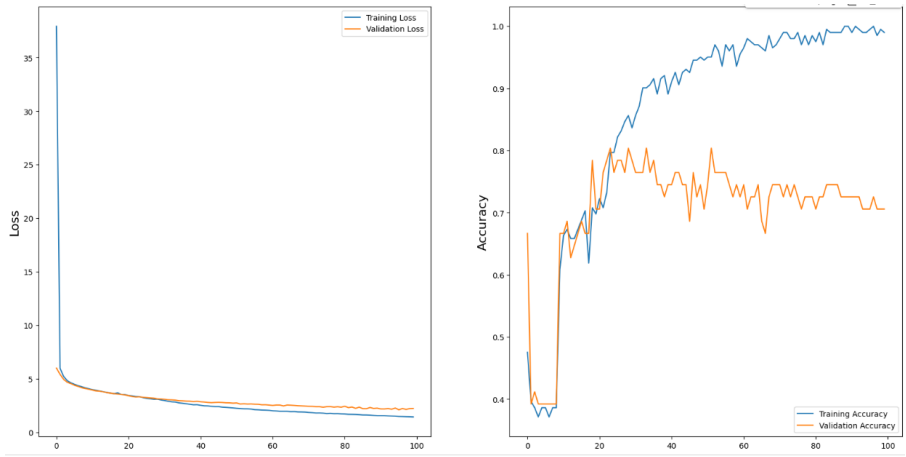

FIGURE 4.9 – CNN model performance evaluation schema

4.7.3.2 ViT Model

plan figure 4.10 The graph shows the results and accuracy and loss measures of a VIT model progression :

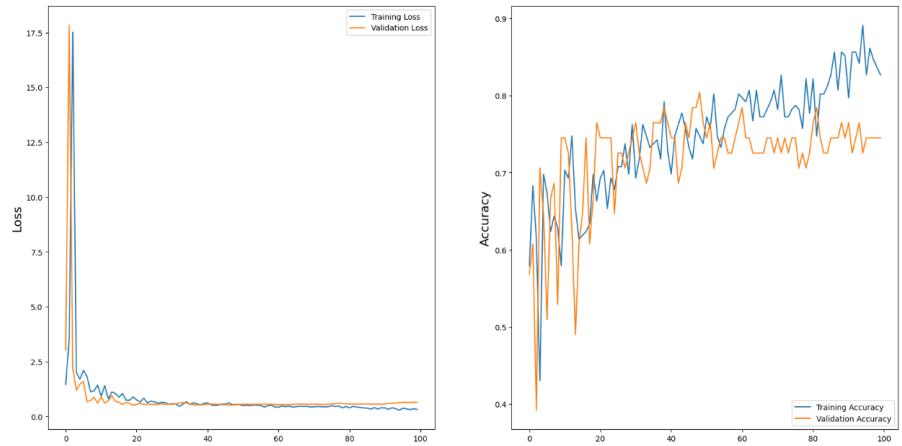

FIGURE 4.10 – ViT model performance evaluation schema

In this evaluation we relied on on two metrics which are training loss and validation loss and the accuracy of each of them on the data.

— Training Loss

This indicator indicates the suitability and learning of the model on the training data, and it is calculated by the average of the loss values obtained from the batch.

— Validation Loss

This measure is an evaluation of the efficiency of the model to generalize on other data that has not been recognized before. It is also an indicator for judging the model whether it is appropriate or inappropriate. The lower the validation loss percentage, the closer the results are to the expected and correct results.

— Training Accuracy

Training accuracy refers to the extent to which the model improves its learning performance on the training data assigned to it. During the training process, the model parameters are updated to reduce training loss and improve its ability to make accurate predictions. As we notice here in the two charts,

there is synchronization in training accuracy and loss. That is, when the training accuracy increases, the model becomes more appropriate and correct.

— Validation Accuracy

The validation accuracy index indicates how well the model has learned data that it has not seen before and how accurate it is in correcting the answer. The purpose of using the validation set is to provide an unbiased estimate of the model's performance on unseen data and to assess its ability to generalize Overfitting refers to the success of the model on training data and its failure on other data that it did not see Whereas, underfitting refers to a situation where the model fails to capture the underlying patterns in the data.

4.8 Discussions

The disparities between CNNs and Vision Transformers are numerous and primarily stem from architectural distinctions, where CNNs are built on a hierarchical structure consisting of convolutional layers followed by pooling layers, which help capture local features and progressively learn more abstract representations. Vision Transformers, on the other hand, are based on the self-attention mechanism, where every image patch attends to all other patches to capture global relationships.. In fact, CNNs yield remarkable outcomes even with smaller-scale training data compared to the extensive data requirements of Vision Transformers. This divergence in behavior seems to arise from the existence of certain inductive biases in CNNs, which may be leveraged by these networks to swiftly comprehend the distinctive aspects of analyzed images. However, these biases also impose limitations and increase the complexity of understanding global relationships.

On the contrary, Vision Transformers are devoid of these biases, enabling them to capture broader and comprehensive relationships. Nevertheless, this advantage comes at the expense of more challenging training in terms of data. Vision Transformers have also demonstrated their efficacy in handling image distortions like adversarial spots or permutations. However, it is not always prudent to favor one architecture over the other, as exceptional results have been achieved in numerous computer vision tasks by employing hybrid architectures that combine convolutional layers and vision adapters.

In addition, the converters have proven their efficiency in computer vision and their ability to overcome CNN in model accuracy, and in the future they can be highly relied upon in the field of image classification through their multiple characteristics and their ability to self-attention with multiple heads.

4.9 The Beginning of our Search

First of all, we started from the idea of merging transformers with CNNs to benefit from the properties of convolutional CNNs while benefiting from the characteristics of transformers with segmenting the patch-like image and the multi head self attention, the Feed Forward network and apply them on a dataset (Brats19) composed of 1677 files each file contains 4 diffirent segmentation and the type of image (NII) when we start with the processing of code in to finish the part of pre-processing and when we start the processing phase we faced hardware problems (GPU) and complexity problems about the dimensions of processed matrices.

these efforts for more than two months of daily work and finally, we decide to change this idea by preparing 2 models CNN and ViT and compare their performance

4.10 Conclusion

In this study, we explored the application of Convolutional Neural Networks (CNN) and visual Transformers (ViT) for brain tumor detection, harnessing the power of deep learning and advanced image analysis techniques. Our goal was to Ensure the most effective model that can aid in the detection of brain tumors and not to make mistakes, leading to improved patient outcomes.

Through the implementation of CNN and ViT architectures, we demonstrated the potential of these models and extract the properties for each model in brain images. By combining the strengths of CNNs and ViTs, we aimed to enhance the detection accuracy and robustness of our modle.

We utilized a comprehensive dataset of brain images, encompassing both tumor and non-tumor samples, for training and evaluation. The training process involved feeding the images through the network, optimizing the model parameters. The model was then evaluated using a separate test dataset, employing metrics such as accuracy, loss.

Our experimental results showcased promising performance in brain tumor detection using CNN and ViT. The architecture VIT effectively captured important features and patterns associated with brain tumors, enabling accurate predictions. We observed high accuracy and favorable evaluation metrics.

However, it is important to note that there are several challenges and considerations in the field of brain tumor detection. The availability of large and diverse datasets, ensuring interpretability of the models, and addressing class imbalance issues are among the key areas for future research.

In conclusion, our study highlights the efficacy of CNN and ViT architectures in brain tumor detection. We concluded that the results of the ViT model are better and more accurate than the CNN model leading to improved accuracy and performance. By advancing the field of medical image analysis, particularly in brain tumor detection, we can to early diagnosis without mistakes, timely interventions, and ultimately, better patient outcomes.

Conclusion General

In conclusion, the fields of artificial intelligence, computer vision, convolutional neural networks (CNNs), and transformers have all made significant contributions to advancing the capabilities of visual data analysis and processing.

Artificial intelligence has revolutionized various industries by enabling machines to mimic human intelligence and perform complex tasks. In computer vision, AI techniques have played a crucial role in interpreting and understanding visual data, allowing machines to recognize objects, detect patterns, and make informed decisions based on visual information.

CNNs have emerged as a powerful deep learning architecture specifically designed for computer vision tasks. Their hierarchical structure, featuring convolutional and pooling layers, enables automatic feature extraction from images, leading to remarkable performance in tasks such as image classification, object detection, and image segmentation. CNNs have greatly advanced the field of computer vision, providing robust and efficient solutions for analyzing visual data.

More recently, transformers have gained significant attention in the field of computer vision. Initially popularized in natural language processing, transformers have shown great potential in visual tasks as well. Transformers excel in capturing long-range dependencies and modeling global context, making them suitable for tasks like image captioning, image generation, and visual question answering. Their self-attention mechanism allows them to analyze relationships between different elements in an image, providing a holistic understanding of the visual content and This is the secret of its superiority over CNN in terms of accuracy. For this in this thesis we focused on the comparison of pre-trained CNN and visual transformer (ViT) architectures for a brain tumor classification task using medical MRI datasets.

After training both the CNN and ViT models on the same dataset, the results show that the performance of the visual transformers surpasses CNNs on various evaluation metrics, including accuracy and loss. This indicates that the ViT architecture is better suited for the specific challenges of the brain tumor classification task.

Looking ahead, the integration of CNNs, transformers, and artificial intelligence will continue to drive advancements in computer vision. The ongoing research and development efforts in these areas hold great promise for tackling more complex visual tasks, such as video understanding, 3D scene reconstruction, and visual reasoning. Additionally, the interpretability and explainability of these models will be critical for building trust and understanding their decision-making processes, paving the way for responsible and ethical deployment of computer vision systems.

Bibliographie

[1] "Machine learning in the context of ai," https://levity.ai/blog/how-do-machines-learn, accessed APRIL 16, 2023.

[2] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis : A review. arxiv 2022," *arXiv preprint arXiv :2202.12165*.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available : http://arxiv.org/abs/1706.03762

[4] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *CoRR*, vol. abs/2111.06091, 2021. [Online]. Available : https://arxiv.org/abs/2111.06091

[5] "Deep learning and machine learning," https://levity.ai/blog/difference-machine-learning-deep-learning#:~:text=Machine%20Learning%20means%20computers%20learning,documents%2C%20images%2C%20and%20text., accessed APRIL 16, 2023.

[6] "Ibm," https://www.ibm.com/topics/machine-learning, accessed APRIL 15, 2023.

[7] D. Kitaguchi, N. Takeshita, H. Hasegawa, and M. Ito, "Artificial intelligence [U+2010] based computer vision in surgery : Recent advances and future perspectives," *Annals of gastroenterological surgery*, vol. 6, no. 1, pp. 29–36, 2022.

[8] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—a contemplative retrospection," *Engineering applications of artificial intelligence*, vol. 122, 2023.

[9] R. M. Hussien, K. Q. Al-Jubouri, M. A. Gburi, A. G. Hussein Qahtan, and A. H. Duaa Jaafar, "Computer vision and image processing the challenges and opportunities for new technologies approach : A paper review," *Journal of Physics : Conference Series*, vol. 1973, no. 1, p. 12002, 2021.

[10] E. Elyan, P. Vuttipittayamongkol, P. Johnston, K. Martin, K. McPherson, C. F. Moreno-García, C. Jayne, M. K. Sarker *et al.*, "Computer vision and machine learning for medical image analysis : Recent advances, challenges, and way forward," *Artificial Intelligence Surgery*, vol. 2, no. 1, pp. 24–45, 2022.

[11] "Computer aided diagnosis - medical image analysis techniques," https://www.intechopen.com/chapters/56615, accessed APRIL 15, 2023.

[12] R. A. Castellino, "Computer aided detection (cad) : an overview," *Cancer Imaging*, vol. 5, no. 1, p. 17, 2005.

[13] M. Firmino, G. Angelo, H. Morais, M. R. Dantas, and R. Valentim, "Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy," *Biomedical engineering online*, vol. 15, no. 1, pp. 1–17, 2016.

[14] "dataset," https://byjus.com/maths/data-sets/?fbclid=IwAR2VlxpHrURRJA6HcgdvSQUKO8eCqMFAJmgqCpJizGBN accessed APRIL 17, 2023.

[15] Y. Xia, "Towards robust deep learning for medical image analysis," Ph.D. dissertation, Johns Hopkins University, 2021.

[16] "viso.ai," https://viso.ai/deep-learning/vision-transformer-vit/, accessed APRIL 17, 2023.

[17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on visual transformer," *arXiv preprint arXiv :2012.12556*, vol. 2, no. 4, 2020.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words : Transformers for image recognition at scale," 2021.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available : https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[20] S. Tuli, G. Casale, and N. R. Jennings, "Tranad : Deep transformer networks for anomaly detection in multivariate time series data," *CoRR*, vol. abs/2201.07284, 2022. [Online]. Available : https://arxiv.org/abs/2201.07284

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[22] J. Hu, H. Qin, T. Yan, and Y. Zhao, "Corrected bayesian information criterion for stochastic block models," *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1771–1783, 2020.

[23] V. L. L. Thing, "Deepfake detection with deep learning : Convolutional neural networks versus transformers," 2023.

[24] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers : Token-based image representation and processing for computer vision," *arXiv preprint arXiv :2006.03677*, 2020.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020. [Online]. Available : https://arxiv.org/abs/2005.12872

[26] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR : deformable transformers for end-to-end object detection," *CoRR*, vol. abs/2010.04159, 2020. [Online]. Available : https://arxiv.org/abs/2010.04159

[27] Z. Yao, J. Ai, B. Li, and C. Zhang, "Efficient DETR : improving end-to-end object detector with dense prior," *CoRR*, vol. abs/2104.01318, 2021. [Online]. Available : https://arxiv.org/abs/2104.01318

[28] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence : Rethinking transformer in vision through object detection," *CoRR*, vol. abs/2106.00666, 2021. [Online]. Available : https://arxiv.org/abs/2106.00666

[29] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr : Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1601–1610.

[30] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr : Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 619–13 627.

[31] A. Adhikari, A. R. Choudhuri, D. Ghosh, N. Chattopadhyay, and R. Chakraborty, "Breast cancer histopathological image classification using convolutional neural networks," in *Proceedings of International Conference on Innovations in Software Architecture and Computational Systems : ISACS 2021*. Springer, 2021, pp. 183–195.

[32] T. A. Sadoon and M. H. Ali, "Deep learning model for glioma, meningioma and pituitary classification," *Int. J. Adv. Appl. Sci. ISSN*, vol. 2252, no. 8814, p. 8814, 2021.

[33] W. Grisold and A. Grisold, "Cancer around the brain," *Neuro-oncology practice*, vol. 1, no. 1, pp. 13–21, 2014.

[34] M. Kheirollahi, S. Dashti, Z. Khalaj, F. Nazemroaia, and P. Mahzouni, "Brain tumors : Special characters for research and banking," *Advanced biomedical research*, vol. 4, 2015.

[35] J. Lukoff and J. Olmos, "Minimizing medical radiation exposure by incorporating a new radiation "vital sign" into the electronic medical record : quality of care and patient safety," *The Permanente Journal*, vol. 21, 2017.

[36] R. Vankdothu and M. A. Hameed, "Brain tumor mri images identification and classification based on the recurrent convolutional neural network," *Measurement : Sensors*, vol. 24, p. 100412, 2022.

[37] T. Rahman and M. S. Islam, "Mri brain tumor detection and classification using parallel deep convolutional neural networks," *Measurement : Sensors*, vol. 26, p. 100694, 2023.

[38] S. Irsheidat and R. Duwairi, "Brain tumor detection using artificial convolutional neural networks," in *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2020, pp. 197–203.

[39] Y. Jiang, J. Hou, X. Xiao, and H. Deng, "A brain tumor segmentation new method based on statistical thresholding and multiscale cnn," in *Intelligent Computing Methodologies : 14th International Conference, ICIC 2018, Wuhan, China, August 15-18, 2018, Proceedings, Part III 14*. Springer, 2018, pp. 235–245.

[40] D. Liu, D. Zhang, Y. Song, F. Zhang, L. J. O'Donnell, and W. Cai, "3d large kernel anisotropic network for brain tumor segmentation," in *Neural Information Processing : 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part VII 25*. Springer, 2018, pp. 444–454.

[41] Y. Suter, A. Jungo, M. Rebsamen, U. Knecht, E. Herrmann, R. Wiest, and M. Reyes, "Deep learning versus classical regression for brain tumor patient survival prediction," in *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries : 4th International Workshop, BrainLes 2018, Held in Conjunction*

with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer, 2019, pp. 429–440.

[42] Y. Bhanothu, A. Kamalakannan, and G. Rajamanickam, "Detection and classification of brain tumor in mri images using deep convolutional network," in *2020 6th international conference on advanced computing and communication systems (ICACCS)*. IEEE, 2020, pp. 248–252.

[43] A. Marefat, M. Marefat, J. Hassannataj Joloudari, M. A. Nematollahi, and R. Lashgari, "Cctcovid : Covid-19 detection from chest x-ray images using compact convolutional transformers," *Frontiers in Public Health*, vol. 11, p. 581, 2023.

[44] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet : Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv :2102.04306*, 2021.

[45] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, "Transclaw u-net : Claw u-net with transformers for medical image segmentation," *arXiv preprint arXiv :2107.05188*, 2021.

[46] G. Xu, X. Wu, X. Zhang, and X. He, "Levit-unet : Make faster encoders with transformer for medical image segmentation," *arXiv preprint arXiv :2107.08623*, 2021.

[47] Z. Fu, J. Zhang, R. Luo, Y. Sun, D. Deng, and L. Xia, "Tf-unet : An automatic cardiac mri image segmentation method," *Mathematical Biosciences and Engineering*, vol. 19, no. 5, pp. 5207–5222, 2022.

[48] C. You, R. Zhao, F. Liu, S. Chinchali, U. Topcu, L. Staib, and J. S. Duncan, "Class-aware generative adversarial transformers for medical image segmentation," *arXiv preprint arXiv :2201.10737*, 2022.

[49] Y. Zhang, H. Liu, and Q. Hu, "Transfuse : Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 14–24.

[50] A. M. Al-Hejri, R. M. Al-Tam, M. Fazea, A. H. Sable, S. Lee, and M. A. Al-Antari, "Etecadx : Ensemble self-attention transformer encoder for breast cancer diagnosis using full-field digital x-ray breast images," *Diagnostics*, vol. 13, no. 1, p. 89, 2022.

[51] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, "Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus," *arXiv preprint arXiv :2103.07055*, 2021.

[52] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, and E. Vezzetti, "Vision transformer for femur fracture classification," *Injury*, vol. 53, no. 7, pp. 2625–2634, 2022.

[53] G. van Tulder, Y. Tong, and E. Marchiori, "Multi-view analysis of unregistered medical images using cross-view transformers," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 104–113.

[54] G. S. S. Costa, A. C. Paiva, G. B. Junior, and M. M. Ferreira, "Covid-19 automatic diagnosis with ct images using the novel transformer architecture," in *Anais do XXI simpósio brasileiro de computação aplicada à saúde*. SBC, 2021, pp. 293–301.

[55] X. Gao, Y. Qian, and A. Gao, "Covid-vit : Classification of covid-19 from ct chest images based on vision transformer models," *arXiv preprint arXiv :2107.01682*, 2021.

[56] L. Zhang and Y. Wen, "Mia-cov19d : a transformer-based framework for covid19 classification in chest cts," in *Proceeding of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021, pp. 513–8.

[57] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE transactions on medical imaging*, vol. 41, no. 1, pp. 213–224, 2021.

[58] B.-H. Kim, J. C. Ye, and J.-J. Kim, "Learning dynamic graph representation of brain connectome with spatio-temporal attention," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4314–4327, 2021.

[59] S. Wang, Z. Zhuang, K. Xuan, D. Qian, Z. Xue, J. Xu, Y. Liu, Y. Chai, L. Zhang, Q. Wang *et al.*, "3dmet : 3d medical image transformer for knee cartilage defect assessment," in *Machine Learning in Medical Imaging : 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 347–355.