Understanding Human Imagination Through Diffusion Model

Minh Nguyen Pham

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Computer Engineering

Creed F. Jones, Chair
Ryan K. Williams
Thinh T. Doan

December 7
Blacksburg, VA

Keywords: Artificial Intelligence, A.I, Machine Learning, Computer Vision, Artificial Neural Network, Human Vision

Understanding Human Imagination Through Diffusion Model

Minh Nguyen Pham

(ABSTRACT)

This paper develops a possible explanation for a facet of visual processing inspired by the biological brain's mechanisms for information gathering. The primary focus is on how humans observe elements in their environment and reconstruct visual information within the brain. Drawing on insights from diverse studies, personal research, and biological evidence, the study posits that the human brain captures high-level feature information from objects rather than replicating exact visual details, as is the case in digital systems. Subsequently, the brain can either reconstruct the original object using its specific features or generate an entirely new object by combining features from different objects, a process referred to as "Imagination".

Central to this process is the "Imagination Core," a dedicated unit housing a modified diffusion model. This model allows high-level features of an object to be employed for tasks like recreating the original object or forming entirely new objects from existing features. The experimental simulation, conducted with an Artificial Neural Network (ANN) incorporating a Convolutional Neural Network (CNN) for high-level feature extraction within the Information Processing Network and a Diffusion Network for generating new information in the Imagination Core, demonstrated the ability to create novel images based solely on high-level features extracted from previously learned images. This experimental outcome substantiates the theory that human learning and storage of visual information occur through high-level features, enabling us to recall events accurately, and these details are instrumental in our imaginative processes.

Understanding Human Imagination Through Diffusion Model

Minh Nguyen Pham

(GENERAL AUDIENCE ABSTRACT)

This study takes inspiration from how our brains process visual information to explore how we see and imagine things. Think of it like a digital camera, but instead of saving every tiny detail, our brains capture the main features of what we see. These features are then used to recreate images or even form entirely new ones through a process called "Imagination." It is like when you remember something from the past – your brain does not store every little detail but retains enough to help you recall events and create new ideas.

In our study, we created a special unit called the "Imagination Core," using a modified diffusion model, to simulate how this process works. We trained an Artificial Neural Network (ANN) with a Convolutional Neural Network (CNN) to extract the main features of objects and a Diffusion Network to generate new information in the Imagination Core. The exciting part? We were able to make the computer generate new images it had never seen before, only using details it learned from previous images. This supports the idea that, like our brains, focusing on important details helps us remember things and fuels our ability to imagine new things.

Contents

List of Figures

Chapter 1

Introduction

1.1 A brief history of AI

The domain of Artificial Intelligence (AI) has been a wellspring of inspiration across generations, with the notion of crafting a fully automated entity surfacing in the writings of scholars and inventors throughout history. However, it did not formally emerge as a research area until the 1950s when Alan Turing introduced the concept of applying mathematical principles to AI [1]. Turing, observing the human ability to acquire information and solve problems through reasoning, sought to instill a similar problem-solving approach in machines. Despite the novelty of this idea, computers during Turing's era were costly and functionally limited, accessible only to select high-profile groups with specific research resources. Shortly after Turing's groundbreaking concept became public, Allen Newell, Cliff Shaw, and Herbert Simon developed the Logic Theorist, the pioneering program designed to emulate human problem-solving processes [1][2]. While the program did not establish a standard for AI, it marked an important milestone and inspired numerous subsequent projects. The inaugural AI conference in 1955 garnered global interest, showcasing promising projects such as ELIZA, a language interpretation agent by Newell and Simon, Deep Learning introduced by John Hopfield and David Rumelhart, and Expert System, a program simulating human expert decision-making processes created by Edward Feigenbaum [3]. Despite the challenges posed by expensive and limited computers, the enhanced computing power, increased accessibility, and reasonable pricing of the new generation of computers propelled a

monumental leap in AI development between 1957 and 1974 [2]. However, after Marvin Minsky's unfulfilled claim that machines would match the intelligence of an average human within three to eight years, government interest waned, leading to a gradual decrease in AI research funding [1]. Nevertheless, AI persevered through this challenging period, as demonstrated in 1997 when IBM's Deep Blue triumphed over world chess champion Gary Kasparov, a landmark achievement frequently cited by AI researchers [1][2]. Subsequently, in tandem with the evolution of computer hardware, AI has flourished into a robust research field marked by innovation and continuous discoveries. The trajectory of AI's development demonstrates its resilience and adaptability in overcoming obstacles and advancing our understanding of intelligent systems.

Continuing into the present day, the landscape of Artificial Intelligence (AI) has undergone remarkable transformations, with advancements in technology, algorithmic improvements, and an exponential increase in computational power. The integration of machine learning, particularly deep learning techniques, has propelled AI into new frontiers, enabling it to surpass human-level performance in various tasks [1]. In recent years, breakthroughs in natural language processing, computer vision, and reinforcement learning have led to the development of sophisticated AI applications. Models like OpenAI's GPT-3 (Generative Pre-trained Transformer 3) [4] have demonstrated an unprecedented ability to understand and generate human-like language, revolutionizing language-based AI systems. Computer vision models, such as those used in facial recognition and image classification, have achieved remarkable accuracy, contributing to advancements in fields like healthcare, autonomous vehicles, and surveillance. The application of AI has permeated numerous industries, including finance, healthcare, education, and entertainment [1][3]. In finance, AI algorithms analyze vast datasets to inform investment decisions and optimize trading strategies. In healthcare, AI plays a crucial role in medical imaging analysis, drug

discovery, and personalized medicine. Educational platforms leverage AI for adaptive learning, tailoring educational content to individual needs. AI-driven content recommendation systems in entertainment platforms enhance user experience by predicting preferences and suggesting relevant content.

Given that AI is fundamentally inspired by human intelligence, there is a distinct trajectory of research known as brain-inspired AI, which centers on exploring the research of brain architecture and the fundamental aspects of intelligence [30]. This avenue aims to develop AI systems capable of showcasing more organic behaviors derived from consciousness rather than relying solely on predefined algorithms. The goal is to imbue artificial intelligence with a capacity for natural responses and actions, aligning more closely with the nuanced and adaptive characteristics observed in conscious human behavior. Brain-inspired AI, or neuromorphic computing, traces its roots to the 1940s with the inception of artificial neural networks (ANNs) [30][31]. Pioneered by Warren McCulloch and Walter Pitts, these early models laid the foundation for subsequent developments. The field experienced a setback in the 1960s when the limitations of single-layer perceptrons were exposed. However, the resurgence of interest in the 1980s, fueled by the backpropagation algorithm and advancements in computational capabilities, set the stage for the deep learning era of the 2000s. This period witnessed the ascent of deep neural networks, powered by abundant data, potent GPUs, and refined algorithms [30]. In the 2010s, the focus shifted towards neuromorphic computing, seeking inspiration from the brain's architecture both in hardware and software [31]. IBM's TrueNorth chip exemplified this approach, emphasizing energy efficiency and cognitive capabilities. Presently, ongoing research explores biologically realistic models and spiking neural networks, while projects like the Human Brain Project and Neuromorphic Computing Initiative strive to comprehend and replicate the intricacies of the human brain.

However, the widespread adoption of AI has not been without challenges. Ethical considerations, biases in AI algorithms, and concerns about job displacement have prompted increased scrutiny [4]. The responsible and ethical development of AI systems has become a focal point for researchers, policymakers, and industry leaders. Looking forward, the future of AI holds exciting prospects and significant challenges. Continued research aims to make AI more interpretable, explainable, and accountable. Addressing the ethical implications of AI, ensuring fairness, transparency, and privacy will be critical for its responsible deployment [4]. As AI technologies evolve, interdisciplinary collaboration between researchers, policymakers, and ethicists will be essential to navigate the complex and evolving landscape of artificial intelligence.

1.2 Motivation

Throughout history, human intelligence has been a captivating focus of research, owing to its extraordinary and complex structure that facilitates cognitive processes like learning, reasoning, problem-solving, perception, and language understanding. Consequently, the primary motivation driving AI research is the aspiration to comprehend and emulate the processes of human intelligence to some extent [5][6]. The encompassing objective for numerous researchers is to develop artificial systems capable of showcasing intelligence and adaptive behavior by understanding the cognitive functions inherent in the human brain [6].

Understanding human intelligence involves delving into the inner workings of the brain which is responsible for complex cognitive functions. Through the neural pathway, information being processed and transformed to meet the current task requirement, which when combined together, exhibit intelligence-like behavior that human possess [5]. Thanks to many insights that neuroscience provides, computational models and algorithms that mimic these biological such as artificial neural networks are getting closer to the goal of understanding human intelligence. Replicating human intelligence in artificial systems has several objectives by imbuing AI systems with cognitive abilities similar to those of humans [6]. This spans from basic abilities such as perceiving vision and gathering information through sensors to more sophisticated tasks such as natural language processing, enabling AI to comprehend and generate human-like language, and logical reasoning and problem-solving for more complex planning tasks [5][6].

While AI has made significant advancements, it still faces challenges, and many of these challenges are rooted in the fundamental principles upon which many AI algorithms are built, particularly the statistical nature of decision-making [5]. Specifically, many AI algorithms, especially those based on machine learning, operate on statistical principles [6]. They learn

patterns and make decisions based on probabilities derived from training data. This statistical approach can lead to challenges, especially in scenarios where data is biased or unrepresentative. Additionally, AI systems often lack common sense reasoning and the ability to understand context in a way that humans do. This limitation hinders their ability to navigate complex, real-world situations effectively [6]. Moreover, the reliance on statistical patterns and data-driven decision-making raises ethical questions about accountability, responsibility, and transparency [5][6]. Addressing these concerns is crucial for the ethical deployment of AI technologies. Mitigating these challenges involves ongoing research in areas like explainable AI, robust machine learning, and the development of ethical guidelines for AI deployment. Balancing the power of statistical decision-making with ethical considerations and a deeper understanding of causation remains a key focus for the AI community.

To address the constraints within AI and move towards achieving general intelligence, I propose that it is essential to deepen our comprehension of human intelligence and leverage its biological framework. Thus, the objective of my research is to apply insights from the biology of the human brain to enhance the complexity of AI functionality. This approach seeks to enable more flexible responses rooted in a genuine comprehension of situations, moving beyond reliance on simplistic statistical metrics. Moreover, the strategy strives to reduce bias stemming from training data by integrating principles derived from the inherent architecture of human intelligence. Given the scope of the research, I have segmented the topic into two parts. The initial phase, which is the focal point of this thesis, entails comprehending the workings of human vision, leading to an understanding of the imagination process within the brain. This knowledge is then utilized to recreate the imagination process within a structure known as the Imagination Core, employing Stable Diffusion as an image synthesis model. The subsequent phase involves leveraging the

Imagination Core to construct a visual state space containing the relative locations and properties of each element within the environment. This spatial awareness facilitates the Reinforcement Learning process, enabling the agent to learn to make informed decisions and predict the next state while navigating unfamiliar surroundings.

Although the aspiration to emulate human intelligence is ambitious, it is crucial to recognize that AI systems do not necessarily have to duplicate every facet of human cognition to be beneficial [5]. Researchers, instead, strive to grasp fundamental principles and mechanisms that play a role in intelligent behavior. The interdisciplinary character of AI research, integrating insights from neuroscience, psychology, computer science, and diverse fields, highlights the comprehensive endeavor to comprehend and reproduce human intelligence in artificial systems [6]. This pursuit not only propels technological progress but also enriches our comprehension of the very essence of intelligence.

1.3 Thesis Organization

Chapter 2 provides a foundational understanding derived from relevant literature to establish the necessary groundwork for the simulation.

Chapter 3 provides an in-depth exploration of the simulation process, encompassing the preparation of the dataset, the thought process guiding the workflow of the program, and a step-by-step explanation of the implementation of the simulation.

Chapter 4 entails a comprehensive examination of the simulation's final results, drawing comparisons with the initial expectations to formulate a conclusion.

Chapter 5 provides a summarization of the experiment and outlines the future work, paving the way for the second phase of the research topic.

Chapter 2

Review of Literature

2.1 Human Brain System

The human brain is a marvel of biological complexity, serving as the central organ of the nervous system and the seat of consciousness, intelligence, and control over bodily functions [7][8]. Comprising approximately 86 billion neurons, the brain is a highly intricate network of cells that communicate through electrical and chemical signals. Neurons, the primary building blocks of the brain, form connections called synapses, allowing them to transmit information to one another.

The brain is divided into different regions, each responsible for specific functions. The cerebral cortex, the outermost layer, is associated with higher cognitive functions such as perception, memory, language, and decision-making [6][8]. The limbic system, situated deeper within the brain, plays a crucial role in emotions, motivation, and memory formation.

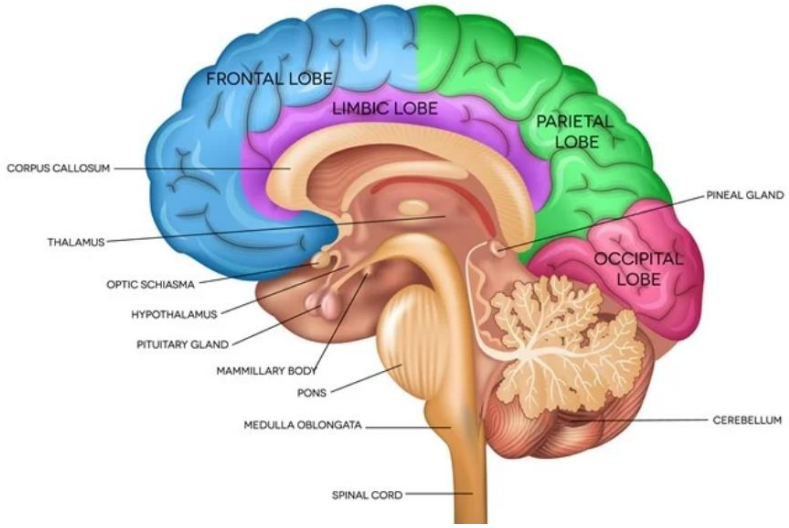

Figure 2.1 The brain anatomy. *Image Credit:Tefi/Shutterstock.com*

The brainstem, connecting the brain to the spinal cord, regulates basic physiological functions like breathing, heartbeat, and sleep. Neurotransmitters, chemical messengers released by neurons, play a vital role in communication between brain cells [7]. Dopamine, serotonin, and acetylcholine are examples of neurotransmitters that influence mood, motivation, and learning. The brain's plasticity, or ability to adapt and reorganize itself, allows for learning and memory formation [6]. Structural and functional changes occur in response to experiences, and this adaptability underlies the brain's capacity to learn new skills and recover from injury.

Another pivotal aspect influencing the functionality of the brain is the connectivity of synapses. Synapses serve as the communication junctions between neurons in the brain, facilitating the transmission of nerve impulses between neurons and muscle cells. This complex process translates the brain's intentions into actions, enabling different types of activity. The firing speed of synapses in the human brain can vary significantly based on the type of synapse and the specific neural circuit in question. Typically, synaptic transmission takes place within the range of milliseconds. The mechanism involves the release of neurotransmitters, their binding to receptors on the postsynaptic membrane, and the generation of postsynaptic potentials, collectively contributing to the overall speed of synaptic communication. While certain synapses operate with durations in the range of a few milliseconds, others may have longer timeframes. The precise timing of these synaptic events is crucial for ensuring the proper functioning of neural circuits and the effective processing of information within the brain.

The study of the brain has been advanced by various technologies, including functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), enabling researchers to observe brain activity in real-time [7][8]. Disorders affecting the brain, such as Alzheimer's disease, Parkinson's disease, and psychiatric disorders, highlight the complexity of the brain's functioning

and the challenges in understanding and treating neurological conditions [8]. Ongoing research in neuroscience continues to unravel the mysteries of the brain, contributing to our understanding of consciousness, cognition, and the intricate interplay between biology and behavior [6].

2.2 Human Vision

Human vision is the sensory ability that allows individuals to perceive and interpret the surrounding world through the sense of sight [9]. It is a complex process that involves the eyes, the optic nerves, and the brain working together to gather, process, and make sense of visual information [9][10]. The journey of visual perception begins with the eyes, which act as the primary receptors equipped with specialized cells known as cone and rod cells [9]. These cells play a pivotal role in capturing light, converting it into electrical signals that serve as the language of communication within the visual system. The cones are responsible for detecting color variations, allowing us to perceive a rich spectrum of hues, while the rods facilitate vision in low-light conditions, contributing to our ability to see in dim environments [9][10]. These signals are then transmitted through the optic nerve to the brain, where a complex network of neural processes occurs. The brain's visual cortex plays a crucial role in organizing and interpreting the incoming signals, forming the basis for our perception of colors, shapes, depths, and motion [6][10]. From the complexity of color vision, depth perception, and motion detection to the fascinating phenomenon of visual illusions, human vision is a subject of continuous exploration and scientific inquiry. Understanding the mechanisms behind how we see not only sheds light on the marvels of our biological design but also has deep implications for fields ranging from psychology and neuroscience to artificial intelligence and virtual reality [6][7].

Human vision serves as a profound source of inspiration for the field of computer vision, which endeavors to replicate and complement the intricate processes involved in visual perception [9]. The biological foundation of human vision, encompassing the eyes, optic nerves, and cognitive processing in the brain, guides the development of algorithms that emulate perceptual abilities. Researchers closely examine how humans recognize objects, interpret scenes, and extract meaning

from visual cues, paving the way for the creation of computational models that share parallels with the complexity of human vision [10]. The synergy between human and computer vision becomes evident in the domain of image processing and feature extraction. Both systems employ techniques to make sense of visual data, such as edge detection and color analysis. Feature extraction, a fundamental aspect of pattern recognition, is a shared trait as humans naturally focus on key features for object recognition, while computer vision algorithms extract relevant features to identify patterns and objects [9]. Object recognition and classification represent another convergence point between human and computer vision. While human vision excels at recognizing patterns and objects in the visual field, computer vision leverages advanced technologies like deep learning and convolutional neural networks (CNNs) to achieve human-level performance in tasks such as image classification and object detection. Although CNNs may not attain the complexity of biological neural networks, they effectively emulate the brain's ability to establish correlations among groups of pixels and derive meaningful patterns from these associations [9]. This is achieved through the use of convolutional layers that systematically scrutinize local regions of input, mirroring the human brain's hierarchical and interconnected approach to recognizing and interpreting visual information. The reciprocal relationship between human and computer vision is underscored by applications in medical imaging, autonomous vehicles, and augmented reality, where computer vision not only replicates but also enhances human vision capabilities in diverse fields [9][10].

2.3 Human Imagination

Human imagination is a complex cognitive process that involves the intricate blending and recombination of features derived from a multitude of objects and experiences encountered over time [11]. It is a remarkably creative aspect of human cognition, allowing individuals to transcend the boundaries of reality and generate novel mental representations.

The process of imagination begins with the collection of sensory information from the environment [11][12]. As we interact with the world, our senses gather details about the shapes, colors, textures, and other distinctive features of various objects. These features, stored in our memory, become the building blocks for the imaginative process. Imagination unfolds as a dynamic interplay of these stored features. The mind engages in a creative synthesis, combining elements from different objects and experiences to construct mental images or concepts that may not exist in the physical realm [11][13]. This process is not constrained by the literal representation of individual objects but rather involves a fluid and flexible rearrangement of their attributes. Furthermore, the ability to imagine is not limited to the mere replication of observed objects; it extends to the generation of entirely new ideas, scenarios, or entities [12]. Through the creative amalgamation of features, the mind can conceive unique and innovative possibilities, contributing to artistic expressions, problem-solving, and the development of imaginative narratives [6][11].

In summary, human imagination is a multifaceted cognitive phenomenon that draws upon the rich repository of features gathered from our experiences [11]. It involves a creative synthesis, allowing us to transcend the limitations of individual objects and envision novel mental constructs that fuel creativity, innovation, and the expansive realm of human imagination.

2.4 Computer Vision

Computer vision is an interdisciplinary domain that enable machines to comprehend and interpret visual information extracted from the outside world, mimicking human visual perception [14]. It encompasses a range of tasks, including image and video recognition, object detection, image segmentation, and scene understanding [14][15]. The goal is to enable machines to extract meaningful insights and make decisions based on visual data. Computer vision finds applications in diverse domains, from healthcare and automotive industries to security and entertainment [15]. One fundamental aspect of computer vision is image recognition, where algorithms analyze and classify images into predefined categories. Convolutional Neural Networks (CNNs) have been particularly successful in image recognition tasks, learning hierarchical features from data and achieving state-of-the-art results in image classification challenges like ImageNet [14]. Object detection extends image recognition by not only classifying objects in an image but also locating and delineating their positions. This is crucial for applications such as autonomous vehicles, where identifying and tracking objects in real-time is essential [15]. Image segmentation involves dividing an image into meaningful regions, allowing for a more detailed understanding of its content. This is especially useful in medical imaging for identifying and analyzing specific structures within organs. Computer vision techniques are also employed in facial recognition systems, enabling machines to identify and verify individuals based on facial features [14]. While this technology has various applications, concerns related to privacy and ethical considerations have prompted discussions and debates regarding its widespread adoption.

The development of computer vision is closely tied to advances in deep learning, where neural networks are trained on large datasets to automatically learn hierarchical representations of visual features [15]. Transfer learning, a technique where pre-trained models are fine-tuned for specific

tasks, has accelerated progress in computer vision by leveraging knowledge gained from one domain to improve performance in another [14]. Despite significant advancements, challenges persist, including the need for robustness to variations in lighting and viewpoints, as well as addressing bias in datasets to ensure fair and ethical applications of computer vision technologies.

2.5 Machine Learning Techniques

Machine learning encompasses computational algorithms or systems engineered to acquire patterns and formulate predictions or decisions devoid of explicit programming tailored for a specific task [16][17]. These models form a crucial element of the broader realm of machine learning, a subset of artificial intelligence (AI). Within machine learning, a diverse array of techniques empowers computers to glean insights from data and formulate predictions or decisions without the need for explicit instructions.

One fundamental approach within machine learning is supervised learning, where models are trained on labeled datasets to understand the mapping between input features and corresponding output labels [16][17][18]. This method is extensively applied in tasks such as classification and regression. Unsupervised learning, in contrast, involves training models on unlabeled data to uncover inherent patterns, structures, or relationships within the data [16][17]. Common unsupervised learning techniques include clustering and dimensionality reduction.

Another essential technique is reinforcement learning, where an agent learns to make decisions by interacting with an environment [16][17]. The agent receives feedback in the form of rewards or penalties based on its actions thus allowing it to learn optimal strategies over time. Reinforcement learning has shown remarkable success in applications such as game playing and autonomous systems.

Last but not least is the technique called transfer learning which involves fine-tuning a model pre-trained on one task for a related but different task [16][18]. This leverages knowledge acquired from a source task to enhance performance on a target task with limited data. Transfer learning has found applications in diverse domains, including natural language processing and computer vision. Ensemble learning, another noteworthy technique, combines multiple models to enhance

overall performance. Techniques such as bagging (Bootstrap Aggregating) and boosting create ensembles by training models independently or sequentially and combining their outputs [16]. Ensemble methods are recognized for their ability to bolster model robustness and generalization. Machine learning techniques often involve iterative optimization processes, adjusting model parameters based on feedback to minimize errors or improve performance. Gradient descent is a common optimization algorithm used to find the minimum of a cost or loss function [16][17]. Regularization techniques, such as L1 and L2 regularization, help prevent overfitting by penalizing complex models. Hyperparameter tuning is another crucial aspect, involving the search for optimal settings that govern the learning process [16]. Machine learning techniques continue to evolve, driven by research advancements, increasing availability of data, and improvements in computational resources, leading to their widespread application in diverse fields such as healthcare, finance, and autonomous systems.

2.6 Deep Learning

Deep learning constitutes a subset within machine learning, emphasizing the creation and training of artificial neural networks, with a specific emphasis on deep neural networks [19][20]. The term describes a process in which data undergoes transformation across multiple layers, facilitated by weighted connections and activation functions between each perceptron. This structure is distinguished by the inclusion of several hidden layers positioned between the input and output layers [20]. The fundamental building block of deep learning is the artificial neuron or perceptron, which processes information and learns from data through a process known as training allowing the model to learn and make decisions [19]. Deep learning has demonstrated exceptional accomplishments across diverse domains, encompassing computer vision, natural language processing, speech recognition, and beyond [20].

In deep learning, the hierarchical architecture of deep neural networks empowers them to autonomously learn and extract advanced features from raw data, leading to their exceptional performance in tasks like image and speech recognition, natural language processing, and reinforcement learning. [19][20]. The depth of the network allows it to capture hierarchical representations of features, with each layer learning increasingly abstract and complex patterns. This capability has led to significant breakthroughs and advancements in various domains, transforming the landscape of artificial intelligence [19].

Training deep neural networks involves an iterative process known as backpropagation, where the model adjusts its internal parameters based on the error between predicted and actual outputs. Advances in deep learning have been fueled by the availability of large datasets, powerful computing hardware and innovations in neural network architectures [19][20]. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are two common types of deep

neural networks that have proven particularly effective in computer vision and sequential data processing, respectively [19]. Deep learning continues to be at the forefront of research and application development, pushing the boundaries of what is possible in artificial intelligence and machine learning.

2.6.1 Artificial Neural Network

An artificial neural network (ANN) is a computational model that draws inspiration from the structure and functioning of biological neural networks, akin to the human brain. [6][19]. It is a key component of the broader field of machine learning, particularly in the subfield of deep learning. ANNs consist of interconnected nodes, also known as artificial neurons or perceptrons, organized into layers [19][20]. The fundamental components of an artificial neural network include: Perceptrons, Layers, Connection and Weights, Activation Function, Feedforward, Backpropagation and Learning. Each of these component plays an important role that enable ANN to learn and make complex decision [19].

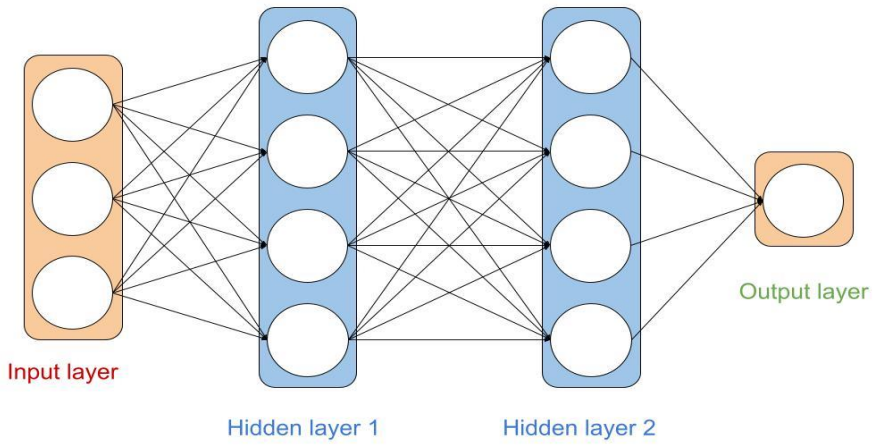

Figure 2.2 An ANN with two fully connected layers

Perceptrons

Perceptrons are the basic computational units within the network [19][20]. These nodes are inspired by the structure and functioning of biological neurons found in the human brain. In an ANN, nodes process and transmit information, allowing the network to learn patterns and make predictions. Each node receives one or more input signals, performs a mathematical operation on these inputs, and produces an output signal.

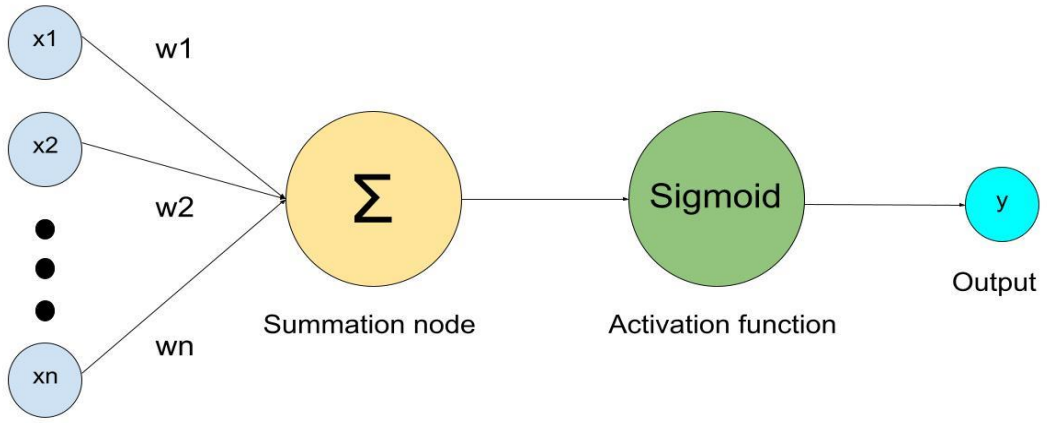

Figure 2.3 Structure of a Perceptron

Perceptrons, with their ability to process information and adjust their behavior during training, are the building blocks that enable neural networks to learn and perform a wide range of tasks, from image recognition to natural language processing [20]. The complex interactions between nodes in deep neural networks contribute to the network's capacity to understand and represent intricate patterns in data.

Layers

ANNs are organized into layers, layers are organized and structure the flow of information through the network [19][21]. Each layer in a neural network has a specific role in processing and transforming input data to produce the final output or prediction. The standard configuration of the network comprises an input layer, one or more hidden layers, and an output layer. The input layer takes in the initial data and the output layer generates the final result or prediction.

The input layer serves as the starting layer that accepts the raw input data. Each node in the input layer represents a feature or attribute of the input data [19]. The dimensionality of the input data dictates the number of nodes in the input layer. Hidden layers are located between the input and output layers. They play a crucial role in learning complex representations and patterns from the input data. A neural network with one or more hidden layers is referred to as a "deep" neural network [20]. The nodes in these layers process information and transmit it to subsequent layers. Finally, the output layer is the final layer of the neural network. It produces the network's output or prediction based on the processed information from the preceding layers [20]. The number of nodes in the output layer is dictated by the nature of the task, such as the number of classes in a classification task or the count of output values in a regression task.

Connections and Weights

Connections between nodes in adjacent layers are represented by weights [19][20]. The strength of the relationship between connected nodes is determined by the weight assigned to each connection. During training, these weights are adjusted to optimize the network's performance. In addition to the input, hidden, and output layers, many neural networks include bias nodes. Bias nodes contribute a constant value to the input of nodes in the following layer, helping the network account for variations and improve flexibility during training [20]. The architecture of a neural

network, defined by the number of layers, the number of nodes in each layer, and the connections between nodes, is crucial in determining the network's capacity to learn and generalize from data. The depth of a neural network—determined by the number of hidden layers—often influences its ability to capture hierarchical representations of complex patterns.

Activation Function

An activation function is a crucial component of artificial neural networks (ANNs) that introduces non-linearity to the model [19]. It operates on the weighted sum of the inputs to a node (or neuron) and determines the node's output. Activation functions play a key role in allowing neural networks to learn complex patterns and relationships in data. Here are some commonly used activation functions:

The step function is a simple binary activation function [19][21]. It outputs 1 if the input is greater than or equal to a certain threshold and 0 otherwise. While historically used in early models, it is less common in modern neural networks due to its lack of differentiability.

The sigmoid function squashes the input to a range between 0 and 1. It is particularly useful in the output layer of binary classification tasks, where the network needs to produce probabilities. However, it can suffer from the vanishing gradient problem during training. The formula for sigmoid function is as below:

$$Sigmoid(x) = \frac{1}{1+e^{-x}}$$

Similar to the sigmoid function, the tanh function squashes the input, but it ranges between -1 and 1 [19][20]. Tanh is often used in hidden layers of neural networks because it helps mitigate the vanishing gradient problem better than the sigmoid. The formula for tanh function is as below:

$$Tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$$

ReLU is a widely used activation function that outputs the input directly if it is positive, and zero otherwise [19]. ReLU is computationally efficient and helps address the vanishing gradient problem. However, it can suffer from the "dying ReLU" problem, where neurons may become inactive during training. The formula for ReLU function is as below:

$$ReLU(x) = \max(0, x)$$

Leaky ReLU is a variant of ReLU that allows a small, positive slope for negative inputs [19][20]. This helps mitigate the dying ReLU problem by allowing a small gradient for negative values. The formula for leaky ReL function can be written as:

$$Leaky\ ReLU(x) = \max(\alpha x, x)$$

where α is small positive constant.

PReLU is an extension of Leaky ReLU where the slope is learned during training rather than being a fixed hyperparameter [20]. The formula for leaky ReLU function can be written as:

$$PReLU(x) = \max(\alpha x, x)$$

where α is a learnable parameter.

The softmax function is often used in the output layer for multi-class classification tasks [19][20]. It converts the raw output scores into probabilities, ensuring that the sum of the probabilities across classes is equal to 1. Formula for softmax function can be written as:

$$Softmax(x) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Choosing an activation function depends on the specific task, the nature of the data, and considerations such as the potential for vanishing gradients or dead neurons. Experimentation and tuning are often required to determine the most suitable activation function for a particular neural network architecture.

Feedforward

The process of passing data through the network from the input layer to the output layer is known as feedforward [19]. During training, backpropagation is used to adjust the weights based on the calculated error, minimizing the difference between the predicted and actual outputs. The process begins with the input layer, where the network receives the initial data. Each node in the input layer represents a feature or attribute of the input data [20][21]. The input data is multiplied by weights associated with the connections between nodes in the input layer and the nodes in the next (hidden) layer. These weights represent the strength of the connections. The weighted sum of the

inputs is passed through an activation function at each node in the hidden layers. The activation function introduces non-linearity, allowing the network to learn complex patterns and relationships [19][20]. The processed information from the hidden layers is further multiplied by weights and passed through an activation function in subsequent hidden layers until reaching the output layer. The output layer produces the final result or prediction. The nodes in the output layer produce the final output based on the processed information. The number of nodes in the output layer depends on the nature of the task, such as the number of classes in a classification task or the number of output values in a regression task [20]. The feedforward process is essentially a series of matrix multiplications and non-linear transformations. During training, the network learns optimal weights through a process called backpropagation, where the difference between the predicted and actual outputs (error) is used to adjust the weights and improve the model's performance [19].

Loss Function

In machine learning, a loss function, also referred to as a cost or objective function, serves as a pivotal metric for evaluating the alignment between a model's predictions and the actual values of the target variable [19]. During the model training phase, the overarching objective is to minimize the value of this function. The loss function essentially quantifies the dissimilarity between predicted and actual outcomes, providing a numerical benchmark for optimization algorithms to systematically reduce this disparity [20]. The choice of a specific loss function is contingent upon the type of machine learning task at hand, such as classification or regression, and the inherent characteristics of the problem under consideration. Common loss functions include Binary Cross-Entropy for tasks such as binary classification, Mean Squared Error for regression tasks and Categorical Cross-Entropy for multi-class classification tasks [19][20].

Each loss function caters to specific machine learning scenarios and problem domains. For instance, Mean Squared Error gauges the average squared deviation between predictions and actual values in regression tasks, while Binary Cross-Entropy quantifies the divergence between binary labels and predicted probabilities in binary classification [19]. Similarly, Categorical Cross-Entropy extends this concept to multi-class classification settings. The diverse landscape of loss functions allows practitioners to tailor their choice based on the details of the data and the comprehensive goals of the machine learning endeavor, ensuring a nuanced and effective approach to model optimization.

Backpropagation

Backpropagation, is a commonly used technique that used to train artificial neural networks (ANNs) [19]. It is a key component of the learning process in which the model learns to adjust its parameters to minimize the difference between its predicted outputs and the true labels of the training data. Backpropagation is a form of gradient descent optimization. During the forward pass, input data is passed through the neural network, layer by layer, to produce predictions. The weighted sum of inputs is computed at each neuron, and an activation function is applied to produce the output [19][20]. The predicted outputs are compared to the true labels, and a loss function (also known as a cost function or objective function) is used to quantify the difference between the predictions and the actual values. The loss function represents how well or poorly the model is performing on the training data. The backward pass involves computing the gradient of the loss with respect to the model's parameters [19]. This process starts from the output layer and moves backward through the network to update the weights and biases. The partial derivatives of the loss with respect to each parameter are calculated using the chain rule of calculus. The weight update can be calculated using the following formula:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

where:

- E is the error or loss function that measures the difference between the predicted output and the actual output.
- w_{ij} is the weight connecting the i-th perceptron in the previous layer to the j-th perceptron in the current layer.
- η is the learning rate.
- $\frac{\partial E}{\partial w_{ij}}$ is the partial derivative with respect to weigh w_{ij}.

The gradients obtained in the backward pass are used to update the model's parameters (weights and biases) in the direction that reduces the loss. This update is performed using an optimization algorithm, typically a variant of stochastic gradient descent (SGD). These steps are repeated iteratively for multiple epochs or until convergence [19][20]. In each iteration, the model refines its parameters to improve its ability to make accurate predictions on the training data. Backpropagation is crucial for training deep neural networks, especially those with multiple layers (deep learning models). It allows the model to learn and adjust the weights and biases in a way that minimizes the error between predicted and actual values [19]. The backpropagation algorithm is computationally efficient and has been a key factor in the success of deep learning applications. Additionally, variants of backpropagation may include techniques like mini-batch training (updating weights based on a subset of the training data at a time) and regularization methods to prevent overfitting [19].

Learning

Artificial neural networks learn from data by adjusting the weights of connections between nodes [19]. This learning process involves iteratively presenting input data, making predictions, calculating errors, and updating the weights to improve performance until satisfactory performance is achieved. Continuous learning may occur as the model encounters new data, allowing it to adapt to evolving patterns and trends.

Summary

Artificial neural networks can be designed for various tasks, including classification, regression, pattern recognition, and more [19][21]. The depth of an artificial neural network, determined by the number of hidden layers, distinguishes shallow networks from deep networks. Deep neural networks, often referred to as deep learning models, have demonstrated remarkable success in tasks such as image recognition, natural language processing, and speech recognition.

The structure and functioning of artificial neural networks draw inspiration from the way biological neurons in the human brain process and transmit information [20]. However, it is essential to note that while ANNs are inspired by biological systems, they are highly simplified and not direct replicas of the complex neural networks found in living organisms.

2.6.2 Convolutional Neural Network

A Convolutional Neural Network (CNN or ConvNet) is a specialized type of artificial neural network designed for processing and analyzing visual data, such as images and videos [22]. CNNs are particularly effective in tasks related to computer vision, including image recognition, object detection, segmentation, and image generation. They have played a crucial role in advancing the state of the art in these domains.

The architecture of a CNN is characterized by the use of convolutional layers, which apply convolution operations to input data using filters or kernels. These filters detect local patterns and features in the input, allowing the network to automatically learn hierarchical representations. The convolutional layers are typically followed by activation functions, such as Rectified Linear Unit (ReLU) [22][23], which introduce non-linearity to the model and enable it to capture complex relationships in the data.

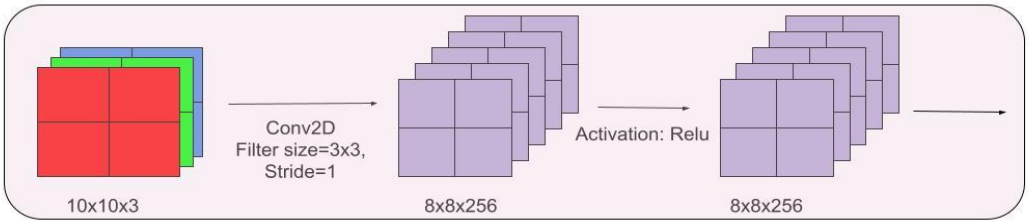

Figure 2.4 An example of a CNN structure

In addition to convolutional layers, CNNs often incorporate pooling layers, such as max pooling or average pooling, to downsample the spatial dimensions of the data and reduce computational complexity. This pooling operation helps the network focus on the most essential information while preserving important features [21][22]. Fully connected layers at the end of the network process the high-level abstractions learned by the previous layers to make final predictions or classifications. The strength of CNNs lies in their ability to automatically learn spatial hierarchies of features, allowing them to excel in tasks such as image recognition, object detection, and image segmentation.

Several influential CNN architectures have contributed to the advancement of computer vision tasks. AlexNet, which won the ImageNet Large Scale Visual Recognition Challenge in 2012,

played a pivotal role in popularizing CNNs [21]. Following that, models like VGG, GoogLeNet (Inception), and ResNet have further pushed the boundaries of accuracy and efficiency. These architectures differ in terms of depth, complexity, and design principles, but they all share the fundamental concepts of convolution, pooling, and hierarchical feature learning.

The success of CNNs extends beyond static images to tasks involving temporal data, such as video analysis, and their application has become pervasive in fields ranging from autonomous vehicles to medical image analysis [22][23]. Continued research and development in CNNs and related architectures continue to drive innovations in computer vision and shape the landscape of artificial intelligence.

2.6.3 U-Net

U-Net is a specialized architecture in the field of computer vision and image segmentation, designed to address challenges in medical image analysis [24]. Developed by researchers at the Computer Science Department of the University of Freiburg, U-Net gets its name from its distinctive U-shaped architecture. The primary application of U-Net is semantic segmentation, where the goal is to classify each pixel in an image into specific classes. This architecture has found widespread use in medical image segmentation tasks, such as identifying and delineating structures in MRI or CT scans, due to its ability to handle limited labeled data effectively [25].

The U-Net architecture consists of a contracting path, a bottleneck, and an expansive path. The contracting path, resembling an encoder, captures contextual information by using convolutional and pooling layers to reduce spatial dimensions. The bottleneck, often referred to as the central layer, retains high-level features and acts as a bridge between the contracting and expansive paths. The expansive path, resembling a decoder, uses transposed convolutions and upsampling to restore spatial dimensions while refining the segmentation output. Skip connections connect

corresponding layers in the contracting and expansive paths, facilitating the transfer of high-resolution information and aiding in precise localization.

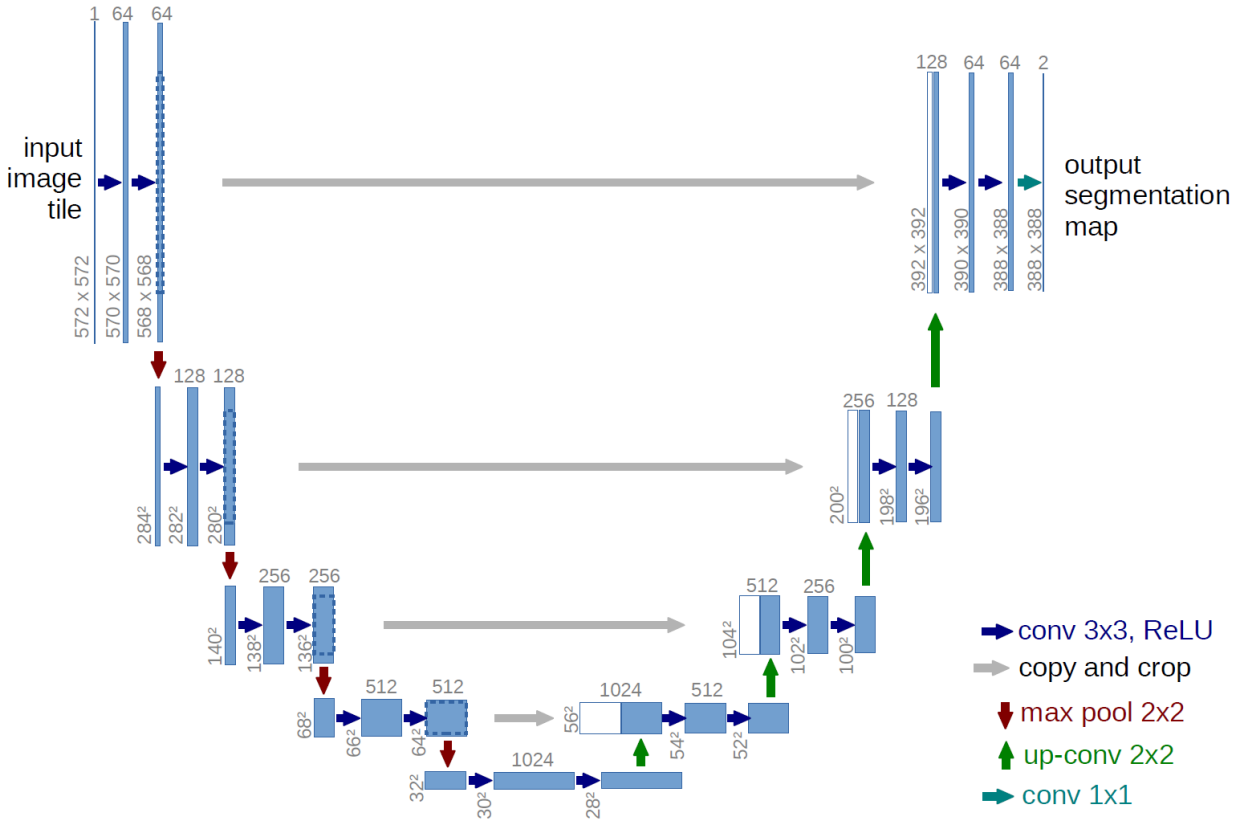

Figure 2.5 An example of U-Net architecture. *Image Credit: [29]*

One of the distinctive features of U-Net is its skip connections, which allow the network to combine both low-level and high-level features during the upsampling process, preserving fine details in the segmentation [24]. This architecture mitigates the common issue of information loss during downsampling, making U-Net particularly effective for tasks where precise localization is crucial. U-Net's success has spurred various modifications and adaptations in the design of neural network architectures, influencing the development of subsequent models for semantic segmentation and image-to-image tasks.

2.6.4 Diffusion Model

Diffusion models represent a category of generative AI models capable of producing high-resolution images [26]. Their functioning involves introducing Gaussian noise to the original data and then learning to eliminate this noise through both forward and reverse diffusion processes (Figure 2.4). Alternatively termed noise conditional score networks (NCSN) or score-matching with Langevin dynamics (SMLD), diffusion models excel in generating data resembling that on which they were trained, capable of transforming noise into coherent images [26][27].

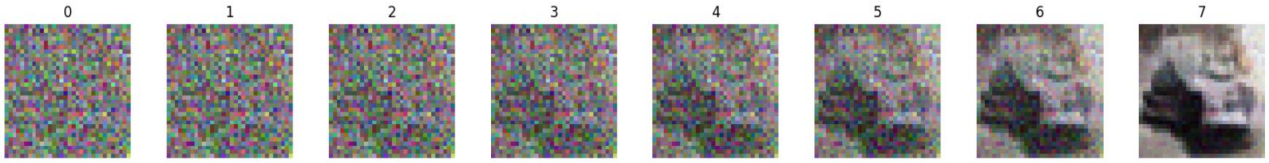

Figure 2.6 Denoising process of diffusion model

Notably distinct from normalizing flow, where both forward and backward processes are deterministic, diffusion models employ a fixed forward process and a trainable backward process, both of which are stochastic. This unique characteristic allows diffusion models to collaborate with large language models (LLMs) to interpret text prompts and produce high-quality images [27]. Leveraging the strengths of LLMs in comprehending natural language prompts, diffusion models contribute their expertise in generating visually compelling images.

2.7 Brain Inspired Neural Network Architectures

A human-inspired neural network architecture seeks to mimic the structure and functionality of the human brain, particularly the neocortex, which is integral to higher-order cognitive functions. One such architecture is the Hierarchical Temporal Memory (HTM), developed by Numenta [28]. The neocortex is known for its hierarchical organization, sparse connectivity, and remarkable ability to learn and recognize complex patterns, making it a rich source of inspiration for creating intelligent systems. The HTM architecture incorporates several key principles observed in the neocortex. Firstly, it embraces the concept of sparse distributed representations, where only a small fraction of neurons is active at any given time [28]. This sparsity enables efficient memory storage and retrieval, allowing the system to represent a vast array of patterns without overwhelming computational resources. Additionally, the HTM architecture is designed with a hierarchical structure, resembling the layered organization of the neocortex. Information is processed in a series of levels, with each level capturing increasingly abstract features of the input data.

Temporal memory is a critical component of HTM, enabling the model to understand and predict sequences of patterns over time [28]. This aligns with the brain's natural ability to comprehend and learn temporal dependencies, crucial for tasks involving sequential data. Moreover, HTM supports online learning, enabling the model to adapt and learn in real-time as it encounters new data. This online learning capability reflects the brain's continuous learning process, allowing it to dynamically adjust to changing environments.

While human-inspired neural network architectures like HTM draw inspiration from the brain's structure, it is essential to note that these models are abstractions and simplifications of the complex biological system [28]. Researchers and engineers leverage these principles to develop

intelligent systems that can perform tasks such as pattern recognition, anomaly detection, and prediction.

Although not as mainstream as traditional neural network architectures, human-inspired models contribute to the exploration of novel approaches in the field of artificial intelligence, aiming to capture the essence of how the human brain processes information and learns patterns.

Chapter 3

Methodology

Having established a fundamental understanding of the literature and the individual components utilized in the experiment, we can now delve into the methodology. Our chosen method involves the application of an artificial neural network, more specifically the U-Net architecture, to execute the diffusion model. The objective is to derive insights into the functioning of the human imagination process. This section will first outline the data collection procedure, followed by a detailed description of the simulation subsection. Within this subsection, a conceptual process will be outlined, illustrating the formation of the workflow that resulted in the implementation of the experiment.

3.1 Data Collection

To conduct the model training, I employed the CIFAR-10 dataset sourced from the TensorFlow library. This publicly accessible dataset comprises a total of 60,000 color images, each with dimensions of 32x32, and spans across 10 distinct classes. The classes are systematically labeled from 0 to 9, representing diverse categories such as Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, and Truck. Within each class, there are 6,000 images. For both the training and testing phases of the AI, the dataset includes 50,000 training images and 10,000 test images. The dataset is organized into five training batches and one test batch, each containing 10,000 images. The test batch is composed of 1,000 randomly selected images from each class. In the training batches, the remaining images are arranged in a random order. Although there may be variations in the number of images from different classes in specific training batches, collectively, the

training batches encompass 5,000 images from each class. To streamline the simulation process, only four out of the ten available classes were utilized and selected in pair. This decision was driven by time and resource constraints during the model training, while still ensuring the sufficiency of data to validate the proposed idea. Specifically, the four selected classes for the simulation were (bird, horse) and (airplane, automobile). These two pairs were selected to introduce more variability in the outcomes and to explore potential differences in learning rates and loss metrics when the details are more organic in one pair compared to the other.

3.2 Simulation

This section provides a detailed account of the complex process through which results are generated to validate the original theory. It begins by presenting the thought process, offering a comprehensive overview of the conceptual framework. Subsequently, it delves into the finer details, presenting a step-by-step breakdown of the algorithm's implementation. By thoroughly examining each phase of the process, the section aims to offer a better understanding of how the results were derived, thus establishing the foundation of the initial theoretical framework.

3.2.1 Conceptual Process

The diffusion model operates through a systematic process of refining a fully noisy image to enhance its overall quality gradually. This method involves a series of steps aimed at reducing or eliminating the inherent noise present in the initial image. The noise removal process is executed in a step-by-step manner, where each iteration contributes to the gradual improvement of the image's visual clarity.

At the initial stage, the image is inundated with noise, resulting in a visual representation that lacks clarity and coherence. The diffusion model then strategically employs algorithms and techniques to identify and diminish these undesirable elements within the image. This iterative refinement process continues until the image quality reaches a point of clarity where the inherent noise is significantly reduced, if not entirely eliminated.

The core concept underlying the diffusion model is the gradual progression from a visually obscured state to a clearer and more refined representation. This iterative approach allows for a meticulous reduction of noise, ensuring that the final output is a high-quality image with improved visual fidelity. In essence, the diffusion model acts as a transformative mechanism, progressively

enhancing the image by strategically addressing and mitigating the noise elements present in the initial, fully noisy state (Figure 3.1).

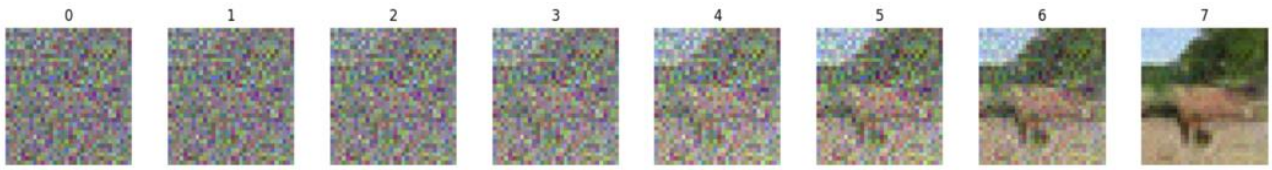

Figure 3.1 Recreate the image with diffusion model

With this understanding, it is feasible to construct a Deep Learning model designed specifically to elevate image quality. This model is envisioned to navigate the transformation process from a state of complete noise to the attainment of a clear and refined image, guided by a well-defined conceptual flow.

The fundamental idea underlying this Deep Learning model revolves around its ability to iteratively enhance image quality through learning to identify noise drawn from a normal distribution at each time step then subtract it from the noisy image to recover the previous less noisy version following Markovian property. The process commences with the introduction of a fully noisy image, characterized by visual distortions and imperfections. The model is then set to work through a sequence of algorithmic operations to identify and reduce the noise elements within the image (Figure 3.2). The flow is structured in a way that each stage of the model's operation contributes to the gradual refinement of the image. This includes the identification and suppression of noise patterns, the enhancement of crucial features, and the overall optimization of visual fidelity. As the model iterates through these stages, the image quality undergoes a transformative progression, ultimately reaching a state of clarity and improved visual integrity.

This conceptual framework highlights the Deep Learning model's proficiency not only in mitigating noise but also in progressively refining image quality. By integrating the insights discussed about the nuances of the human imagination process in section 2.3 on Human Imagination into the implementation phase, the model establishes a correlation between the creative facets of human imagination and the learning process of AI in synthesizing images. This correlation prompts an exploration to ascertain if the AI can demonstrate the ability to recreate images or even generate entirely new visual compositions through a process akin to human "imagination." In essence, this research represents a biologically-inspired approach within the realm of Deep Learning, seeking a deeper understanding of human intelligence.

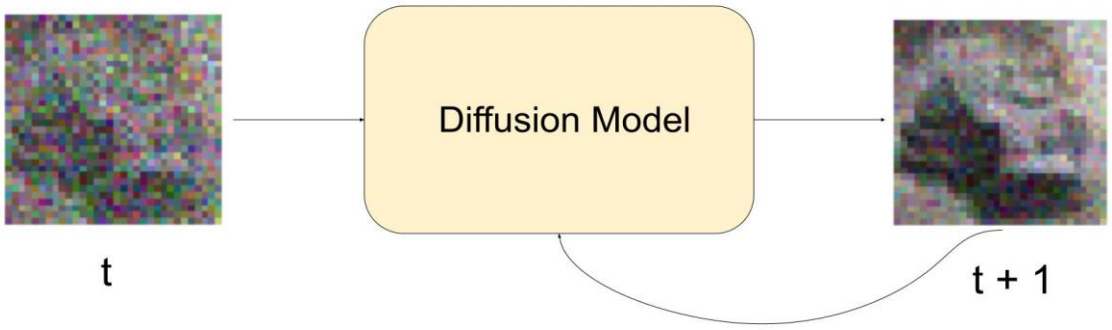

Figure 3.2 The iterative process of diffusion to eliminate noise from the image

3.2.2 Implementation

This section provides a detailed, step-by-step explanation of the implementation of the idea flow we have discussed thus far. Beginning with data preparation, the CIFAR-10 dataset is imported and split into training and testing sets. The training set consists of 50,000 sample images (5,000 per class), while the test set contains 10,000 sample images. Since our focus is on four classes, (bird, horse) and (airplane, automobile), the effective image count is 20,000 (Figure 3.3).

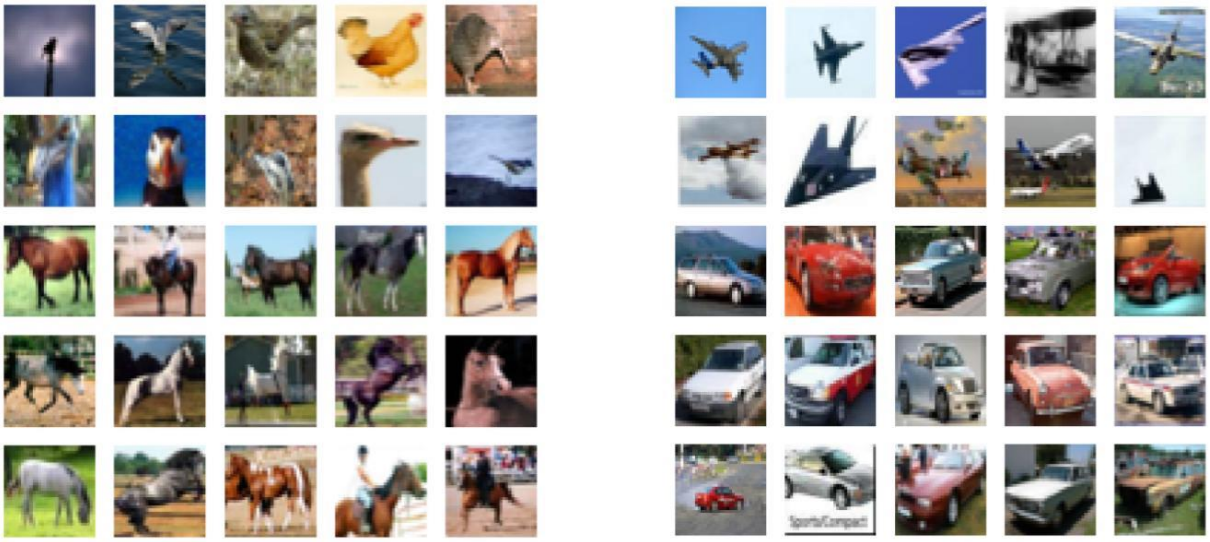

Figure 3.3 A batch of images from the CIFAR-10 dataset

Moving forward, we define several training constants crucial for our model. Firstly, the input image size is set to match the dimensions of CIFAR-10 images, which is 32x32 pixels. Subsequently, the batch size constant is determined to balance learning diversity and training efficiency. After testing, a batch size of 256 is chosen to enable the model to see a diverse range of images from each classes within reasonable training time and resource constraints.

Another vital constant is the time step, indicating how many steps our model takes to learn the transformation from noisy to clear images during the training process. After considering image size and final clarity through testing, a time step constant of 16 is selected. Lastly, the linear Beta scheduler constant determines the noise level at each time step, playing a crucial role in the forward process when introducing noise to the image.

With these constants established, the forward diffusion process is implemented. The approach involves passing the image through the forward diffusion function and applying the noise-inducing equation to obtain two images from random time points. At time t, one image is noisy, while at time t+1, it is a clearer version. The model learns to transform t into t+1 by removing the noise at that specific time point (Figure 3.4).

Algorithm 1: Simplified Forward Diffusion algorithm

Initialization: Given x,t:
(1) beta(t), beta(t+1) ⟵ beta scheduler value at t and (t+1)
(2) noise $\sim \mathcal{N}(\mu, \sigma^2)$
(3) img_t ⟵ x * (1-beta_t) + noise * beta_t
(4) img_(t+1) ⟵ x * (1-beta_(t+1)) + noise * beta_(t+1)
return img_t, img_(t+1)

Figure 3.4 The algorithm for forward diffusion process. *Based on:[26]*

Our subsequent objective involves the implementation of a modified U-Net model, incorporating multiple Convolutional Neural Network (CNN) layers for image downscaling and upscaling, along with dense layers for the multilayer perceptron section. The diagram of this model is outlined below:

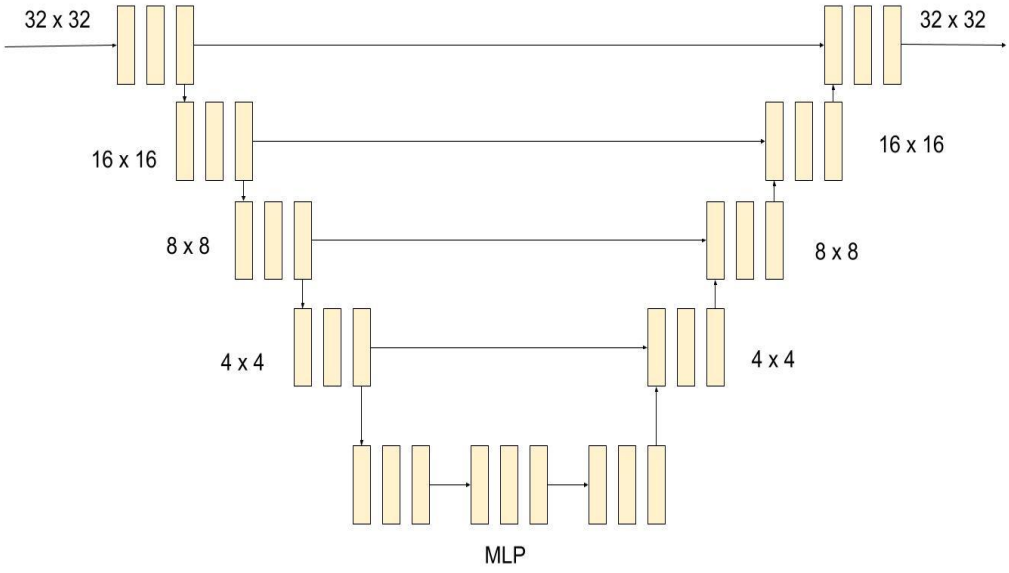

Figure 3.5 U-Net structure for the implementation

As all CNN blocks (Figure 3.6) share the same layer composition, we can consolidate this composition into a single function. This function incorporates two convolutional networks with a time parameter, enabling the network to discern its current time step and generate corresponding output information.

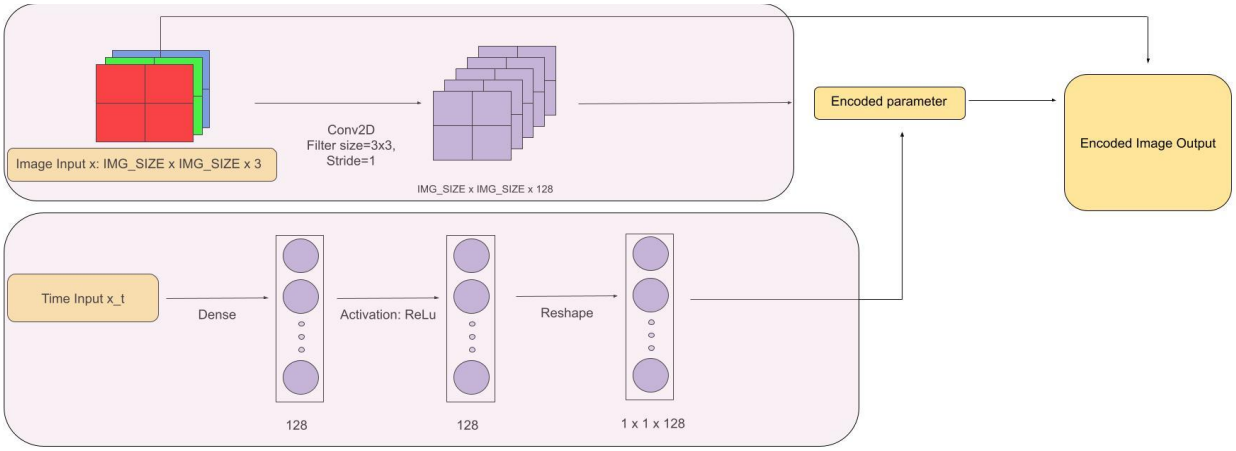

Figure 3.6 CNN block structure

Now that we have successfully implemented the fundamental building block, we can proceed to create the modified U-Net specifically designed for the reverse diffusion process. This involves leveraging the established building block to construct a U-Net architecture tailored to the task of transforming clear images into their corresponding noisy versions. The modified U-Net integrates the essential components, such as convolutional layers for upscaling, downscaling, and dense layers for the multilayer perceptron section, to facilitate the reverse diffusion process effectively. This network architecture is pivotal in enabling the model to learn the complex task of adding noise to images while maintaining clarity and coherence in the information it processes.

Upon completing the model implementation, we can proceed to create functions for both training and predicting the noisy image. The prediction process involves generating a noisy image and, for each time step, inputting both the image and its current time step into our model for prediction until we reach the final time step.

Algorithm 2: Prediction function algorithm

noisy image $\sim \mathcal{N}(\mu, \sigma^2)$
for *i in range time step* **do**
    time step t ⟵ i
    Predict the noisy image at the current time step t
end

Figure 3.7 The algorithm for prediction function

The training process is straightforward, with the provision of image input and time step input, enabling our model to learn the process of generating the denoised image with each batch of images. To penalize the error during training, a loss function is introduced to help with the learning process. This function plays a crucial role in penalizing any deviations between the model's output

and the target image. In this context, the mean absolute zero loss is chosen as a metric to measure the errors during the training sessions and help contribute to the model's ability to improve its performance and accurately generate denoised images over successive iterations.

Algorithm 3: Training function algorithm

```
Specify number of epochs
for i in range epochs do
    Get a batch of images specified by batch constant
    Call the train on batch function for training model
end
```

Figure 3.8 The algorithm for training function

At the heart of this process is the integration of image input, which serves as the raw data that our model utilizes to grasp the visual elements and patterns within the dataset. Simultaneously, the time step input plays a crucial role in guiding the model through the sequential progression of learning, allowing it to adapt and refine its denoising capabilities over time. As the model processes each batch of images, the iterative nature of the training process ensures that the neural network getting better and improve the denoising process. By training the model on diverse sets of images and associated time step inputs, we create a robust learning environment that enable the model to generalize its denoising skills, thus enhancing its proficiency in generating clear and refined images.

Chapter 4

Discussion

4.1 Experimental Results

After conducting the simulation for 15 epochs, which took approximately 5 hours on Google Colab, I have obtained multiple sets of denoised images. The loss exhibits a consistent decrease and the clarity of each output image progressively improves with each epoch (Figure 4.1). This trend indicates that the model is enhancing its performance with increased training iterations and exposure to a greater number of images (Figure 4.2, 4.3, 4.4). It is essential to highlight that, despite training on two distinct pairs of classes, the loss per epoch remains roughly consistent. This suggests that the learning process is not contingent on the specific class being trained, emphasizing a generalizability across different class pairs.

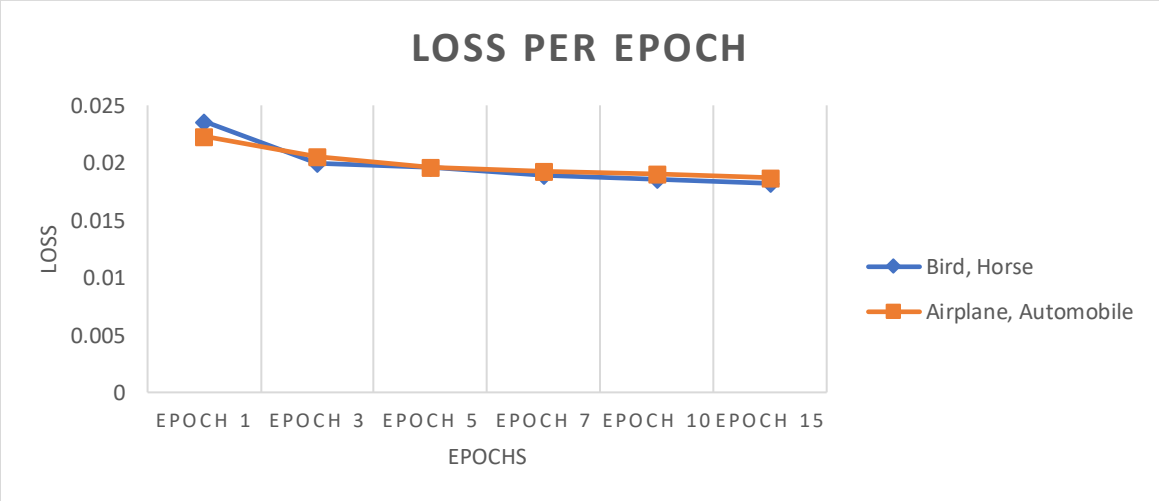

Figure 4.1 The loss decreases as the number of training epoch increases

Figure 4.2 Epoch 1 output result

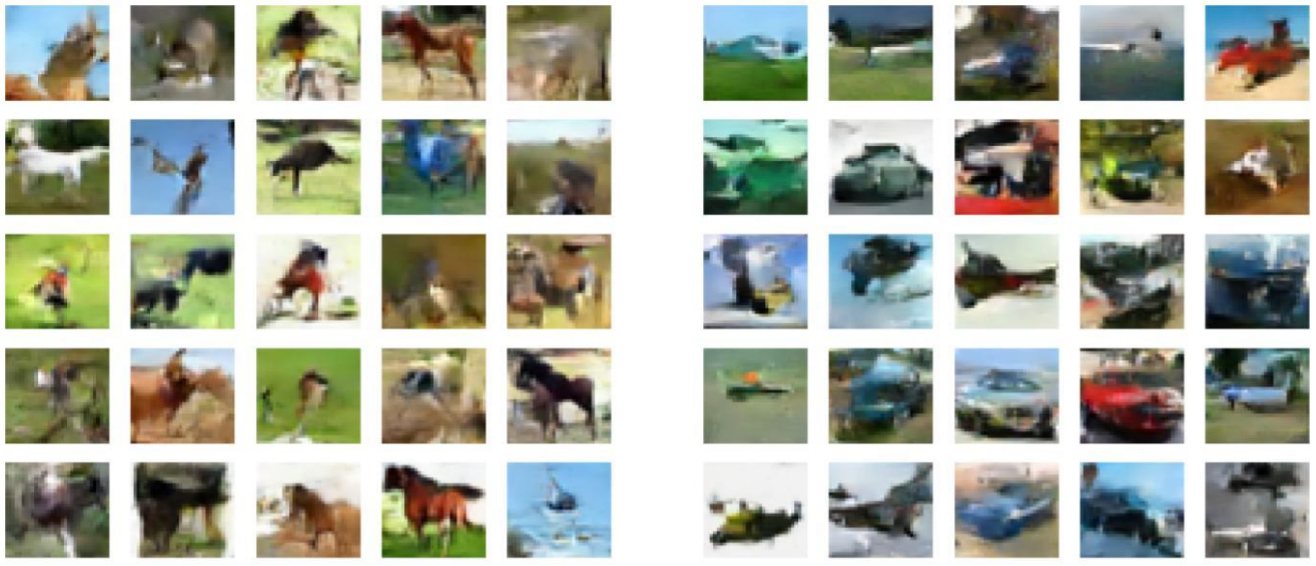

Figure 4.3 Epoch 7 output result

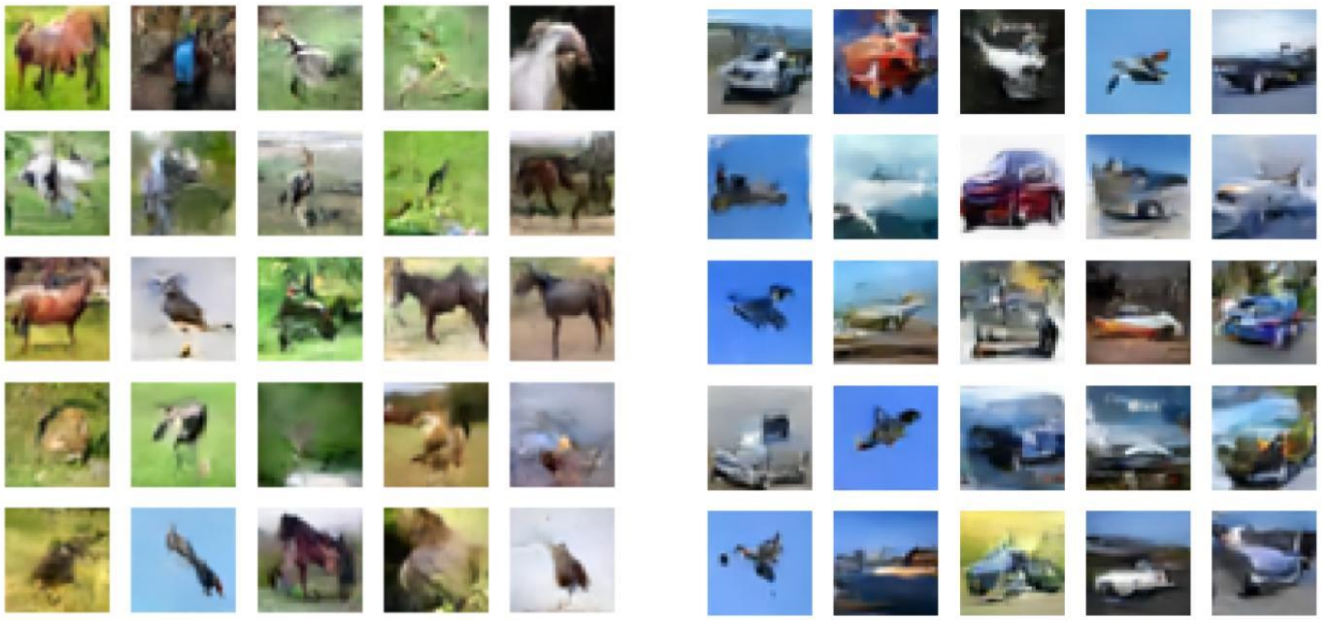

Figure 4.4 Epoch 15 output result

4.2 Actual result vs. Expectation

The initial hypothesis suggests that the human imagination process occurs when we observe the environment. When we see an object, our brain stores only the high-level features of the object, such as shape, texture, and color, rather than storing the object pixel by pixel, as computers do. With these high-level features, the brain can reconstruct a mental image of the object by combining these features and predicting finer details to complete the object. While this biological approach may not create the exact same image as the computer approach, it allows for more customization with the object information, leading to the combination of features from different objects to create entirely new objects and the ability to imagine new things.

With this concept in mind, utilizing our simplified diffusion model, we aim for the AI to learn to abstract important details from each image related to the object during the denoising process. Subsequently, it should apply these details when attempting to reconstruct the image from complete noise. Furthermore, to validate the theory of human imagination, the model was trained on different datasets of objects—specifically, the bird and horse datasets. This was done to demonstrate that the model can leverage its past knowledge, which comprises the high-level features it has learned, to combine and generate new images. Given that the details come from bird and horse images, the expectation is for the model to create an image that combines features from both animals.

Considering the low resolution of the results, attributed to the dataset consisting of 32 x 32 images and the absence of any implemented upscaling technique to enhance image quality, we still obtained satisfactory outcomes that validate the theory to some extent. As depicted in Figure 4.5 below, some images have been successfully recreated to depict a horse, a bird, an automobile, or an airplane.

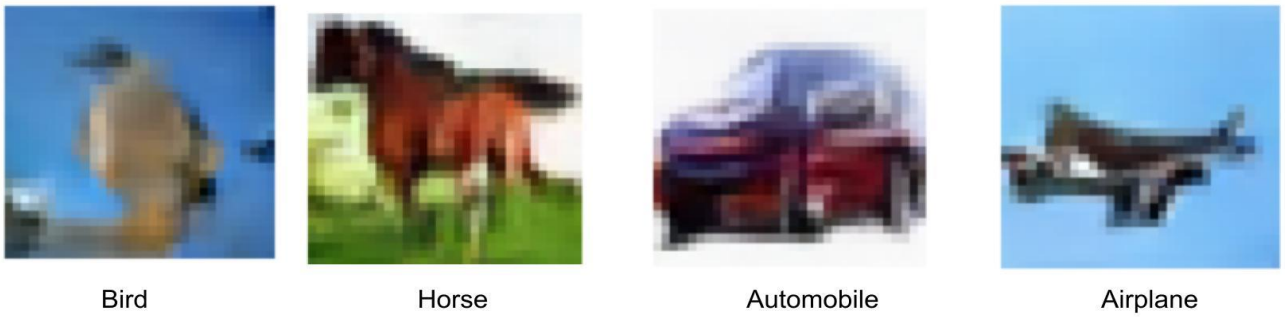

Figure 4.5 Single class objects

While Figure 4.6 displays images that exemplify a fusion of characteristics from diverse classes, such as a horse adorned with bird features on the far left and middle left pictures, or an automobile incorporating airplane attributes on the middle right and far right pictures.

Figure 4.6 Combined classes objects

Chapter 5

Conclusion & Future Works

5.1 Conclusion

Overall, this outcome demonstrates that, although the theory has not been conclusively proven, it has provided valuable insights into the complex capacity of the human mind within the realm of imagination. Through the application of the diffusion model, we have not only expanded our comprehension but have also been able to illustrate elements of the theory. This exploration has deepened our understanding of how the human brain processes information and reconstructs mental images, shedding light on the nuanced interplay of features in the imaginative process.

One notable aspect is the capacity of the diffusion model to abstract essential details from each image during the denoising process. By leveraging high-level features learned during training, the model demonstrates an ability to reconstruct images from complete noise, akin to the human brain's process of combining features to predict finer details and complete an object. This parallels the theoretical framework proposing that human imagination involves storing and combining high-level features rather than pixel-level details, allowing for more customized and creative reconstructions.

Despite the challenge of low resolution due to the dataset's limitations and the absence of upscaling techniques, the obtained results showcase the potential of the diffusion model to capture and utilize past knowledge. Training the model on different datasets, such as those featuring birds and horses, demonstrates its ability to combine high-level features from diverse sources and generate entirely

new images. These findings open avenues for further exploration, emphasizing the need for continued research to unravel the intricacies of human imagination and enhance the capabilities of artificial systems in this realm.

5.2 Future Works

As highlighted in section 1.2 on Motivation, this study is an integral part of my ongoing exploration into biologically inspired artificial intelligence. Specifically, it delves into the realm of Concept Learning (Figure 5.1), proposing an algorithm built upon the foundations of Reinforcement Learning. In this approach, the process commences with the analysis of visual data received by the agent's sensors. Subsequently, this visual information is harnessed and stored to reconstruct a representation of the surrounding environment, facilitating the navigation of the agent and culminating in the establishment of the agent's location awareness. This environmental representation encompasses aspects such as depth, properties of elements within the environment, and more. New objects are dissected into detailed components for the sake of storage and recall, while familiar objects aid the agent in its decision-making processes. The logical unit is engaged during the visual processing phase to predict and interpret the environment, such as determining that solid objects should rest on stable surfaces, or that elastic objects should rebound when thrown onto hard surfaces. Following the processing of sensor information to construct a coherent environment for the agent, the control center generates a state space that encapsulates the relative locations and properties of all elements. Leveraging reinforcement learning, the agent then learns to navigate and accomplish the intended tasks.

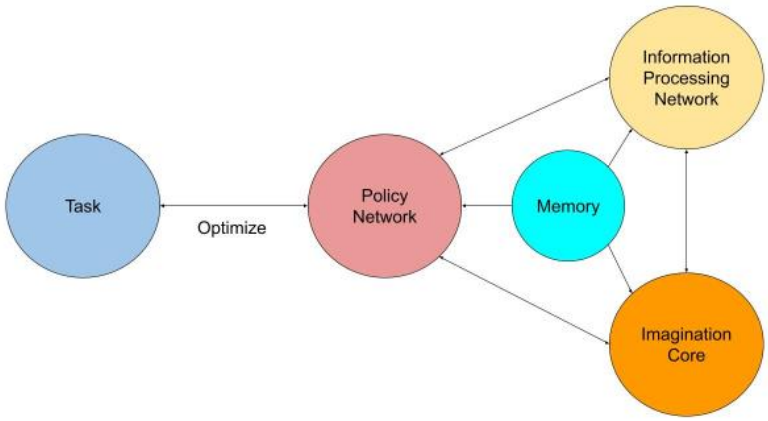

Figure 5.1 The flow diagram of Concept Learning

As part of future endeavors, the imagination process can extend its utility by aiding the agent in envisioning potential actions and their consequences in future states, fostering a deeper understanding of the environment and enhancing decision-making capabilities to accomplish current tasks more effectively.

Bibliography

[1] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," California Management Review, vol. 61, no. 4, pp. 5-14, 2019.

[2] N. Muthukrishnan et al., "Brief history of artificial intelligence," Neuroimaging Clinics, vol. 30, no. 4, pp. 393-399, 2020.

[3] I. M. Cockburn, R. Henderson, and S. Stern, "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis," in The Economics of Artificial Intelligence: An Agenda, University of Chicago Press, pp. 115-146, 2018.

[4] T. Phillips et al., "Exploring the Use of GPT-3 as a Tool for Evaluating Text-Based Collaborative Discourse," in Companion Proceedings of the 12th, vol. 54, 2022.

[5] A. Konar, "Artificial Intelligence and Soft Computing: Behavioral and Cognitive Modeling of the Human Brain," CRC Press, 2018.

[6] J. Fan et al., "From Brain Science to Artificial Intelligence," Engineering, vol. 6, no. 3, pp. 248-252, 2020.

[7] D. S. Bassett and M. S. Gazzaniga, "Understanding complexity in the human brain," Trends in Cognitive Sciences, vol. 15, no. 5, pp. 200-209, 2011.

[8] B. J. Baars and N. M. Gage, Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience. Academic Press, 2010.

[9] A. Rose, Vision: Human and Electronic. Springer Science & Business Media, 2013.

[10] S. Buetti, et al., "Towards a Better Understanding of Parallel Visual Processing in Human Vision: Evidence for Exhaustive Analysis of Visual Information," Journal of Experimental Psychology: General, vol. 145, no. 6, pp. 672, 2016.

[11] E. Pelaprat and M. Cole, "“Minding the gap”: Imagination, creativity and human cognition," Integrative Psychological and Behavioral Science, vol. 45, pp. 397-418, 2011.

[12] O. Vartanian, A. S. Bristol, and J. C. Kaufman, Eds., Neuroscience of Creativity. MIT Press, 2013.

[13] A. Fuentes, "The evolution of a human imagination," in The Cambridge Handbook of the Imagination, pp. 13-29, 2020.

[14] A. Voulodimos et al., "Deep learning for computer vision: A brief review," Computational intelligence and neuroscience, 2018.

[15] R. Szeliski, Computer Vision: Algorithms and Applications. Springer Nature, 2022.

[16] B. Mahesh, "Machine learning algorithms-a review," International Journal of Science and Research (IJSR), vol. 9, no. 1, pp. 381-386, 2020.

[17] M. Mohammed, M. B. Khan, and E. B. M. Bashier, "Machine learning: algorithms and applications," CRC Press, 2016.

[18] B. Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), vol. 9, no. 1, pp. 381-386, 2020.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[20] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," MIT Press, 2016.

[21] O. I. Abiodun, et al., "State-of-the-art in Artificial Neural Network Applications: A Survey," Heliyon, vol. 4, no. 11, 2018.

[22] Z. Li et al., "A survey of convolutional neural networks: analysis, applications, and prospects," IEEE Transactions on Neural Networks and Learning Systems, 2021.

[23] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput., 29(9):2352–2449, September 2017.

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

[25] N. Siddique et al., "U-net and its variants for medical image segmentation: A review of theory and applications," IEEE Access, vol. 9, pp. 82031-82057, 2021.

[26] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems 33, pp. 6840-6851, 2020.

[27] A. Jacob, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured Denoising Diffusion Models in Discrete State-Spaces," in Advances in Neural Information Processing Systems, 2021.

[28] J. Wu, W. Zeng, and F. Yan, "Hierarchical temporal memory method for time-series-based anomaly detection," Neurocomputing, vol. 273, pp. 535-546, 2018.

[29] Ronneberger, O. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/, 2015. University of Freiburg. Accessed on: 2023-12-04.

[30] M. Poo, "Towards brain-inspired artificial intelligence," National Science Review, vol. 5, no. 6, pp. 785-785, 2018.

[31] L. Zhao et al., "When brain-inspired AI meets AGI," Meta-Radiology, vol. 100005, 2023.