**DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA**

**MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH**

**FACULTY OF MATHEMATICS AND INFORMATICS**
**DEPARTMENT OF COMPUTER SCIENCE**
**N° :……28……..**


**DOMAINE: Mathematics and Informatics**

**FILIERE: Informatics**
**OPTION: INFORMATION SYSTEMS AND SOFTWARE ENGINEERING**
**……..**

**A Dissertation in Fulfillment**
**For the Requirements of the Degree of Master**
**In Computer Science**
**By:**

Khoudour aya nor elhouda

Nasri nesrine

**Entitled**

# Application of ensemble Learning in visual question-answering

**Presented publicly to the jury:**

| | | |
|---|---|---|
| Hichem Debbi | University of M'sila | Superviser |
| Meliouh Amel | University of M'sila | President |
| Hamani Said | University of M'sila | Examiner |

**Academic Year: 2022/2023**

**DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA**

**MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH**

**FACULTY OF MATHEMATICS AND INFORMATICS**
**DEPARTMENT OF COMPUTER SCIENCE**
**N°:……28……..**


**DOMAINE: Mathematics and Informatics**
**FILIERE: Informatics**
**OPTION: INFORMATION SYSTEMS AND SOFTWARE ENGINEERING**

**A Dissertation in Fulfillment**

**For the Requirements of the Degree of Master**

**In Computer Science**

**By:**

**Khoudour aya nor elhouda**

**Nasri nesrine**

**Entitled**

# Application of ensemble Learning in visual question-answering

## Presented publicly to the jury:

| | | |
|---|---|---|
| Hichem Debbi | University of M'sila | Superviser |
| Meliouh Amel | University of M'sila | President |
| Hamani Said | University of M'sila | Examiner |

## Academic Year: 2022/2023

## *Dedication*

Mom and Dad, my brothers and all my family, I present to you my graduation with all his efforts and perseverance, I present to you who I have drawn the strength from until I reach this stage.

*Nasri Nesrine*

## *Dedication*

I dedicate this work to the two most precious people to my heart, and who was one of the most important reasons why I succeeded in life and got to where I am today, my father and mother.

And my whole family.

*Khoudour aya nor elhouda*

## ACKNOWLEDGEMENT

First of all, we say thanks to God FOR agreeing us to do this work and providing us with health and wellness so that we can endure the problems and difficulties for writing this memo and continue until the end.

We extend our gratitude to Supervisor Hicham Debbi for his guidance and assistance and provide us with his expertise and knowledge throughout this work.

Finally, we would like to express our deep gratitude to our family, friends, and all those who have contributed to us with the support and encouragement that has always been present with us.

Thank you so much.

## *Table of Content*

## *List of tables:*



## *List of figures:*

## *List of abbreviations:*

# General Introduction

# General Introduction

In recent years, Artificial intelligence especially deep learning has rapidly advanced and has found applications in the industrial field and many various domains. It involves development of computer systems capable of performing tasks that usually require human intelligence. AI includes various sub-fields and techniques, such as Computer Vision (CV), which enables a machine to understand visual data, and Natural language processing (NLP), which enables machines to understand, interpret, and generate human language in various forms, such as text or speech. Deep learning has significantly advanced the fields of NLP and CV, this advance created increased interest in the integration of vision and language.

Vision and language is a domain that combines CV and NLP fields and it aims to enable machines to comprehend and generate visual content through the use of natural language. VL involves various tasks such as image captioning, where machines generate textual descriptions or captions for given images, and Visual Question Answering (VQA), when there is a question about an image, and the answer is from the image.

VQA it aims to build systems that enable us to understand and answer questions about visual content, such as images or videos. VQA processes both images and questions input using the CV and NLP technique, to predict the answer. A lot of research and developers have contributed to building various VQA algorithms and approaches, also Datasets to generate VQA model.

VQA task has a widespread interest, which makes the recent VQA research work to improve the performance of VQA models to better understand the image and questions and predict the correct answer. The top performing VQA systems are ensembles of neural networks that perform substantially better than any of the underlying individual models [1]. VQA is a challenging task; it necessitates the model to understand both the visual content of an image and the textual content of a question and must integrate information from both visual and textual modalities to produce precise and reliable answers. Ensemble learning is a technique that combines the predictions of multiple models to improve the overall accuracy of the individual model. By combining the predictions of different architecture models, it increases the diversity of the predictions and makes the system more robust to predict more accurate answers.

In this dissertation, our objective is to explore and apply ensemble models to the VQA problem. Through it, we aim to overcome the limitations of single-model approaches and advance the state-of-the-art in this domain. Our approach presents enhancing diversity on the side of vision input within an ensemble model of various ResNet architectures. By utilizing a

diversity of ResNet architecture, we can capture more of visual characteristics and increase the overall understanding of the image content.

To evaluate the effectiveness of our ensemble models, we will employ the CLEVR dataset, a widely used benchmark in the VQA domain. Through extensive experimentation and analysis, we compare the results with single-model baselines and provide the advantages and limitations of our approach.

This work is organized into 4 chapters, as a VQA area combines the both CV and NLP therefore the first chapter is about Computer vision, and the second chapter is about Natural language processing. Chapter 3 provides the Visual Question Answering task and recent research and development engagements. In the last chapter, chapter four, we provide the realized work and the obtained results.

# Chapter 1:

# Computer Vision

# Chapter 1: Computer Vision

## 1. Introduction:

Computer vision is study and technology field, which focuses to make computers interpret and understand visual information from digital images or videos. This chapter involves the definition and historical evolution of computer vision, also various applications of this field. In addition, the convolution neural network is the most important network in the field, also the algorithms, computer vision pipeline, and pre-trained models.

## 2. Definition:

Computer vision it is a science of how to make computers "see" and understand the world through images and videos with the ability to take the appropriate actions.

Computer vision is a technical field of artificial intelligence which is about developing technologies, through these technologies the computer understands the images pixel by pixel and extracts the visual data, manages them, and analyzes the results with the help of advanced software.

Computer vision can have a lot of similarities to human vision, but there are big differences between them. For a human, the vision systems consist of an eye that captures images and the brain to process and interprets the image. To emulate this system, the machine needs an alternative to each one of the two components, a sensing replacing the eye and a powerful algorithm to mimic the brain function in interpreting and classifying the image content.

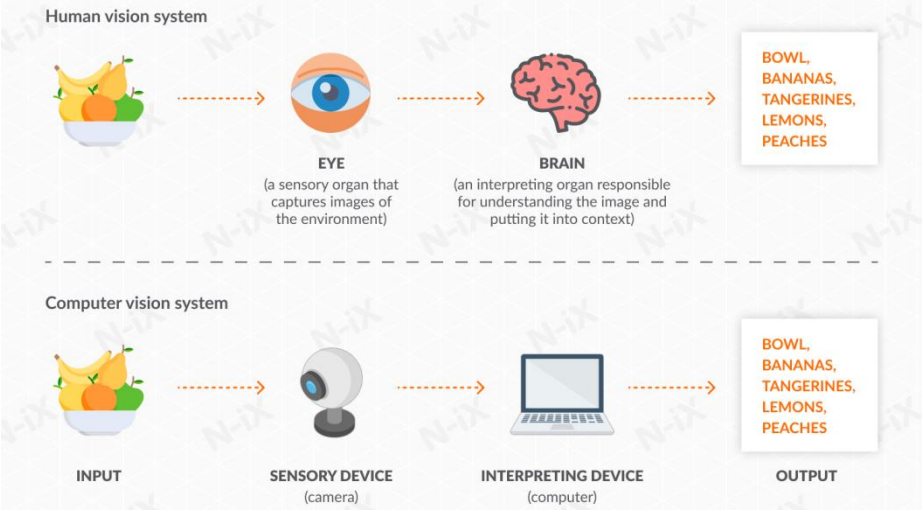



**Figure 1. 1: Human vision system vs Computer vision system [2] .**

## 3. History of Computer Vision:

For about 60 years, engineers and scientists have been working on developing vision systems; the story begins in **1960**, when Computer Vision father Larry Roberts' PhD was discussed, at MIT discussed the possibilities of extracting 3D geometrical information from 2D perspective views of blocks (polyhedral) [3].

- In **1978** David Marr's representational Framework for vision.
- In **1980** Japanese computer scientist Kunihiko Fukushima built the "Neocognitron" network, includes several convolutional layers whose receptive fields had weight vectors.
- A few years later, in **1989** Yann LeCun applied a backpropagation style learning algorithm to Fukushima's convolutional neural network architecture, after working on the project, LeCun released LeNet-5 the first modern convolutional network.
- Around **1999** lots researchers stopped trying to reconstruct object by creating 3D models of them (the path proposed by Marr), and that's why David Low worked on "Object Recognition from Scale-Invariant features".

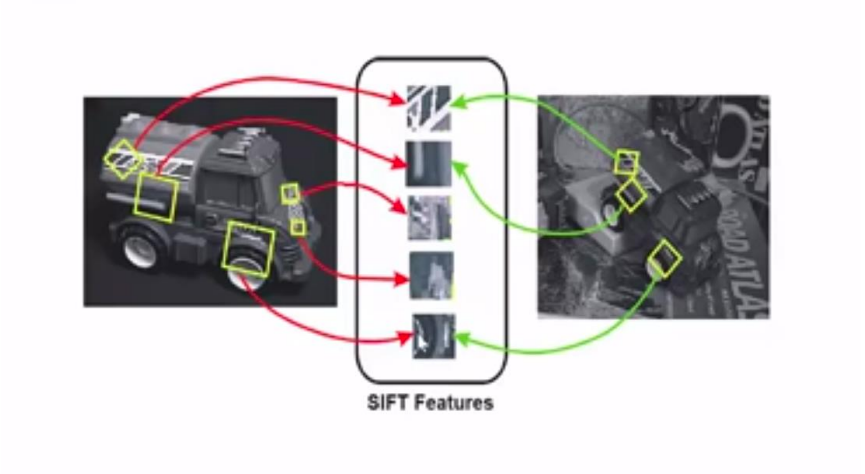



**Figure 1. 2: Object Recognition from Scale-Invariant features [4]**

- The ImageNet Large Scale Visual Recognition Competition (ILSVRC), started in **2010**.ImageNet containing millions of tagged images across various object classes that provided the foundation of CNNs and other deep learning models used today[5].
- Beginning in 2010 appearances deep networks such as Alex-Net (2010), VGG-Net (2015), and Res-Net (2016), they achieved outstanding results in image recognition and classification tasks.

## 4. Computer Vision Application:

Now our daily life is hard without computer vision, it plays a major role in all fields of industry, health, transport, study..... There are many other examples. The most prominent examples [6]:

- Self-Driving Cars: Helping autonomous vehicles navigate the road, detect and perceive objects, and understand their environment.
- Healthcare: Precise diagnoses and early detection of diseases, Medical imaging with high accuracy and quality, nuclear medicine, and so on.
- Defense and security: Facial detection, unmanned military vehicles, weapon defect detection.
- Manufacturing and Retail: Assist in inventory management by monitoring the quantity and quality of products and classifying them, conducting bar code analysis …
- Agriculture: Drone-based crop monitoring and damaged crop detection, automated pesticide spraying, livestock count detection, and smart systems for crop grading and Sorting.


**Figure 1. 3: Some computer vision applications.**

## 5. Convolution Neural Network (CNN):

Convolutional neural network (CNN or conv_net) is a deep learning algorithm and one of the artificial neural network categories. It deals with data that has a grid-like structures such as images, consisting of structured pixels are processed in a grid-like manner. The structure of CNN consists of three layers each one having a different role. The first layer is the Convolution Layer. It is the main element of the CNN structure, consisting of filters also called “kernels” that extract features from the input image.

- The second layer is the pooling layer, which reduces the number of parameters the network needs to learn by inserting the image into the filter and reducing the features in the input, and streamlining the outputs.
- The third layer is fully connected, inside this layer image classification happens in the CNN based on the features extracted in the previous layers.

## 6. Computer Vision Algorithms:

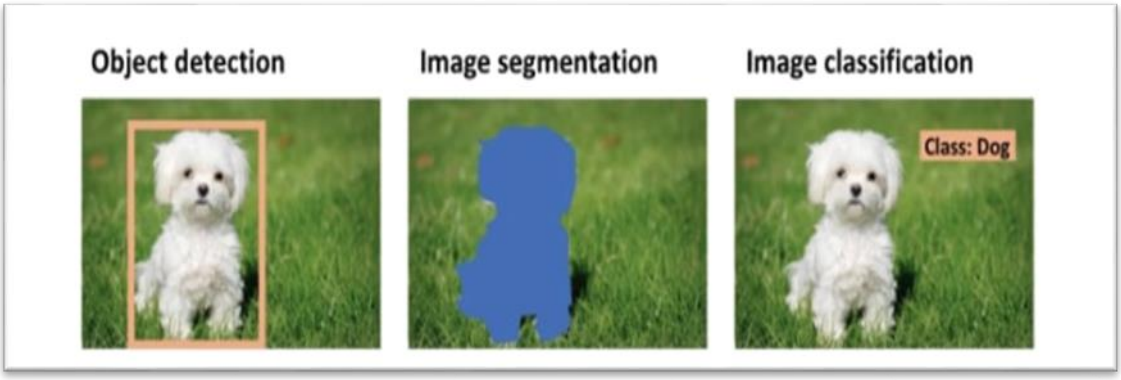



**Figure 1. 4: Object detection Vs Image segmentation vs Image classification.**

Researchers in computer vision aspired to develop algorithms for such visual perception tasks [7] include different method involved in recognizing things in images, the most common algorithms are:

### 6.1 Image classification :

The main objective of this technique refers to a process in computer vision that can classify sets of images into classes using classifier model. For decades, researchers have laid path in developing advanced techniques to improve the classification accuracy [7]. In parallel with the creation of a large-scale image dataset "ImageNet", the classification of images has changed significantly and started to show great performance in classification, this is because of the development of profound deep learning technology.

Before deep learning came out, there were other methods such as designing scale-invariant feature and classifications such as a Support Vector Machine SVM learning algorithm, but all these traditional methods became weak in front of complex natural image features.

Today the classification model based on deep CNN performs much more robustly than old methods. The convolutional Neural Network is a kind of multi-layer neural networks, and is very effective in recognizing optical patterns of pixel images with minimal pre-processing.

### 6.2 Object Detection:

As one of the most important computer vision technique; object detection works to identify objects in an image or video and draws frames around those detected objects, this lets us locate them in a given scene. Object Detection Methods is divided into two approaches, Neural network-based and Non-neural approaches. Before deep learning, object detection was done through classical machine learning techniques, such as: Scale Invariant feature transform (SIFT), these algorithms would define features then using a technique such as SVM to do the classification. Either Neural network approaches or Deep learning-based approaches; is use neural network architecture to detect object without specifically defining features. Object detection has applications in many areas as Retail Use Cases, health care, Industry, Self-driving cars, Security & Safety Use Cases such as Video surveillance, Crowd counting.

### 6.3 Image Segmentation:

Another important track within computer vision is image segmentation, it aims to divide an image into several components based on pixel characteristics to identify the boundary of object to simplify an image and analyze it more efficiently. The result of segmenting is a set of segments has a set of pixels, each of the pixels in a region is similar with respect to any characteristic or computed property [8].Image segmentation is divided into three categories, first the Semantic Segmentation, it classifies each pixel into semantic classes. The algorithm takes an image to categorize similar items in the same class (i.e. all trees types in the tree class, all car types in the vehicle class). Another category is the instance segmentation; it classifies each pixel according to "Instances". Instance segmentation it combines object detection and semantic segmentation. Finally the Panoptic segmentation is a combination of semantic and instance segmentation, In this case, each instance of an object in the image is separate and the identity of the object is provided.

(a) Image (b) Semantic Segmentation

(c) Instance Segmentation (d) Panoptic Segmentation

**Figure 1. 5: Semantic, Instance, Panoptic Segmentation [9] .**

## 7. Computer Vision Pipeline:

We spoke before about the computer vision, consisting of sensing device and interpreting device. Our focus now is on interpreting device component, to take a look at the process of the pipeline, a Computer Vision Pipeline is a series of steps that most computer vision applications will go through [10]. The General Pipeline is consisting of:

### 7.1. Input image:

The image size represented by width and height, for example 24*24 there are 24 pixels horizontally and 24 vertically, this means that there are total 576 (24*24) pixels. Every image consists of a set of pixels; the pixel value represents the intensity of light, 0 represent very dark pixel until 255 is very bright. In Grayscale image we use (0-255) different brightness (gray) levels. Represent by single color plane with 8 bits. There are many types of a color image we will consider the RGB system. In RGB color space the value of the pixel is represent by 24-bit with the amount of its compounds red (R), green (G) and blue (B).

For a computer, the image resembles a 2D matrix of pixel values that represent the intensities in the color spectrum. For color image, it can be represented using three 2D arrays of same size, each color array element has an 8-bit value, and indicates how much red, green or blue there is on a scale of [0, 255].

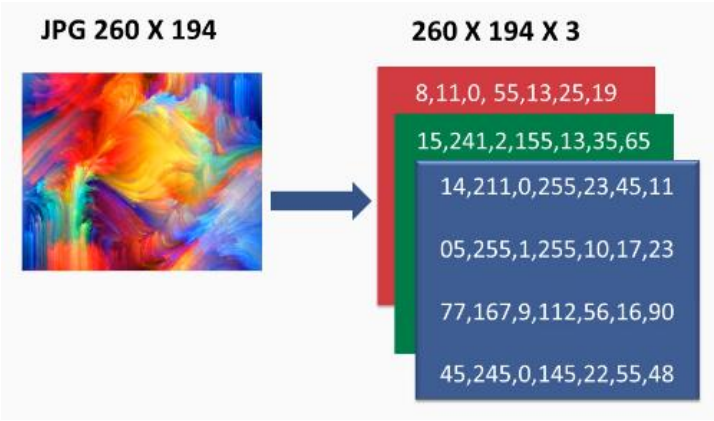



**Figure 1. 6: what we see Vs what the computers see.[11]**

### 7.2. Image pre-processing:

Acquired data is usually disorderly and comes from different sources and must be standardized and cleaned up. Preprocessing is used to perform steps that will reduce complexity and improve the accuracy of the applied algorithm, thus, convert images in a form that would enable a general algorithm to solve it. During image preprocessing, different techniques are used according to the specific application and the desired result; such as rescaling and resizing, gray scaling is converting images to grayscale, also image

augmentation or augmentation techniques includes rotation, translation, scaling, flipping, can be applied to generate input image variations.

The purpose of pre-processing is getting the data ready for training by removing undesirable distortions and improved specific qualities needed for application. An image requires preprocessing before being used for model training and inference, to make the model work correct and provide the target results.

### 7.3. Feature Extraction:

In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon [12]. Feature extraction is a technique to detect features in the image. Through these features, we effectively reduce the quantity of data to be processed while accurately and comprehensively describing the original data set.

Machine learning and deep learning have evolved over the past few years, finding a new method to process images directly through deep learning and automatically extracting feature of the image. This method does not detect the feature points alone, but adaptively finds the feature description of the image according to the task by gradient descent according to the network structure and loss function [13].

Using CNN's architecture, we can easily extract image features. By training, the model learns to extract features through using convolution of the image and filters to generate fixed features, and use those features to train an image classifier, in the last gets final output.

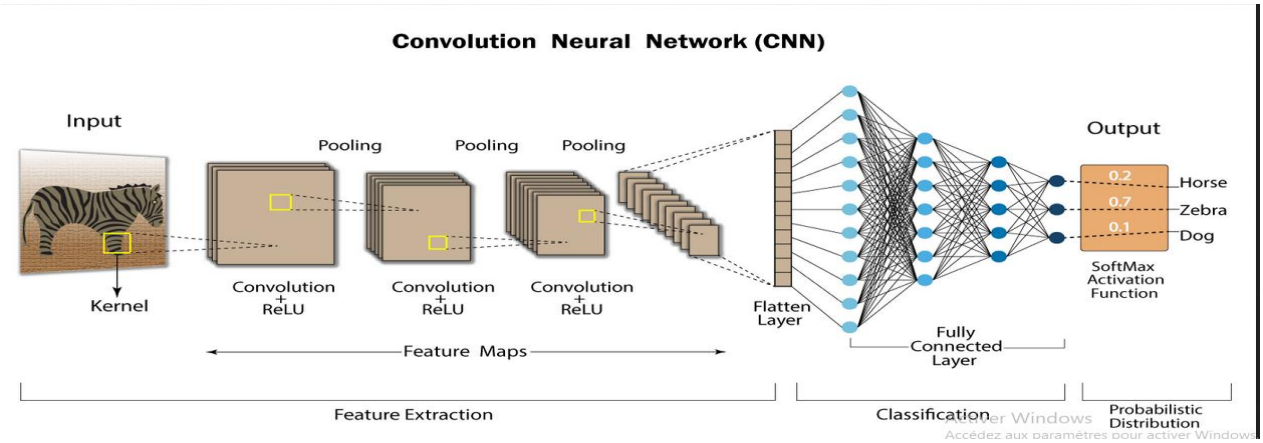



**Figure 1. 7: CNN Model Feature Extraction and Classification [14] .**

## 8. Transfer learning:

Transfer learning [15]is a subfield of artificial intelligence which aims to enable use a pre-trained model, trained on a large dataset, for applying it to a specific task or applications. Here, some several pre-trained models trained on large datasets and are available for computer vision tasks:

**8.1 VGG-Net:**

VGG [16] represents the Visual Geometry Group; it is a group of researchers proposes very deep convolutional network architecture of multiple convolution layers with small filters (3×3). The architecture of VGG it consists of input layer has a fixed-size of 224×224 RGB image, and a stack of convolution layers followed by RELU activation layer. Spatial pooling is carried out by five max-pooling layers with 2×2 pixel, which follow some of the convolutions [16]. VGG-net has 3 fully connected (FC) layers; the two first layers have shape of 1×1×4096. And in the last, it has the soft-max layer. VGG propose different configurations, such as VGG16 the number 16 means a deep neuron network of 16 layers; 13 convolutions layers and 3 fully connected layers. VGG achieved top-5 test accuracy of 92.7% in ImageNet, and can classify the images to 1000 classes.

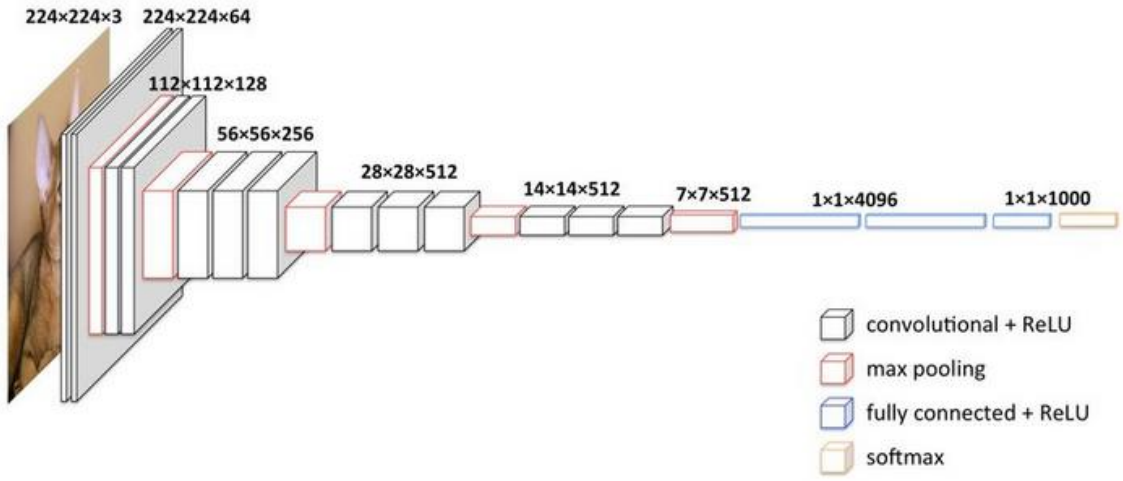



**Figure 1. 8: VGG16 architecture.[17]**

**8.2 Res-NET :**

Residual Network (Res-NET) [18] is a deeper neural network easy to train more than previous architectures. The main building blocks of the ResNet architecture is The Residual Block; where this block consists of multiple 3x3 convolutional layers and ReLU activation functions, batch normalization. The architecture is typically consists of multiple residual blocks, the number of layers blocks based on the specific ResNet configurations (e.g., ResNet-18, ResNet-50, etc.) Connected to the global average pooling and Fully Connected Layer, Softmax activation for classification tasks. Res-NET solves the vanishing gradient problem which usually occurs in very deep network. The problem was mitigated by introducing known as “skip connections”; is mechanisms that directly connect the output of one layer to the input for a short cut, and therefore allow the gradient to flow directly out from the previous layers to the next layers during backpropagation, this accelerates the training process and reduces the vanishing gradient problem. ResNet models trained and evaluated on

the ImageNet 2012 classification dataset, and have been shown to achieve state-of-the-art performance on image classification, object detection, and image segmentation.

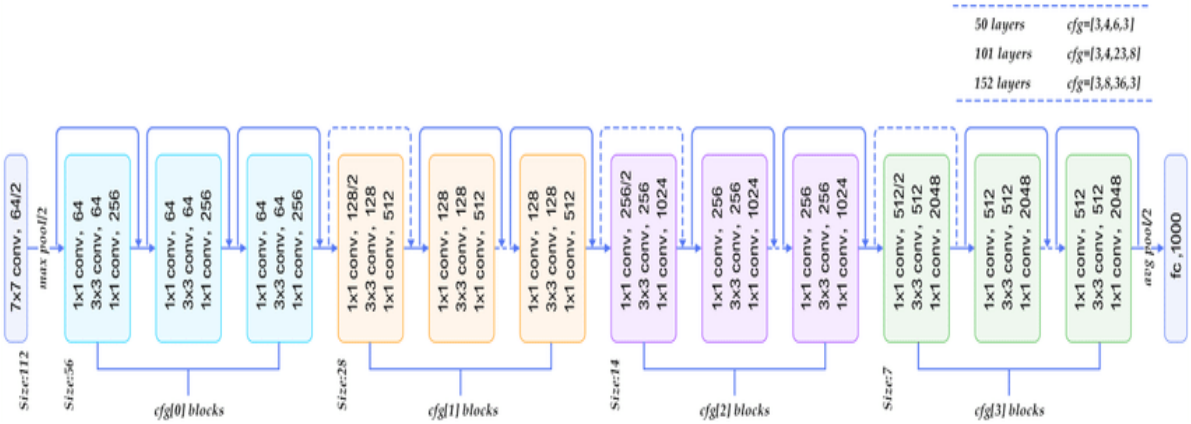



**Figure 1. 9: ResNet architecture. [19]**

## 9. Conclusion:

Computer vision is famous field and widely used on many applications across various industries, including healthcare, automotive, agriculture, robotics, security, entertainment, and more. The advancement of the field of Computer vision provided great breakthroughs in fields such as autonomous driving, medical imaging, augmented reality, and visual search engines, among others.

# *Chapter 2:*

# *Natural Language Processing (NLP)*

## 1. Introduction:

Natural language processing is one of the most important technologies of the information age. It is the point of understanding between the machine and the human. In this chapter we will talk about NLP and see how it works and what techniques it uses to make humans and the machine understand each other's language and see what the results and applications of this technique are at the level of normal life.

## 2. Definition:

Natural language processing is one of the most important technologies in the field of artificial intelligence that gives computers the ability to deal with human language, spoken or written with understanding and interpreting it[20], thus facilitating the process of communication between computer and human.

## 3. NLP Component:

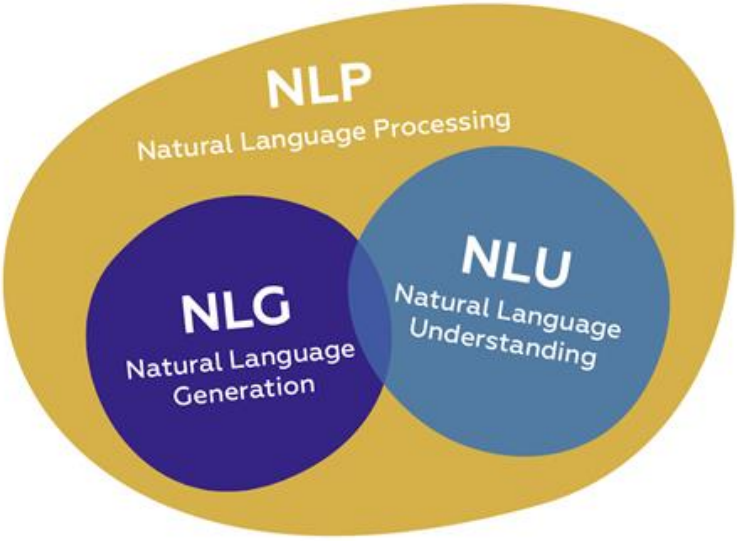



**Figure 2. 1: Natural Language Processing (NLP) Component.**

### 3.1. Natural Language Understanding (NLU):

Natural Language Understanding is one of the branches of artificial intelligence responsible for making human language understandable in devices, analyzing natural language and interpreting key concepts and terminology as well as emotions and transforming them into a form that devices can understand [21].

### 3.2. Natural Language Generation (NLG):

Natural Language Generation Another branch of AI in the field of language processing, lies in its role in generating an understanding language for humans, by using algorithms and

techniques that convert computerized information and data into a spoken or written natural language[22].

## 4. Applications of Natural language processing:

All text-based applications are candidates for natural language processing, but the most important areas in which natural language processing has progressed significantly are:

### 4.1 Chat-Bot:

A Chat Bot is an automated conversation system that replies to users' queries by analyzing them using NLP and assists them in every way it can [23]. Chat bots take care of frequent and time-consuming communications, saving people more time and fewer human resources. Chat bots can be found in various platforms, such as messaging apps, websites, and social media, and are commonly used for customer support, information queries, and personal assistance.

### 4.2 Sentiment Analysis:

Sentiment Analysis is a process of identification and extraction of subjective information from a lot of data available on the web to determine the positive, negative, or neutral sentiment feelings of the public towards a particular topic or entity. The Sentiment Analyst helps a great deal in e-commerce in terms of tracking the mood of the public in a product or service, and can be useful in many tasks such as marketing, customer service, and political analysis.

### 4.3 Language translation:

Language translation uses NLP to translate text or speech from one language into another. This application is very useful in education, communication and E-commerce. Many Language translation type applications are available such as Google Translate, Bing Microsoft Translator, DeepL, and a lot of other apps. Language translation allows people from different regions of the world to communicate with ease.

### 4.4 Speech recognition:

Speech recognition is the software that allows converting speech from a verbal format into usable structured data typically in the form of readable text. Speech recognition learns over time instead of feeding it experiences we feed its data into the form of audio and transcripts that enable artificial intelligence to distinguish between things such as age, gender, accents, and even different languages. The Applications of Speech Recognition enable the people to interact with their devices by voice such as Alexa, Cortana, Google Assistant and Siri.

### 4.5 Question answering:

Question answering imports the answer to the question from a specific text or can be created from scratch. Answering questions consists of extracting an answer from a given document; question answer templates take a context and a question and return an answer. Question answering models are used to automate FAQ replies using an information base, FAQ can automatically answer questions asked by customers it just need a document with information about business and interrogate this document with the questions asked.

## 5. The Top Techniques used in Natural language processing:

Natural language processing has many techniques used to extract data from text efficiently for improved productivity, and reduce the complexity. Below, we define the most techniques used in NLP:

### 5.1 Tokenization:

Tokenization is the most important step in many NLP applications, including machine translation, sentiment analysis, and Chabot development. This technology makes it possible for machines to process and understand human language more effectively. The Tokenization breaks the input text into unites called Tokens suitable for machine learning. This Tokens unites can be words or characters, sub words. There are several types of tokenization, including words Tokenization, character Tokenization, and sub-word Tokenization. Word tokenization breaks a sentence or phrase into individual words, while sub-words tokenization breaks a word into individual parts. Character Tokenization breaks a string of characters into individual characters.

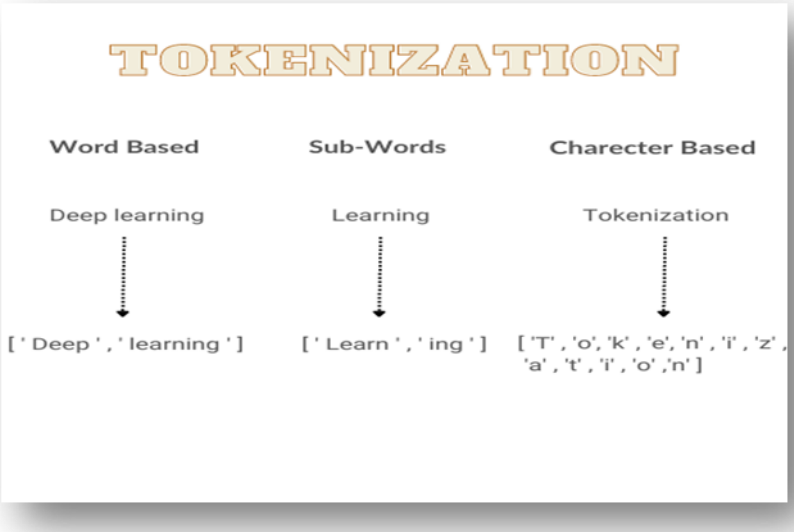



**Figure 2. 2: Types of Tokenization.**

### 5.2 Stemming and Lemmatization:

Lemmatization and Stemming are methods of the Text Normalization techniques. The goal is to generate the root word from the variations of a word. Stemming algorithm extract the root by removing the end of a word (suffix) this algorithm reduces the number of unique words in a text and simplifies analysis. For Lemmatization algorithm, it reduces the words at their base, it identifies the root or the 'lemma' based on the meaning of the word. These two techniques can be useful in a variety of contexts, depending on the specific objectives of the analysis. Both stemming and lemmatization removes inflection from words [24] .

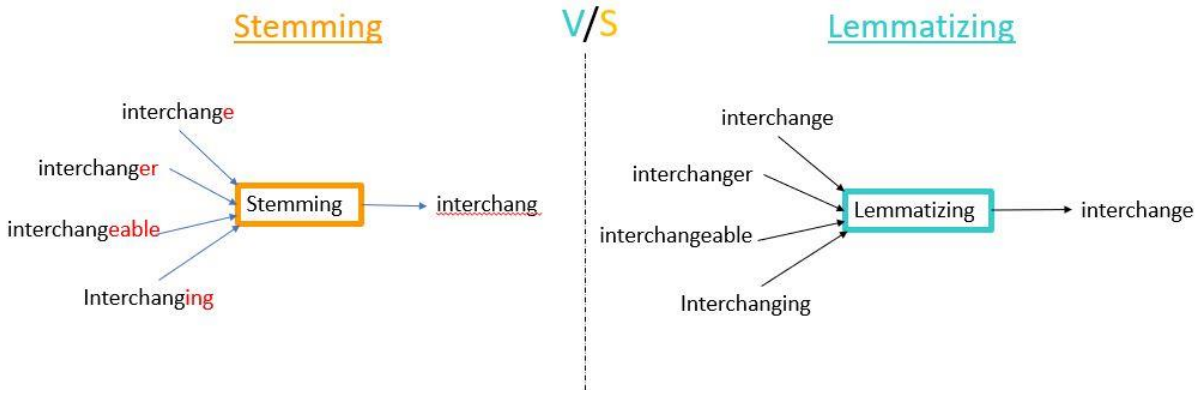



**Figure 2. 3: Stemming Vs Lemmatizing.[25]**

### 5.3 Stop Words Removal:

Stop Words are the words don't make much sense to a sentence (such as "the", "a", "an", "in"). Stop Words removes from text data. Remove the Stop Words allows computer to focus on words with meaning, and reduces the size of the dataset and the training time. And thus it improves machine learning or natural language processing applications.

### 5.4 Syntactic analysis:

Syntactic analysis or sometime called Parsing analysis, it is a phase of Natural Language Processing (NLP) is used to analyze syntax, is to understand the meaning of a sentence and to be able to generate new sentences that follow the same rules of grammar. It Uses the concept of parsing for analyze syntax, this concept works for checking for correct syntax using formal grammar by breaking down a sentence into its constituent parts, such as nouns, verbs, adjectives and prepositions, after that converts input data (text) into a structural representation which can be a parse tree, an abstract syntax tree, or another hierarchical structure for make computer can more easily understand the meaning and intent behind the input. Syntactic analysis is a complex and fascinating process has a crucial role in in enabling computers to understand and process human language and computer programming languages.

### 5.5 Semantic analysis:

Semantic analysis is the most difficult technique that because of difficulty making the computer understands sentences, paragraphs, or whole documents. The main goal of Semantic analysis is drawing meaning from text, and this is done by analyzing the grammatical structure of the text and determines relationships among individual words in a specific context.

### 5.6 Words2Vector:

Word2Vec [26] a word embedding methodology, converts the input words into a set of vectors that are distributed numerical representations, and Word2Vec gives the similar words similar numerical representations. Word2Vec has two strategies to include more contexts: Continuous bag of words, which increases the context by using the surrounding words to predict what occurs in the middle, and the second method, which is called skip the gram, increases the context by using the word in the middle to predict the surrounding word.

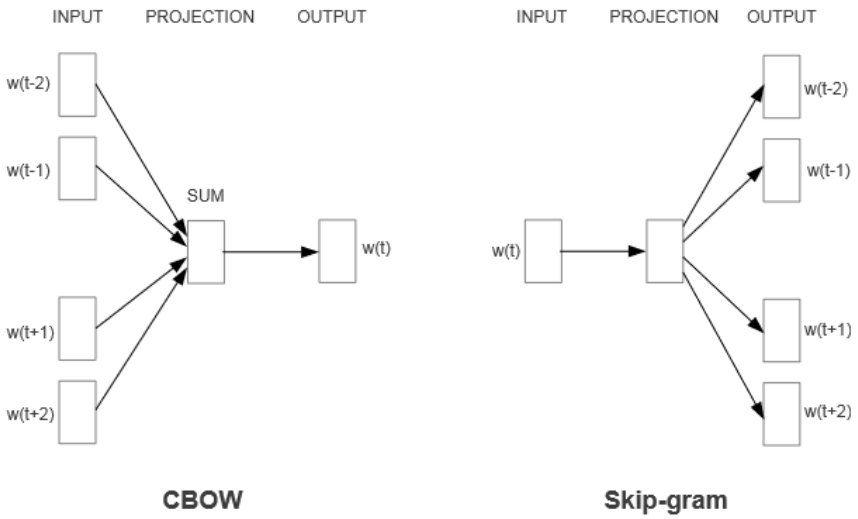



**Figure 2. 4: The CBOW and the Skip-gram architectures.[26]**

## 6. Recurrent Neural Networks (RNN):

Recurrent neural network is a widely used for time series data or sequential data such as text, speech, audio, video among others. RNN is a neural network that captures dynamic information in sequential data periodical connections of hidden layer nodes, and it can classify sequential data.

The important feature of RNN is "memory" that helps to store the output of hidden layer $\mathbf{h_t}$ and feeding this back to the next hidden layer $\mathbf{h_{t+1}}$ comes as input to generate the next output of the $\mathbf{h_{t+1}}$ layer.

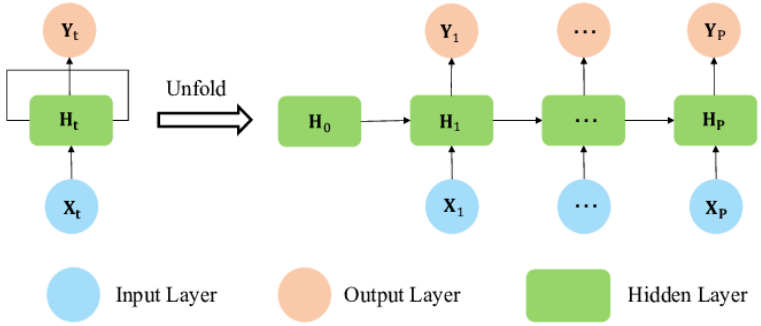



**Figure 2. 5: RNN architecture.**

Recurrent neural networks are most commonly used in the areas of natural language processing and speech recognition. RNN like the neural network has input layer and output layer but in many types –one to one, one to many, many to one, many to many. These types are used for different use cases such as Music generation, Sentiment classification, and Machine translation.

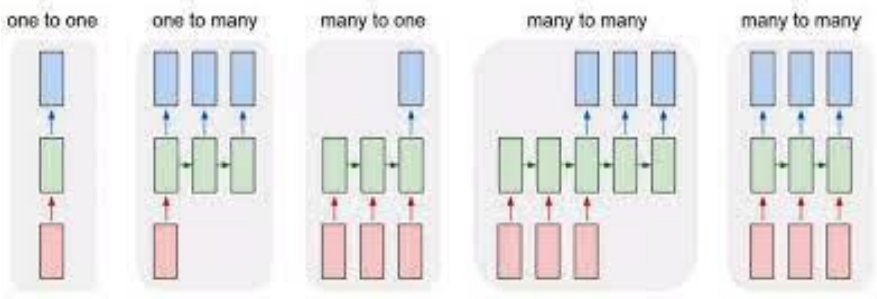



**Figure 2. 6: Types of RNN.**

Recurrent neural networks may not be able to save the information after a long time due to the limited memory capacity, there are other advanced RNN models such as LSTM (Long-Short-Term Memory) and GRU (Gated Recurrent Unit), and they are supposed to do better.

## 7. Long Short-Term Memory (LSTM):

Long Short-Term Memory (LSTM) is an artificial neural network used in the field of artificial intelligence and deep learning, and also is a type of Recurrent Neural Network, which also specializes in sequential data management, but it has a more complex structure, helping it to address problems that RNN faces such as the problem of vanishing gradient.

The LSTM structure consists of three gates (Forget Gate, input Gate, Output Gate).

- Forget Gate: selects important information to memorize and forgets information that is not needed.

- Input Gate: is responsible for providing the cell with the information to be entered.
- The output Gate: is the one that is working to make the information to be transmitted, from the current time to the next time.

This consistency between the gates at the structure level gives lstm the ability to learn long-term dependencies, and control the memory of information for a long time or forget about it [27].

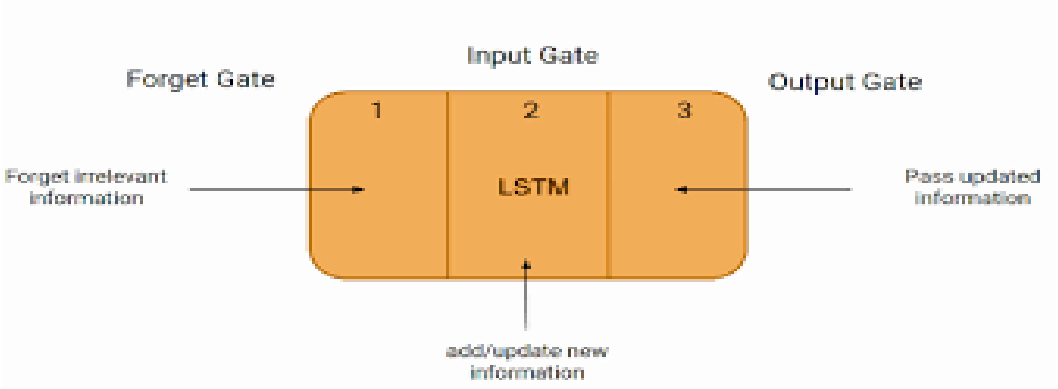



**Figure 2. 7: lstm architecture**

## 8. Gated Recurrent Unit (GRU):

Gated Recurrent Unit (GRU): Another type of (RNN) also aims to solve the problem of vanishing gradient, is very similar in performance to LSTM. It is quick to implement and does not require many memory resources due to its simple and uncomplicated structure [28].

It has only two Gates (Reset Gate and Update Gate):

- Reset Gate determines how much information to forget.
- The Update Gate determines the amount of information needed to extract from one cell to another.

GRU can also keep the information for a very long time, even that insignificant information [29].

## 9. Transformer Models and NLP:

A transformer model is a type of neural network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely [30]. This model has achieved a qualitative shift in deep learning, especially Natural language processing.

The Transformer architecture is Encoder/Decoder architecture based on the attention mechanism, the encoder takes the input and encodes it into a vector, and the decoder takes that vector and decrypts it into output sequence.

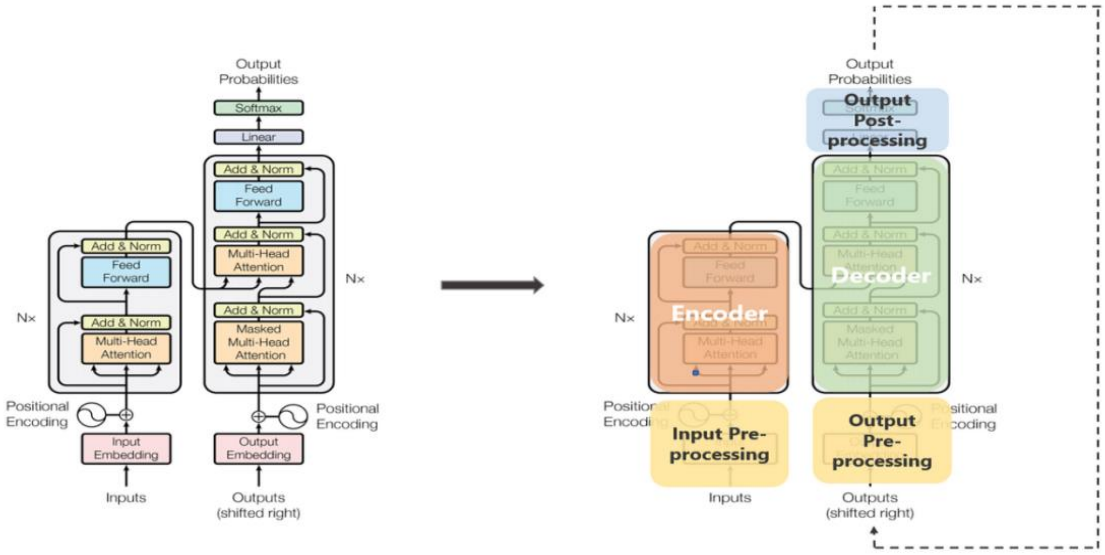



**Figure 2. 8: Transformer architecture. [31]**

There are three main innovations that make this model work so well –Positional Encoding, Attention, and specifically type of attention called self-attention. Attention is a mechanism that focuses on specific parts related to outputs. The true innovations in transformer are self-attention, a twist of traditional attention. Self-attention helps a neural network to understand an input word in the context and can also disambiguate words, recognize parts of speech and even identify word tense.

BERT shortcut for "Bidirectional Encoder Representations from Transformers» transformer-based model developed by Google, BERT was trained on massive text about 3.3 billion words of Wikipedia and thousands of books. It is used in a Google search to help understand search queries and it powers a lot of Google cloud's NLP tools

Another model has been developed based on the transformer is called Generative pre-trained transformer or GPT. It is a transformer-based language model uses deep learning algorithms and natural language processing techniques to produce text that resembles that of a human; GPT released by Open AI.

The open AI team used an approach called semi supervised that divides the training phase into two phases supervised learning and unsupervised learning. First unsupervised learning, the model learns through large unlabeled data, which allows the generative model to learn to condition on long-range information [32].

And then, they're fine-tuned through supervised training to get them to perform better. Open AI released GPT versions with different Parameter count –On 2018 GPT-1 with 117 million of parameter, on 2019 GPT-2 with 1.5 billion of parameter, on 2020 GPT-3 with 175 billion of parameter. The GPT can be used in many various natural language processing tasks

such as text generation, language translation, and text classification. In the last weeks of 2022, the company of Open AI released the powerful chatbot **"**ChatGPT**"** it's an updated version of GPT-3 which make can do a conversational text with human just when he send a message, It's very easy to use and available to the public, thus ChatGPT is the first Chabot that can do more complex conversational with human and The most important stage in the history of the NLP.

**Figure 2. 9: ChatGPT. [33]**

## 10. Conclusion:

NLP is interdisciplinary field of artificial intelligence. It enables machine to understand, interpret, and generate human language through the techniques and algorithms. NLP has the potential to transform the way we communicate and interact with technology, and in the last years we can expect to see even more exciting advancements in the future. So, NLP is a fascinating and rapidly growing field that holds tremendous promise for the future.

# Chapter 3:

# Visual Question Answering

# Chapter three: Visual Question Answering

## 1. Introduction:

Visual Question Answering is a challenging task that combines two main approaches of Artificial intelligence Computer vision and Natural language processing. The main goal of VQA is automatically answer the question based on the context of images or videos. In this chapter we talk about the VQA domain, which is one of the Vision-Language tasks, and the popular datasets and models, also the use cases and definition of the VQA problem.

## 2. Vision-Language Intelligence:

The advances in deep learning in both CV and NLP domains make many IA researchers care about integrating Vision and Language. Vision language is a task that integrates vision and language to simulate human intelligence and to understand the environment and natural language. Vision-language [34] tasks refer to the intersection of computer vision and natural language processing (NLP) and build multimodal which have the ability to process and generate text, images, and other forms of data. Multimodal can combine information from various sources, including images, sound, and text; these multiple modalities enhance real-world applications in various industries. For example, an autonomous vehicle should be able to process human orders (language), traffic signals (vision), and road conditions (vision and sounds) [34]. These Vision-Language problems include Image Captioning, Visual Question Answering (VQA), text-to-image generation, image-text matching, etc.

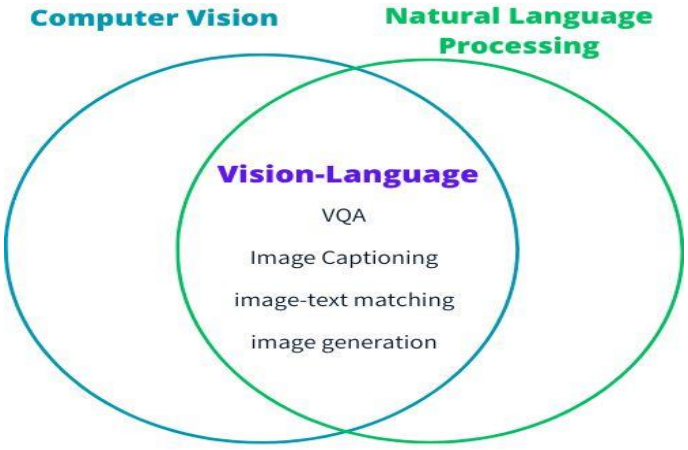



**Figure 3. 1: Vision-language Tasks.**

VQA plays a significant role in vision-language models, it is a benchmark task of evaluating the joint understanding and reasoning capabilities of such models. The VQA is one of the Vision-language tasks that require a high level of visual perception, language understanding, and reasoning, making this a difficult and complex problem.

## 3. Visual question answering Datasets:

VQA research is widely used in several Datasets to train and evaluate VQA models; each dataset has its characteristics, challenges, and uses. Here are some of the most popular VQA datasets:

### 3.1. VQA dataset:

VQA Dataset contains free-form and open-ended questions about images. The goal of free-form and open-ended questions is to increase the variety of knowledge to give accurate answers. VQA takes as input an image and a question about this image, this question requires an understanding of vision, language, and commonsense knowledge to produce an answer as the output. VQA is a large dataset containing 204,721of images from MS-COCO and a newly created abstract scene dataset (50,000 scenes) [35] for performing high-level reasoning; each image or scene has a collection of questions. The VQA dataset includes 1,105,904 questions and 11,059,040 answers [36].


**Figure 3. 2: Sample images and some questions on it. . [35]**

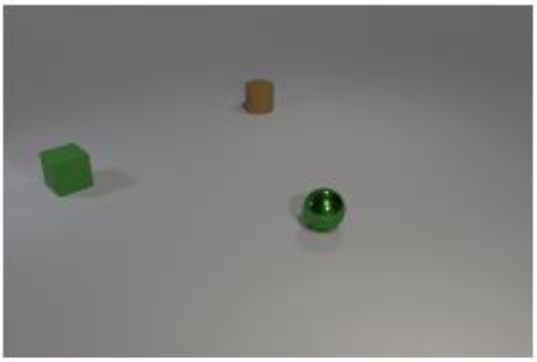
Questions are grouped into various types based on the words that start the question. Each kind of question has a typical answer such as: - the questions start with “Is the…?”, “Are…?”, and “Does…?” the answer must be Yes/No (About 40% of the answers yes or no), other questions such as “What is…?”, “What color…?”, and “Which…?” in many cases have an answer consisting of a single word. The reason for the short answers is that the questions extract specific answers from the images.

### 3.2. DAQUAR Dataset:

The Dataset for Question Answering on Real-world images (DAQUAR). First official dataset for the visual question answering task. It aims to understand and interpret natural language questions and answer them through real world images, especially NYU-Depth dataset images. It is a small size with only 1,449 photos and 12,468 pairs of questions and answers, with 2,483 individual questions [37], which makes it unsuccessful in training and evaluating the most complex models.

### 3.3. CLEVR Dataset:

CLEVR (Compositional Language and Elementary Visual Reasoning) is a synthetic Visual Question Answering dataset. CLEVR provides a dataset that requires complex reasoning to solve and that can be used to conduct rich diagnostics to better understand the visual reasoning capabilities of VQA systems [38]. It contains images of 3D-rendered objects. The CLEVR dataset consists of a training set of 70k images and 700k questions, a validation set of 15k images and 150k questions, and a test set of 15k images and 150k questions [39]. The dataset is generated using three objects in each image, namely cylinder, sphere and cube. These objects are in two different sizes, two different materials and placed in eight different colors. The questions are also synthetically generated based on the objects placed in the image. The dataset also accompanies the ground-truth bounding boxes for each object in the image [37].



**Q:** What is the color of the matte thing that is left of the thing behind the tiny green thing behind the tiny shiny sphere?
**A:** green

**Q:** There is a thing in front of the tiny block; is its color the same as the matte object in front of the tiny brown rubber thing?
**A:** yes

**Figure 3. 3: Sample image and some questions on it from CLEVR dataset [40].**

## 4. Deep Learning Based VQA Models:

Deep learning has played a significant role in the recent advances in VQA research. Here are some common deep learning-based VQA models:

### 4.1 Baseline model :

The VQA task combines two techniques image processing and natural language processing, the aim of VQA is to generate a process that can understand the both terms image and question to provide an answer. To generate this process using deep learning by develop two channels, a channel for vision (images) and a channel for language (questions). Image channel for extracting features from images using pre-trained CNN models. And the other channel, the textual question is processed using NLP techniques the LSTM. Then Fusion of image and question features, the features extracted from the image and question are combined to create a joint representation. This is typically done using techniques such as concatenation, element-wise multiplication, or neural attention mechanisms. In the last, the joint representation is passed through a classifier to generate an answer. classify the answer using softmax.

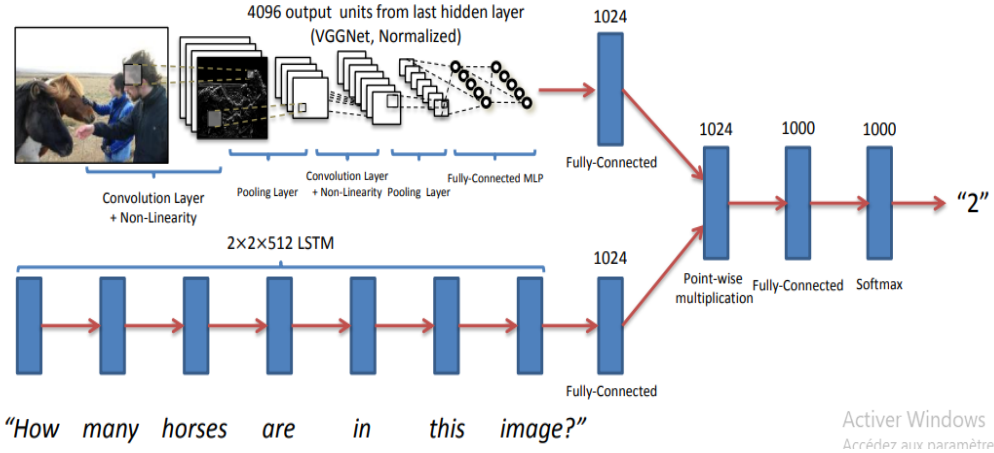



**Figure 3. 4: VQA Network Model.[41]**

Finally, these models often fail to give precise answers when such answers are related to a set of fine-grained regions in an image. [42]

### 4.2 Stacked Attention Networks (SANs) model :

SAN [42] is a deep learning-based model for VQA. It is multi-step reasoning for image QA that uses a series of attention modules to refine her predictions over time. Each attention is attended to different regions of the image and questions to answer. The three major components of SAN are: -image model uses CNN to extract the representation of images, and a question model to extract question semantic vector using LSTM. In the last, the stacked attention network can predict answers via multi-step reasoning. SAN series of stacked attention layers takes the image and question input to capture complex dependencies between the image and question, this process is repeated for multiple attention layers. Finally, the attended image/question features passed through a fully-connected layer, the answer with the

highest probability is then selected as the predicted answer. Yang et al [32] evaluates a SAN model in four QA image datasets and we identify these results in the table below:

| Dataset | Accuracy | Methods |
|---|---|---|
| COCO-QA | 61.6%, | SAN(2, CNN) |
| DAQUAR-ALL | 29.3%, | SAN(2, CNN) |
| DAQUAR-REDUCED | 46.2% | SAN(2, LSTM) |
| VQA | 58.7% | SAN(2, CNN) |

**Table 3. 1: Table of results for all four data sets and their accuracy.**

(The method is a SAN number 2 represents two attention layers used with CNN or LSTM model for questions [42])

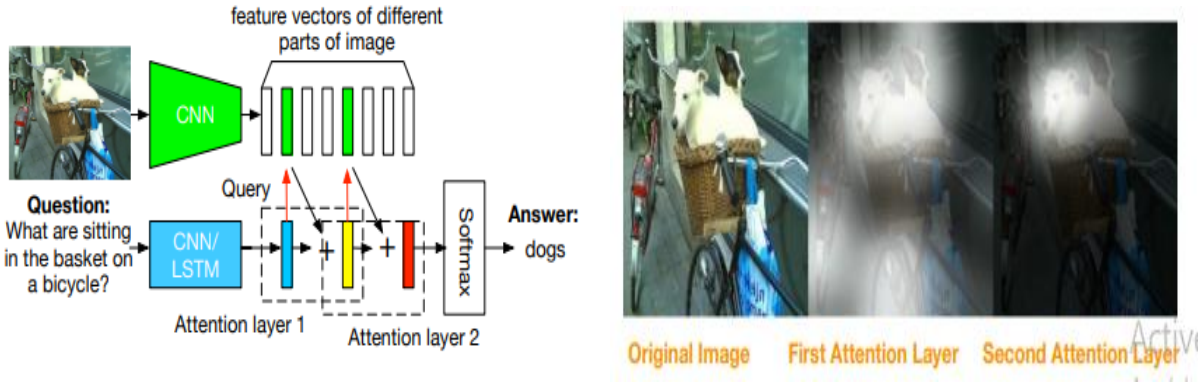



**Figure 3. 5: SAN Model architecture and visualization. [42]**

SAN is a powerful architecture for VQA, as it can reason over multiple steps and capture complex interactions between visual and textual information. However, the downside of SANs is that it can be computationally expensive due to their multi-stage nature.

### 4.3 Bottom-Up and Top-Down Attention model:

Teney et al. Model [43] presents a method for visual question answering. The proposed model is based on two methods, Global Vectors or Glove, the unsupervised machine learning technique for creating word embedding. And the other method is R-CNN is a method for object detection. The questions are split into words, and then the words are embedded by Glove after that the word representation passes through GRU. From the images channel, we get image features extracted by R-CNN based on bottom-up attention techniques. The Bottom-Up and Top-Down Attention mechanism allows the model to attend to different image regions based on the content of the question. The word representation and image features pass are combined to get a vector of both question and image, it is then fed into the output classifier class such as fully connected or softmax layers, which classifies the input to the proposed answer from a set of candidate answers.

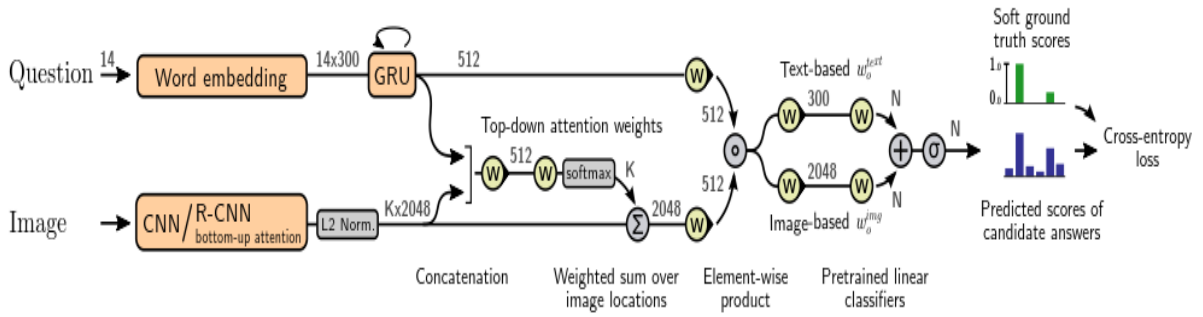



**Figure 3. 6: Teney et al. VQA Model. [43]**

Teney et al. VQA Model gets the first position at the 2017 VQA Challenge; it has achieved state-of-the-art performance on certain VQA datasets, such as the VQA 2.0 dataset. But the complex neural network architecture of the model makes it difficult to interpret or understand and this can make it difficult to diagnose mistakes and improve their performance.

## 5. transformer-based VQA methods:

The transformer is a new kind of deep learning that is composed of an encoder and a decoder. In the current state, the Transformer model is one of the most important deep learning architectures that play a significant role in the development of deep learning solutions for many problems, one of them being the VQA task. Here, are some popular transformer-based VQA methods:

### 5.1 LXMERT Method :

LXMERT [44] or "Learning Cross-Modality Encoder Representations from Transformers" is a transformer based-architecture multimodal that combines the image and text input. LXMERT includes three main encoders, to encode the image the Object-Relationship Encoder, and Language Encoder to encode question. Also, the cross-modality-Encoder is a fusion module that combines image and text representations. Each object relationship encoder and the language encoder consists of self-attention sub-layers; which uses to capture relationships between words in the text and between words and between visual features in the image, and a feed-forward sub-layer. The cross-modality encoder consists of cross-attention sub-layers and two self-attention layers and two feed-forward sub-layer.

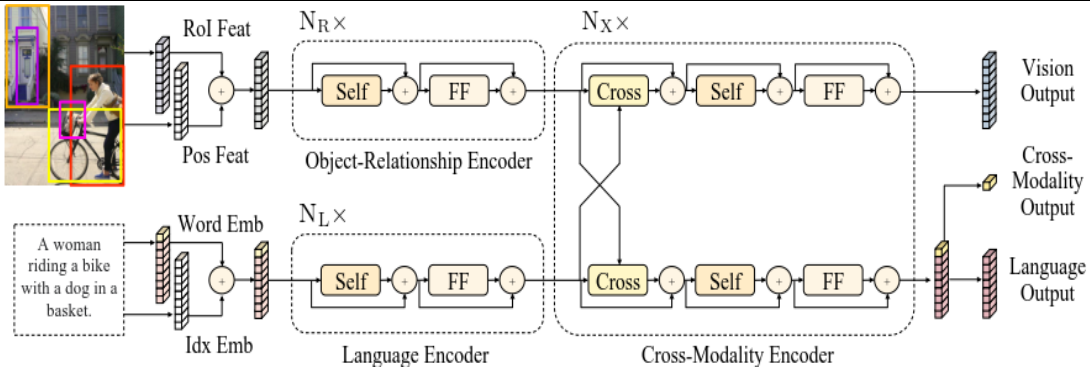



**Figure 3. 7: The LXMERT model [44] .**

The LXMERT model is pre-training on large-scale datasets to learn visual and language representations and performs well on VQA tasks. The LXMERT model has achieved great results for visual question answering, and it represents an important step forward in the development of multimodal can generate more accurate and relevant answers to questions about images and that's because it combines effective components in the model.

### 5.2 ViLBERT :

ViLBERT [45] Short for Vision-and-Language BERT, it is a model expanding the popular BERT architecture to a multimodal model that incorporates both visual and textual information. It is trained on large-scale datasets containing image and text pairs, such as the Conceptual Captions dataset. The model train and evaluate on the VQA 2.0 dataset (supervised learning objective), the model learns to associate visual and textual information to generate accurate answers. The trained ViLBERT model is evaluated on VQA datasets to assess its performance in terms of accuracy, robustness, and generalization of diverse data.

ViLBERT is a popular multimodal for visual question answering (VQA) tasks, which achieves state-of-the-art performance on several VQA benchmark datasets. It enables a joint understanding of image and question information and captures interactions and relationships between them to generate accurate answers in VQA tasks.

## 6. VQA use cases:

VQA is a technology that enables the machine to understand both vision and language, which can be used in a variety of industries and applications that require the processing of both vision and language depending on the condition of use. Here are some possible fields that could take advantage of the VQA task:

### 6.1 Medical Visual Question Answering (VQA):

Med-VQA is a combination of medical artificial intelligence and popular VQA challenges. Helps answer medical questions raised by patients automatically so as to relieve the shortage of experienced doctors. Given a medical image and a clinically relevant question in natural language, the medical VQA system is expected to predict a plausible and

convincing answer. The goal of Med-VQA is to lead to faster and more accurate diagnoses and treatment decisions. Current systems often require large-scale datasets and face challenges such as categorizing clinical questions, selecting relevant images, and capturing context and medical knowledge.

**Figure 3. 8: Medical artificial intelligence.**

### 6.2 VQA for visually impaired people:

Assistance to blind people is among the objectives of several VQA applications proposed in recent years. This is mainly due to the ability of automatic VQA to answer daily questions which may help visually impaired people to live without visual barriers. The VizWiz-VQA-Grounding dataset has been introduced, which is the first dataset that visually grounds answers to visual questions asked by people with visual impairments.

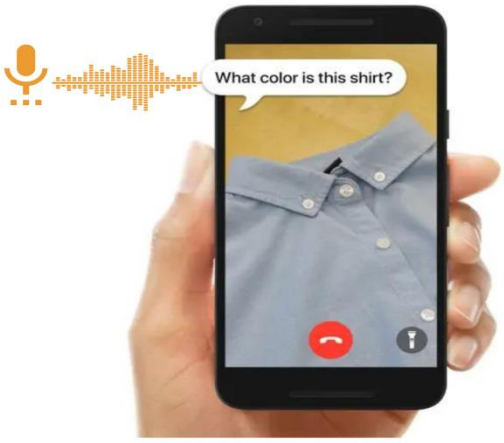



**Figure 3. 9: IA Assistance blind people.**

### 6.3 VQA in video surveillance scenarios:

Video surveillance is an important application of VQA. The adoption of a VQA approach in video surveillance scenarios may help operators to enhance the understanding of a scene, thus helping them to take fair and faster decisions. Video-QA is a question-and-answer format

that is conducted through video content, which enables the viewers to ask many questions about the video in real time.

**Figure 3. 10: video surveillance scenarios.**

### 6.4 VQA and advertising:

Advertising is something strongly related to image understanding. A user looking at an advertisement not only sees the objects in the scene, but also the related text and the relations among the objects, and interprets all such information within.

An advertisement must be quite simple to be understandable for the greatest number of people and at the same time interesting and eye-catching. No surprise then that VQA can find a challenging field of application in advertising. The first task to complete is using VQA to understand the advertiser and, in particular, the underlying communicative strategy.

### 6.5 VQA and Education:

VQA has many potential applications in the field of education. With the VQA education application, the student can interact and ask questions about visual content, such as images or diagrams, and receive immediate answers, thus increasing the learning process. VQA technology can help and improve accessibility for students with diverse learning needs, such as students with visual impairments or language barriers.

## 7. VQA Problem Definition:

VQA is the domain caring for building IA systems that can answer natural language questions about a specific image. For example, an image shows a little girl playing in a Park, and the question is “What is the little girl playing with?” The Expected answer is “jump rope”. The VQA model processes the question and analyzes the given image to generate an appropriate answer. By understanding the visual context, recognizing the girl, and interpreting the question correctly, the model can generate the expected answer “jump rope”.

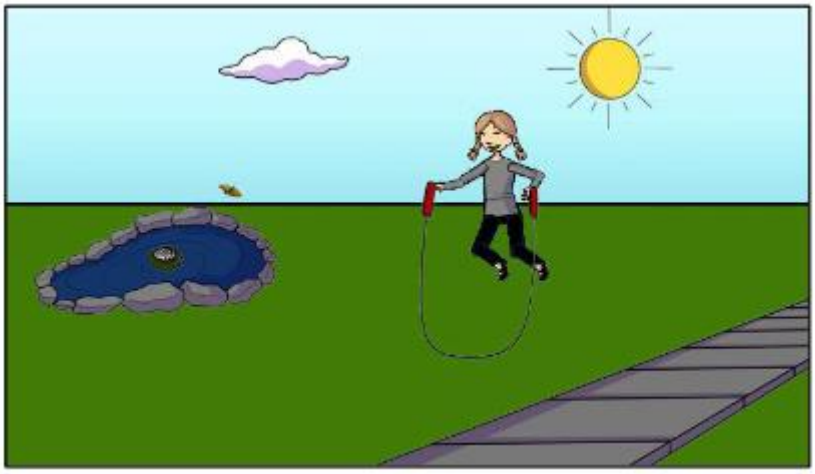


**Figure 3. 11: Example image of the abstract scene. [41]**

VQA model is facing significant challenges as understand both image and the question to determine the connection between them, and then attempt to answer [46] . VQA model suffers from a critical issue which is the languages biases that lead to biased results because (VQA) models strongly rely on learning statistical correlations between questions and answers without understanding the visual content [47] which can limit the accuracy of the models and reduces the robustness and generalization of VQA models also impairs the efficacy of their applications. For example , if we have an image of green banana and a question about it ,and the question is "What color is the banana? ",and the distribution of answers is based on answer yellow more than green. Based on this distribution, the model relies on language bias and response by "yellow" without looking to the image content.

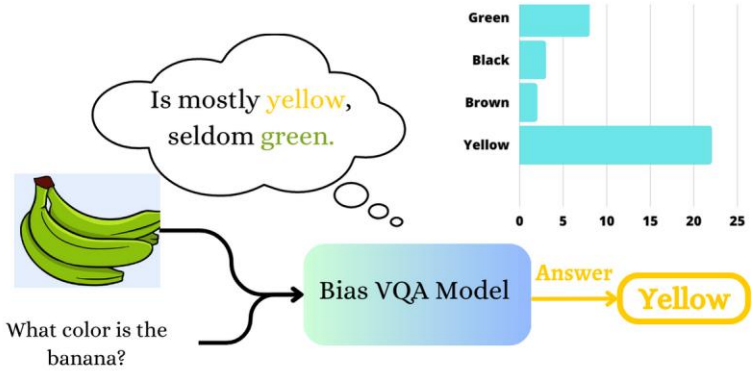



**Figure 3.12: Example of Language Bias.**

Overall, the main problem with VQA models is to generate a system that can understand the both image given and the natural language question about it and provides a natural language answer.

## 8. Conclusion:

VQA is a new technology, it enables the user to interact with both vision and language information. VQA technology is becoming more and more popular in the last few years because of its potential applications in different areas, including security, healthcare, education, and more. Researchers and developers are improving the algorithms and data sets of the VQA, making the future of VQA look promising. As VQA progresses, it can lead to more solutions in many different areas.

# Chapter 4:

# Realized work and obtained results

# Chapter 4: Realized work and obtained results

## 1. Introduction:

Ensemble learning is a technique that aims to strengthen and improve the overall performance of predictive models in the field of machine learning. In this chapter, we will reveal the techniques and steps used in addition to the tools used in this technique to reach the results achieved and analyze and compare the improvement of these results with the results available.

## 2. Ensemble Learning :

Ensemble learning is one of the common techniques of machine learning and deep learning; it combines many individual predictive models to obtain a final predictive model. In ensemble learning, multiple weak models known as "weak learners" are fitted to the same data and then integrate important predictions from these models to get a strong predictive model. This final model solves a lot of problems such as (bias, variance, classification, and regression...). Also, it adds stability and accuracy to predict and improve the overall performance of the ensemble.

### 2.1 Ensemble learning techniques:

Ensemble learning has many methods and techniques to solve a variety of problems, where we have two well-known techniques: Basic ensemble learning techniques and Advanced ensemble learning techniques.

2.1.1 Basic ensemble learning techniques:

- **Maximum Voting:**

Maximum Voting is one of the most famous techniques of ensemble learning, also called plurality voting. It is used to solve problems related to classification. The Maximum Voting technique makes a vote for all predictions resulting from each data point of multiple models and the prediction that gets the highest number of votes is the final prediction.

- **Averaging:**

The average is another simple technique characterized by strength and accuracy, which depends on combining multiple model predictions and taking the average of these predictions from each data point to produce a more accurate prediction.

- **Weighted average:**

The weighted average is a case of the average techniques used when we want to contribute some models more than others and multiply each model by a certain weight according to its ability and importance.

2.2.1 Advanced ensemble learning techniques:

- **Boosting:**

Boosting is a machine learning technique and one of the ensemble learning methods that processes data in a sequenced way, where the next model addresses and corrects the previous models. The boosting technique aims to reduce errors and transform weak learners into a strong learner by focusing on the wrong prediction.

- **Bagging:**

Bagging, also known as bootstrap aggregation is an ensemble learning method used to reduce variance. This technology goes through two phases, the first is bootstrap, which is the creation of a random subset containing projections taken from the data by substitution and then creates a base model for each subset, and then the second phase, the aggregation process, integrates predictions generated by simple techniques such as medium weighted voting.

- **Stacking:**

Ensemble learning technology trains many different models from the same data and uses its outputs as input to an algorithm called a meta-model which makes the final prediction.

## 3. Related works:

Recent studies on VQA generated many different models, including the use of deep learning like the CNN algorithm to extract features from images, and use of RNN or a transformer to process language, also the use of the pre-training multimodal to combine the image and text input. Moreover, there are also researchers who developed VQA models using ensemble model techniques. It is widely used in the field of VQA to improve the performance of VQA models by combining the outputs of several models and averaging these outputs. Here are some related works on the use of ensemble learning in VQA:

| **Author** | **Description** |
|---|---|
| Cappellato et Al. [48] | In this work, they extract features from images using the pre-trained ResNet-152, and BERT models to extract features of questions. Both features pass through the attention layer. For computing output features, they proposed bilinear transformation. Based on the same architecture, they build 11 models and trained on the training set and validation set of the Image CLEF-VQA-Med Dataset with combines the outputs and predict the result by averaging technique. |
| Clark et al. [49] | They propose an ensemble of two models; one naïve model does a good performance in the training set, but the opposite in the test set. The second model was trained in an ensemble with the first model in the training set, but in the test set the second model was trained alone and predicted the results. |

| | |
|---|---|
| Lioutasa et al. [50] | In this work, the authors propose an ensemble of two attention models that extract features of images by ResNet-101 for the first model and ResNet-152 for the second model. And each model embeds input text and then passes both representations of image and question through MLP and then the attention mechanism. This work increases amount 0.01 the accuracy of the single explicit attention model by a combined ensemble of based models. |

**Table 4. 1: related works.**

## 4. The Proposed Approach:

The researchers found that VQA models tended to answer the question blindly without looking at the content of the image in question This is caused by the difficulty of understanding both the picture and the question together, which reduces the robustness of VQA models and impairs the efficacy of their applications. In our proposed approach, we aim to tackle this issue by focusing on enhancing diversity on the side of vision within an ensemble of VQA models. Our approach proposes a novel ensemble method that focuses on the optimization of architectural diversity by integrating the ResNet-101+LSTM, ResNet-152+LSTM, and ResNet-50+LSTM models. The proposed approach is further described in the following section:

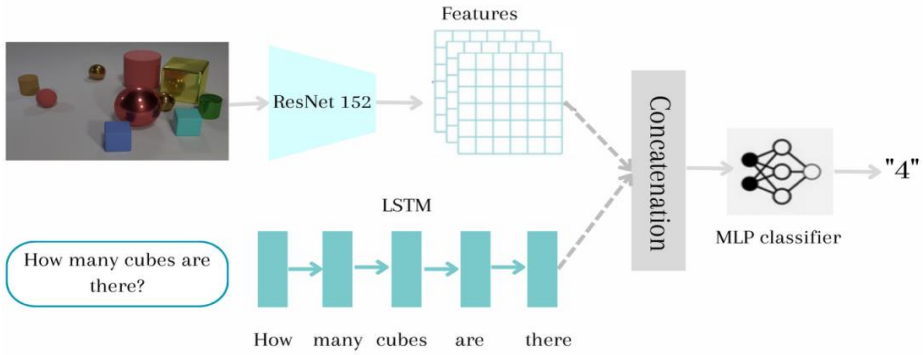



**Figure 4.1: ResNet-152+LSTM model.**

### 4.1 CLEVR Dataset:

The CLEVR dataset consists of synthetic images of 3D objects. The dataset consists of highly compositional questions that fall into different categories, including Exist, Count, Compare Integer, Query Attribute, and Compare Attribute [30]. The CLEVR Dataset includes on a training set 70000 image and 699,989 questions, on a validation set 15,000 images and 149,991 questions, and on a test set 15,000 images and 149,988 questions. Kaggle provides a CLEVR dataset containing folder images and JSON files for questions. The structure of the JSON file is as follows:

"questions" : 103 items
[ 100 items
0 : { 8 items
"image_index" : int 0
"program" : [...] 11 items
"question_index" : int 0
"image_filename" : string "CLEVR_train_000000.png"
"question_family_index" : int 2
"split" : string "train"
"answer" : string "yes"
"question" : string "Are there more big green things than large purple shiny cubes?"
}


**Figure 4.1: Question JSON file.**

### 4.2 Feature Extraction from images:

Feature extraction from images is of utmost importance, it helps to reduce the complexity by extracting meaningful features, and this can allow the model to learn images more abstractly and meaningfully. Generally, the features are extracted using pre-trained CNN models and can capture complex visual features. The proposed CNN models for feature extractions are Resnet101, Resnet152, and Resnet50. Each CNN architecture has its unique characteristics, such as depth and complexity, what happens is diversity in capturing visual information, this diversity improves the ability of the ensemble to generalize and make accurate predictions across a variety of inputs.

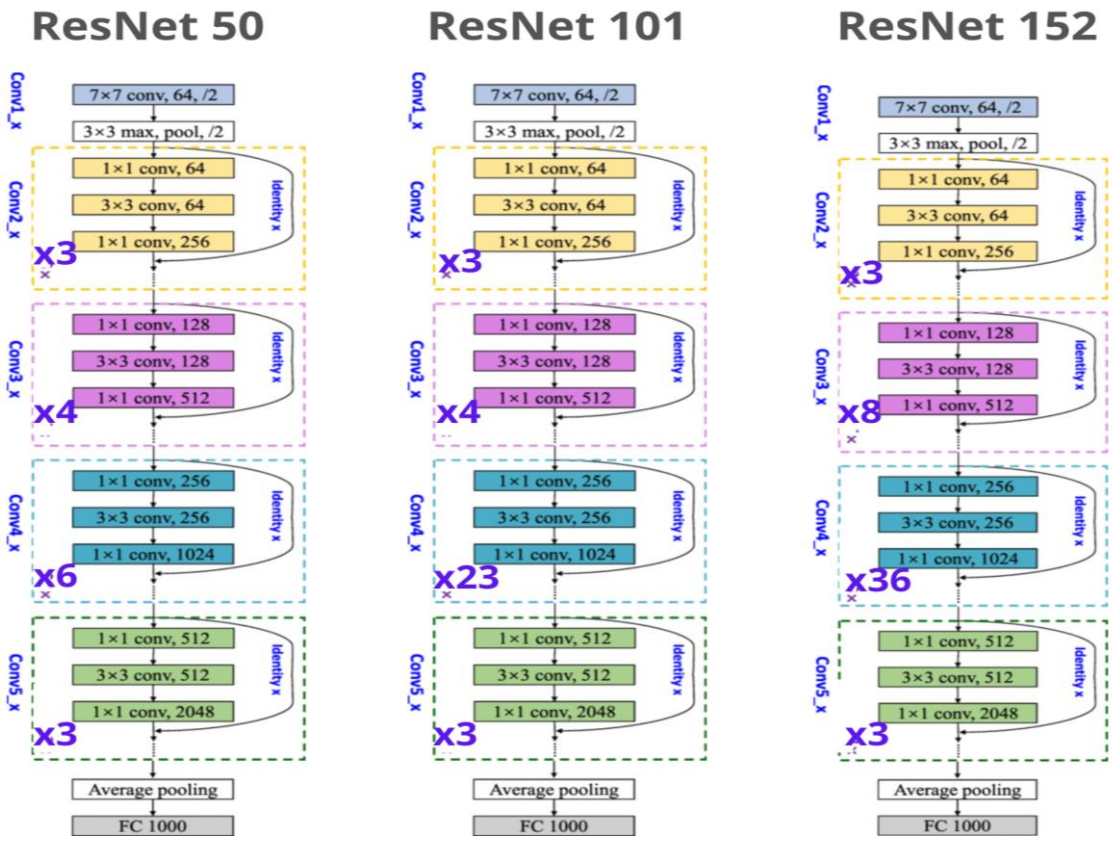



**Figure 4. 3:ResNet50, ResNet101, ResNet152 Architectures.**

- **Implementation Details:**

- Building the model:

Build pre-trained model to extract features from training and validation images. Initializes a ResNet-50 model with pre-trained weights from the torchvision library *"model = torchvision.models.resnet50(pretrained=True)"*. After that modifies the model architecture by *"nn.Sequential(*list(model.children())[:-3])"*, by removing the last 3 layers.*"* *model.cuda()"* this function is used to enable GPU acceleration. And finally return the model.

```
def build_model(args):
    model = torchvision.models.resnet152(pretrained=True)
    model= nn.Sequential(*list(model.children())[:-3])
    model.cuda()
    model.eval()
    return model
```

**Figure 4. 4: Building the model function.**

- Image processing:

Pre-process the images using OpenCV library operations to prepare it for input to a deep learning model. First, load the image in color mode using *"img = cv2.imread(path, 1)"* .And resize the input image to 224*224 size to fit in input size of model. Also change the channel order from height-width-channel (H×W×C) to (C×H×W).

```
# cv2.imread(path, flag) method loads an image from the specified file.
# "1" It specifies that should read image in RGB color way.
img = cv2.imread(path, 1)
# Resize image to 224*224
img = cv2.resize(img, img_size, interpolation=cv2.INTER_CUBIC)
# Returns a tensor that is a transposed version of input
img = img.transpose(2, 0, 1)[None]
```

**Figure 4. 5: Figure: Preprocessing input images.**

- Extract features:

Passes batch of 50 preprocessing images through the pre-trained model to obtain the extracted features. The model processes the input images batch and produces features, converts them to a NumPy array.

```
features = model(image_batch)
features = features.data.cpu().clone().numpy()
return features
```

**Figure 4. 6: Extract features function.**

### 4.3 Preprocess Questions:

The pre-processing converts textual questions into a suitable format that can be fed into a VQA model for training. This step pre-processes the questions and answers for the training, and validation sets. The text data is prepared by the Tokenization technique; it breaks the sentence into words. By Tokenization we get a sequence of tokens, each token mapped to a unique numerical identifier. This encoding allows the VQA model to work with numerical inputs rather than raw text.

- **Implementation Details:**

**Tokenization: *"s.split(delim)"* split a sentence to a sequence of words call tokens.**

```
#split a sentence into a list of words
  tokens = s.split(delim)
```

**Figure 4. 7: Tokenization.**

- Create a vocabulary of unique tokens and assign a numerical index to each token.

*{"question_token_to_idx": {"<NULL>": 0, "<START>": 1, "<END>": 2, "<UNK>": 3, "": 4, ";": 5, "Are": 6, "Do": 7, "Does": 8, "How": 9, "Is": 10, "The": 11, "There": 12, "What": 13, "a": 14, "an": 15, "and": 16, "another": 17, "any": 18, "anything": 19, "are": 20, "as": 21, "ball": 22, "balls": 23, "behind": 24, "big": 25, "block": 26, "blocks": 27, "blue": 28, "both": 29, "brown": 30, "color": 31…….}*

- After getting the sequence of tokens and the vocabulary, each token encoded to a unique numerical identifier.

pre-questions.py +

```
def encode(seq_tokens, token_to_idx):
  seq_idx = []
  for token in seq_tokens:
    if token not in token_to_idx:
    seq_idx.append(token_to_idx[token])
  return seq_idx
```

**Figure 4. 8: Encode function.**

### 4.4 Feature Fusion:

Feature fusion in the context of VQA involves combining visual features extracted from an image and textual features of a question to create a joint representation that incorporates

both formats. The purpose of unified representation is to capture the interactions between the image and the question this allows the model predicts the accurate answer. The technique used to combine the input data is Concatenation; it is a technique to combine the question and image features into a single vector representation.

- **Implementation Details:**

- Embed image representations into fixed-length vectors using CNN. The language representations encode to a single embedding vector using LSTM.
- Fusion: Combines the both image question feature vectors by Concatenation technique. The Concatenation feats pass through Multi-layers perception MLP, to classify the answer.

```
#Concatenation
cat_feats = torch.cat([q_feats, img_feats.view(N, -1)], 1)
scores = self.classifier(cat_feats)
```

**Figure 4. 9: Concatenation.**

- **Training:**

All three baseline models are trained using the Adam Optimizer with a learning rate of 5 × $10^{-4}$ with a batch size of 16. Each epoch of 10000 iterations.

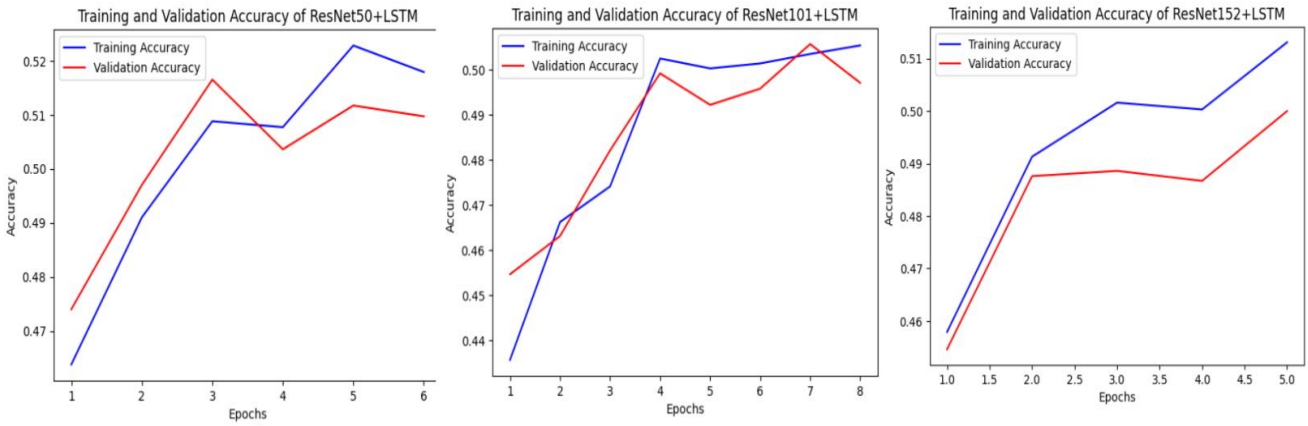



**Figure 4. 10: Training and Validations accuracy.**

- **The evaluation of the model :**

- **Accuracy:**

Accuracy is an evaluation metric used to assess the performance of a classification model. It is calculated using the following formula:

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

- **Adam Optimizer:**

Adam Optimizer is an optimization algorithm commonly used for training neural networks. It is an extension of the stochastic gradient descent (SGD) algorithm to update network weights iterative.

- **Learning Rate:**

The learning rate is a hyperparameter that determines the step size e.g. $5 \times 10^{-4}$ at which the optimization algorithm adjusts the model's parameters during training.

- **Batch Size:**

Batch size is hyperparameter determine the number of samples in each iteration.

- **Epochs:**

Epoch is refers the number of passes a training dataset during the training of a deep learning model.

- **The obtained Results:**

| Models | Training Accuracy | Validation Accuracy |
|---|---|---|
| ResNet 50+LSTM | 51,79% | 50,58% |
| ResNet 152+LSTM | 51,31% | 50,30% |
| ResNet 101+LSTM | 50,54% | 50,64% |

**Table 4. 2: Trainig and Validation Accuracy.**

### 4.5 Ensemble Models:

To combine the predictions of multiple models in an ensemble we use both Average and Weighted Average techniques. Through extensive experimentation and analysis, the weighted averaging technique was found to be more effective than averaging. Weighted Averages determine the most effective ensemble strategy, by assigning different weights to the predictions of each model. Efficiently, weights are determined based on the accuracy of the validation.

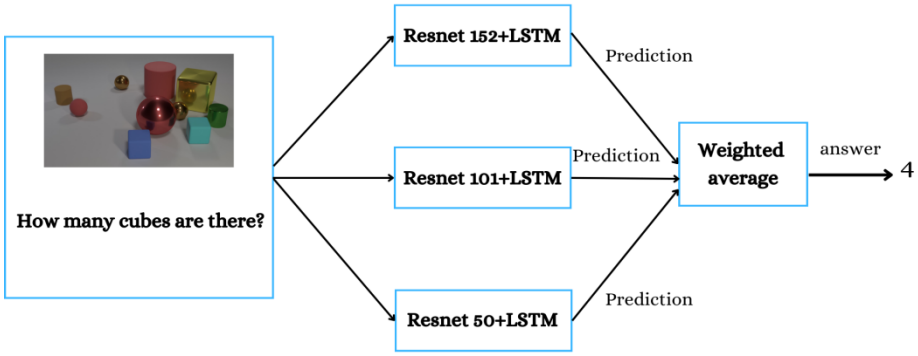



**Figure 4. 11: Ensemble models.**

## 5. Analyses results and discussion:

In comparing our ensemble model's accuracy with the highest accuracy single model, our ensemble outperforms the best model by 0.74%. This result suggests effectivity of diversity on the side of vision using various CNN architectures within an ensemble of models. This approach significant engagement and supports the development of the VQA domain.

There are several interesting future work directions, such as use advance models, e.g. stacked attention models was shown great success in VQA tasks , also more purposeful techniques could also be used to combine models e.g. Boosting. There is also potential applicability of the approach to other VQA datasets, different CNN architectures, or even other computer vision tasks.

Generally, the result indicates the effectiveness of leveraging diversity through the use of various CNN architectures within the ensemble. This strategy can be disseminated to different types of datasets in the VQA domain.

| ▪ **Models** | **Accuracy %** |
|---|---|
| ResNet 50+LSTM | 50,58 |
| ResNet 101+LSTM | 50,64 |
| ResNet 152+LSTM | 50,30 |
| **Our Ensemble models** | **51,32** |

**Table 4. 3: models accuracy.**

## 6. The Used Programing Languages, Libraries and Tools:

Here the Programing Language and Libraries, Tools are used in this work:

### 6.1Programing Languages Python:

Python is a fast and powerful high-level language program. Python is recognized for its simplicity, readability, and learning facility, it’s also open source and is used in a variety of fields such as artificial intelligence, web development, and data science. Python has many

large third-party libraries such as NumPy, Pandas, Keras, and PyTorch. The Python community is large includes enthusiasts, developers, and users who help in the development and promotion of the Python programming language, and can assist in supporting the beginner, the expert, to learn more about this language

### 6.2Libraries and framework:

- **PyTorch:**

PyTorch is an open-source machine learning framework. It is used in search and an application of deep learning because provides a variety of tools and modules for building and training neural networks. Pytorch has a dynamic computational graph system which makes it more flexible and easier to debug. Pytorch is also known for ease of use and has a great community. It can integrate well with other Python libraries and frameworks. Overall, Pytorch is the best choice for data scientists and researchers.


**Figure 4. 12: Pytorch Framework.[51]**

- **Numpy:**

Numpy is a Python library for numerical computing; it provides mathematical functions that allow to operate of multi-dimensional arrays and matrices efficiently. Numpy is a fundamental library for scientific computing in Python.

- **OpenCV:**

OpenCV (Open Source Computer Vision) is an open-source library for computer vision and image processing tasks. It provides functions for image and video analysis.

### 6.3 Tools:

We relied entirely on cloud technology Kaggle, which was very useful and provided a lot of benefits.

- **Kaggle:**

Kaggle is a platform for the data science and machine learning community; it provides a range of datasets, tutorials, and tools for data science and machine learning. Kaggle hosts competitions for the community which makes it a rich medium of knowledge and expertise that can be exchanged among developers. Also, Kaggle integrated with Google Cloud for

more productive tools and services for machine learning and data science. Kaggle provides about 13GB RAM and free GPU access.


**Figure 4. 2: Kaggle Platform.[52]**

## 7. Conclusion:

At the end of this chapter, we see that the results obtained have made remarkable progress and improved compared with the results available. So the ensemble learning technique will be instrumental in the future because of its great impact on machine learning and its ability to solve many problems, especially in fields that run a Visual Question Answering system that requires high-performance techniques and predictions to get accurate answers such as the medical field that requires working on the image and text together as well as the field of education and several other fields.

# General Conclusion

## General Conclusion:

VQA is a challenging task that combines both CV and NLP; it is an important research that has received growing interest from researchers. In this work, we have improved the accuracy of VQA models through ensemble learning and various CNN architectures.

This work can have several practical implications and benefits for real-world VQA applications, such as medical diagnosis. This diversity strategy can assist healthcare professional applications in making accurate diagnoses by accurately analyzing medical images and providing relevant answers to specific questions. Generally, this work can capture more fine-grained vision information, thus it can be effective in fields containing complex images that need accurate understanding and analysis to respond accurately.

VQA is a difficult and complex problem that requires a high level of visual perception, language understanding. It is difficult to bridge the gap between these modalities and integrate them effectively to produce accurate answers. The simple baseline model does not have the capability of capturing the relationships between images and questions, and this may limit the accuracy.

In conclusion, the VQA is a new task that has an impact on many real-world applications. It has an active community that actively contributes to the development of VQA models, datasets, and algorithms, which make the VQA have a promising future. Additionally, our research found the ability to improve the accuracy, by embracing ensemble learning and diverse CNN architectures. These efforts should contribute effectively to the VQA in the future.

# Bibliographie

# Bibliographie

[1] FUKUI, Akira, PARK, Dong Huk, YANG, Daylen, *et al.* Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[2] N-iX, www.n-ix.com consulted on: 13/02/2023.

[3] ALOIMONOS, Yiannis. Preface of special issue on purposive, qualitative, active vision. *Computer Vision and Image Understanding*, 1992, vol. 56, no 1, p. 1-3.

[4] Medium, medium.com consulted on: 13/02/2023.

[5] superannotate, www.superannotate.com consulted on : 10/02/2023

[6] v7labs,www.v7labs.com consulted on : 15/02/2023

[7] FENG, Xin, JIANG, Youni, YANG, Xuejiao, et al. Computer vision algorithms and hardware implementations: A survey. Integration, 2019, vol. 69, p. 309-320.

[8] F. Nielsen and R. Nock, "On region merging: the statistical soundness of fast sorting, with applications," *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Madison, WI, USA, 2003, pp. II-19, doi: 10.1109/CVPR.2003.1211447.

[9] CHEN, Changhao, et al. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. *arXiv preprint arXiv:2006.12567*, 2020.

[10] analyticsvidhya www.analyticsvidhya.com consulted on : 16/02/2023.

[11] packtpub packtpub.com consulted on : 13/06/2023

[12] BISHOP, Christopher M. et NASRABADI, Nasser M. *Pattern recognition and machine learning*. New York : springer, 2006.

[13] LIU, Wen, WANG, Shuo, DENG, Zhongliang, et al. A Review of Image Feature Descriptors in Visual Positioning. IPIN-WiP, 2021.

[14] medium,.medium.com consulted on : 20/02/2023.

[15] DULHARE, Uma N., AHMAD, Khaleel, et AHMAD, Khairol Amali Bin (ed.). Machine learning and big data: concepts, algorithms, tools and applications. John Wiley & sons, 2020.

[16] SIMONYAN, Karen et ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] SHI, Bibo, et al. Learning better deep features for the prediction of occult invasive disease in ductal carcinoma in situ through transfer learning. In: *Medical imaging 2018: computer-aided diagnosis*. SPIE, 2018. p. 620-625.

[18] HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, *et al*. Deep residual learning for image recognition. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.

[19] WAHLANG, Imayanmosha, SHARMA, Pallabi, SANYAL, Sugata, *et al*. Deep learning techniques for classification of brain MRI. *International Journal of Intelligent Systems Technologies and Applications*, 2020, vol. 19, no 6, p. 571-588.

[20] Investopedia,www.investopedia.com consulted on : 16/03/2023.

[21] Simplilearn,www.simplilearn.com consulted on : 16/03/2023.

[22] DONG, Chenhe, LI, Yinghui, GONG, Haifan, *et al.* A survey of natural language generation. *ACM Computing Surveys*, 2022, vol. 55, no 8, p. 1-38.

[23] REGIN, R., RAJEST, S. Suman, SHYNU, T., *et al.* An Automated Conversation System Using Natural Language Processing (NLP) Chatbot in Python. *Central Asian Journal of Medical and Natural Science*, 2022, vol. 3, no 4, p. 314-336.

[24] SAAD ,Tedj, Fake News Detection,Master, University of M'sila,2022.

[25] pluralsight,www.pluralsight.com consulted on : 16/05/2023.

[26] MIKOLOV, Tomas, CHEN, Kai, CORRADO, Greg, *et al*. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[27] Analytics vidhya,www.analyticsvidhya.com consulted on : 16/03/2023.

[28] ur Rehman, S., Khaliq, M., Imtiaz, S. I., Rasool, A., Shafiq, M., Javed, A. R., ... & Bashir, A. K. (2021). DIDDOS: An approach for detection and identification of Distributed Denial of Service (DDoS) cyberattacks using Gated Recurrent Units (GRU). *Future Generation Computer Systems*, *118*, 453-466.

[29] Scalar, www.scaler.com consulted on : 17/03/2023.

[30] VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, *et al.* Attention is all you need. *Advances in neural information processing systems*, 2017, vol. 30.

[31] medium,towardsdatascience.com consulted on : 20/03/2023.

[32] RADFORD, Alec, NARASIMHAN, Karthik, SALIMANS, Tim, *et al.* Improving language understanding by generative pre-training. 2018.

[33] Openai,openai.com consulted on : 27/03/2023.

[34] LI, Feng, ZHANG, Hao, ZHANG, Yi-Fan, *et al.* Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*, 2022.

[35] ANTOL, Stanislaw, AGRAWAL, Aishwarya, LU, Jiasen, *et al.* Vqa: Visual question answering. In : *Proceedings of the IEEE international conference on computer vision*. 2015. p. 2425-2433.

[36] Visualqa,visualqa.org on : 20/03/2023.

[37] Srivastava, Yash, et al. “Visual question answering using deep learning: A survey and performance analysis”. Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5. Springer Singapore, 2021.

[38] Johnson, Justin, et al. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[39] Marois, Vincent, et al. "On transfer learning using a MAC model variant." arXiv preprint arXiv:1811.06529 (2018).

[40] JOHNSON, Justin, HARIHARAN, Bharath, VAN DER MAATEN, Laurens, *et al.* Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 2901-2910.

[41] ANTOL, Stanislaw, AGRAWAL, Aishwarya, LU, Jiasen, *et al.* Vqa: Visual question answering. In : *Proceedings of the IEEE international conference on computer vision*. 2015. p. 2425-2433.

[42] YANG, Zichao, HE, Xiaodong, GAO, Jianfeng, *et al.* Stacked attention networks for image question answering. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 21-29.

[43] Teney D, Anderson P, He X, Van Den Hengel A. Tips and tricks for visual question answering: Learnings from the 2017 challenge. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 4223-4232).

[44] Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv preprint arXiv:1908.07490* (2019).

[45] Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *Advances in neural information processing systems* 32 (2019).

[46] Kansara, Pankti. Visual Question Answering" ,Master, San Jose State University ,(2018).

[47] Hirota Y, Nakashima Y, Garcia N. Gender and racial bias in visual question answering datasets. In2022 ACM Conference on Fairness, Accountability, and Transparency 2022 Jun 21 (pp. 1280-1292).

[48] VU, Minh, SZNITMAN, Raphael, NYHOLM, Tufve, *et al.* Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical

domain. In : *CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, Sept 9-12, 2019*. 2019.

[49] CLARK, Christopher, YATSKAR, Mark, et ZETTLEMOYER, Luke. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[50] LIOUTAS, Vasileios, PASSALIS, Nikolaos, et TEFAS, Anastasios. Explicit ensemble attention learning for improving visual question answering. *Pattern Recognition Letters*, 2018, vol. 111, p. 51-57.

[51] Pytorch, pytorch.org consulted on: 27/06/2023.

[52] Kaggle, www.kaggle.com consulted on: 27/06/2023.

## Abstract

Visual Question Answering (VQA) is a field that combines two different techniques: computer vision and natural language processing. Computer vision is used to process the image or video, and NLP uses for the processing of natural language. VQA is a technology that automatically answers the question based on the context of images or videos. The VQA is one of the Vision-language tasks that require a high level of language and image understanding, making this a difficult and complex problem. In this dissertation, we explore and apply an ensemble of diverse VQA models combined with Weighted Average techniques to increase the accuracy.



**Keywords:** Deep learning, CNN, LSTM, VQA, Ensemble learning, ResNet ,Computer vision, Natural language processing.



## Résumé

La réponse visuelle aux questions (VQA) est un domaine qui combine deux techniques différentes : la vision par ordinateur et le traitement du langage naturel. La vision par ordinateur est utilisée pour traiter l'image ou la vidéo, et NLP est utilisé pour le traitement du langage naturel. La VQA est une technologie qui répond automatiquement à la question en fonction du contexte des images ou des vidéos. La VQA est l'une des tâches en langage de vision qui exige un niveau élevé de compréhension du langage et de l'image, ce qui en fait un problème difficile et complexe. Dans le cadre de ce mémoire, nous explorons et appliquons un ensemble de divers modèles VQA combinés par des techniques de moyenne pondérée pour accroître précision.



**Mots clés:** Deep learning, CNN, LSTM, VQA, Ensemble learning, ResNet ,Computer vision, Natural language processing.



## الملخص

**الإجابة على الأسئلة المرئية هو مجال يجمع بين تقنيتين مختلفتين: رؤية الكمبيوتر ومعالجة اللغة الطبيعية. يتم استخدام رؤية الكمبيوتر لمعالجة الصورة أو الفيديو. وتستخدم معالجة اللغة الطبيعية في معالجة اللغة الطبيعية. الإجابة على الأسئلة المرئية هي تقنية تجيب تلقائياً على السؤال استناداً إلى سياق الصور أو مقاطع الفيديو. وهي من إحدى مهام لغة الرؤية التي تتطلب مستوى عالٍ من فهم اللغة والصورة، مما يجعل هذه مشكلة صعبة ومعقدة. في هذه المذكرة، نستكشف ونطبق مجموعة من نماذج الإجابة على الأسئلة المرئية المتنوعة جنبًا إلى جنب مع تقنيات المتوسط المرجح لزيادة الدقة.**



**الكلمات المفتاحية**

Deep learning, CNN, LSTM, VQA, Ensemble learning, ResNet ,Computer vision, Natural language processing.