Exploring Neural Network Architectures for fMRI Pattern Recognition Using Modern Machine Learning and Deep Learning Techniques

Kathleen Wang

Received September 28, 2023
Accepted February 21, 2024
Electronic access February 29, 2024

As technology progresses, "mind reading" is moving from the realm of science fiction into reality. In this paper, I aim to ask what modern machine learning techniques contribute to decoding fMRI recordings of the brain perceiving images. In the realm of fMRI analysis methodologies, the adoption of various machine learning techniques has proven effective in reconstructing and categorizing images from fMRI data. By applying various machine learning (ML) techniques to fMRI data, researchers have been able to reconstruct and categorize images based only on the brain activity evoked by seeing those images. In some cases, there has even been a relationship between the ML models and the neural underpinnings of hierarchy in human perception. Image categorization has also offered insight into the brain's classification abilities. However, these techniques face limitations: reconstructed images are often imperfect and categorization outcomes vary in accuracy. Looking ahead, using modern machine learning techniques to detect emotions, decode thoughts, and translate abstract concepts could offer promise in potentially aiding medical researchers in the diagnosis of brain disorders and driving advancements in intelligent systems. Integrating fMRI data with artificial intelligence models and combining fMRI with electroencephalography (EEG) data could offer real-time insights into neural perception dynamics, providing deeper cognitive insights. In addition to providing an overview of Convolutional Neural Networks (CNNs), this paper will analyze five research experiments that have been proven successful in either reconstructing or categorizing images based on fMRI data while utilizing different machine-learning techniques.

Introduction

In science fiction, futuristic concepts such as mind-reading and communication via thoughts have long since captivated audiences. These kinds of fascinating ideas have served as inspiration for researchers in neuroscience and artificial intelligence. With the advancement of machine learning and developments in the process of analyzing functional magnetic resonance imaging (fMRI) data, this vision has become more possible to achieve in reality. Already today, researchers are using machine learning algorithms to decode patterns in fMRI data to understand and reproduce specific images.

In the field of modern machine learning, the algorithms and techniques created enable computers to learn from data and make predictions or decisions without being explicitly programmed[1]. It encompasses various approaches, including neural networks, deep learning, random forests, and support vector machines. Modern machine learning techniques have revolutionized numerous fields, such as recommendation systems, image and speech recognition, medical diagnosis, and natural language processing, by extracting meaningful patterns and insights from large datasets. These techniques rely on statistical modeling and optimization algorithms to iteratively learn from examples and improve their performance over time[2]. They have proven to be highly effective in solving complex problems and have contributed to significant advancements in artificial intelligence and data-driven decision-making.

When combined with biology, technology can be used to help uncover the intricate mysteries of the human brain, offering insights into the complex interactions between neural signals, physiological responses, and cognitive functions. The vision we see is images our brain creates from electrical signals passed along by our neurons. During the process, blood flows to sections of the brain that rely on cells working and using oxygen. The more oxygen is consumed the more blood flows to the area. To create an fMRI recording, blood oxygen level dependent signals, an MRI machine is used to track the blood flow in the brain. Sites that appear brighter in the recordings have high blood flow, hence more neural processing.

The following machine learning techniques utilize fMRI recordings as data for their training and testing. Since fMRI data can be noisy and there can be variability across individuals and even scanning sessions, machine learning models may struggle to generalize well across diverse datasets. This is why the success of machine learning models depends heavily on the quality of data preprocessing. Choosing appropriate preprocessing steps is critical, and the impact of preprocessing choices on model performance is not always straightforward. Machine learning models also excel at recognizing complex patterns in large datasets, which is valuable for identifying subtle and

distributed patterns of brain activity associated with specific cognitive processes. The algorithms learn to categorize images by analyzing the fMRI recordings, focusing on the special features in the recordings.

In the brain's hierarchy for processing visual information, there is both sequential and hierarchical organization. Deep learning, particularly in the context of computer vision, attempts to emulate and mimic this natural processing hierarchy. The idea is to create artificial neural networks with multiple layers (hence the term "deep" learning) that progressively extract and learn hierarchical features from raw data, similar to how the human brain processes visual information in a hierarchical manner. This approach allows deep learning models to automatically learn complex representations and patterns from visual data, making them effective for tasks such as image recognition and object detection. Based on the paper "Deep image reconstruction from human brain activity", a similar approach seems effective for recognizing images from fMRI recordings[3].

When humans see or imagine physical objects, their brain reacts in similar ways after being stimulated by viewing or imagining an image of the same object. However, since the fMRI recordings taken of the brain differ from person to person and the resulting model classified images incorrectly often. Further, certain classes are generally harder to recognize. Perhaps, this could be because the classification is not commonly seen in daily life, or the classification is so narrow in scope that the fMRI recordings are too difficult to distinguish. When a classification is not commonly encountered in our daily lives, the brain's response to it might be less distinct and predictable. As a result, the fMRI signals generated could appear more ambiguous and challenging to interpret accurately, leading to potential misclassifications in the model. Additionally, the developmental disparity between the brains of younger and older individuals, coupled with the diverse stressors and life perspectives across different age groups, can lead to differences in fMRI recordings. This necessitates machine learning algorithms to be programmed with an awareness of these variations. Alternatively, a sufficiently diverse training dataset must be compiled to enable the generalization of fMRI training across individuals of all ages and mental states.

A recent breakthrough presented a new frontier in decoding language-related brain activity, moving beyond basic word-level decoding to more complex linguistic units. The research emphasized the significance of understanding the neural mechanisms associated with higher-order language processing. They employed recurrent neural networks (RNNs) to decode the semantic content of continuous sentences, contributing to our comprehension of how the brain processes language on a more sophisticated level[4]. RNNs are a special type of neural network that are able to maintain a memory of previous inputs through the use of connections that loop back on themselves. RNNs are particularly suitable for tasks involving sequences or time-dependent patterns, such as speech recognition, time series analysis, and natural language processing, because of their memory feature. The ability of RNNs to process sequential information also makes them well-suited for tasks where the order of data points matters, as they can capture context and dependencies across different time steps.

While we may not yet be recreating human memories from science fiction, we can now tell when people are looking at specific images and patterns, which is further along than we could've dreamed of forty years ago. In one of the earliest experiments that I reviewed here, Miyawaki et al.[5] chose to use a linear combination method to reconstruct visual images defined by binary contrast patterns because they believed that previous retinotopy based approaches were inadequate. One later paper concludes that it is possible to train a deep neural network (DNN) to reconstruct both natural and artificial images from functional magnetic resonance imaging (fMRI) recordings through the use of hierarchical features and a deep generative network (DGN)[3]. Another paper, meanwhile, shows that it is possible to decode object categories from brain activity using fMRI and visual features through the application of hierarchical visual features and four different ML models[6]. Another group of researchers was able to reproduce a movie by utilizing a Bayesian decoding approach on the invoked fMRI when a human watches a natural movie[7]. In the same vein, a research group used a Generative Adversarial Network (GAN) to successfully reconstruct seen-images from fMRI data[8]. In this paper, I will explore how all these studies extracted information from fMRI recordings.

Bayesian Approach

An early approach to image reconstruction was to reconstruct movie scenes using a Bayesian approach from blood oxygen level dependent (BOLD) signals obtained through fMRI recordings[7]. The researchers employed a new motion energy encoding model on three humans with normal or corrected-to-normal vision, collecting both training and test data over nine sessions, each lasting 10 minutes.

This approach was chosen because while fMRI is the best method to collect brain data from a live human being, it is not very effective in capturing brain activity during natural vision or dynamic mental processes. The new encoding model was devised to overcome these limitations by separating the neural mechanisms mediating visual motion information from the slower hemodynamic mechanisms. By doing so, it focuses on capturing the specific neural activities related to motion perception, while minimizing interference from slower hemodynamic responses. The separation allows for a more accurate representation of brain activity during motion perception. It also provides researchers with a clearer and more precise understanding of how the brain processes visual motion information. By overcoming the limitations of capturing both fast neural responses and

slow hemodynamic responses, the new motion energy encoding model enhances the quality and reliability of brain activity recordings during dynamic visual tasks, contributing to a better understanding of the brain's functioning.

To gauge the performance of their approach, the researchers tested the accuracy of the model's identification. The output timestamp was deemed correct if it was within one second of the actual video timestamp. On natural images, the model was able to determine the specific movie stimulus with an accuracy of around 95% for Subject 1. In addition, the researchers assessed the temporal specificity of the estimated motion-energy encoding model by examining its identification accuracy. Identification accuracy refers to how effectively the model can correctly link an observed BOLD signal pattern to the specific stimulus that elicited it. The identification accuracy for all three subjects was equal to or greater than 75%.

The paper found that using the averaged high posterior (AHP) estimate created a better reconstruction than using the maximum a posteriori (MAP) estimate. The AHP and MAP estimates are statistical concepts used in the context of data analysis, particularly in Bayesian statistics. The MAP estimate represents the most probable value of a parameter in a Bayesian model given the observed data and prior information[9]. The MAP estimate balances the likelihood of the observed data with prior beliefs about the parameter, aiming to find the parameter value that is most consistent with both. The AHP estimate, on the other hand, involves considering a range of parameter values around the MAP estimate. Instead of focusing solely on the single parameter value that maximizes the posterior probability, the AHP averages the values in a distribution of values that are more probable according to the posterior distribution, providing a broader and potentially more reliable estimation of the parameter's value. Adding directional motion signals to the motion energy model also improved performance slightly. The researchers additionally investigated various encoding models that explicitly incorporated color information. Surprisingly, color did not provide substantial improvements in reconstruction, and models that included color did not outperform luminance-only models significantly. However, luminance and color models were effective in reconstructing color borders, such as those between hair and face or face and body.

Such findings provide support for the proposition that fMRI recordings exhibit a stronger representation of shape compared to color in V1, V2, and V3 location[8]. In a study by Taylor and Xu[10], the outcomes indicated heightened sensitivity to orientation changes (shape) rather than color changes in V1, V2, and V3. Conversely, the regions VOT and V4, which exhibited substantial overlap, demonstrated equally robust sensitivity to color and curvature changes but displayed diminished sensitivity to orientation changes. Overall, at the fMRI level, information pertaining to shape predominates over color. Given the greater impact of shape on results, future algorithms should prioritize shape over color during training with fMRI data. By prioritizing shape, algorithms may reproduce images with object structures more closely resembling the stimulus, as shape appears to exert a more pronounced influence on outcomes than color.

While the study generated impressive reconstructions, its model shared no direct analogues with the brain's functioning, and so did not in itself contribute to our understanding of human visual processing. A weakness of the approach is that the model has an inherent knowledge of the correct answer for a given fMRI scan, even without the algorithm's intervention. This issue occurs because the model was trained using fMRI data and its corresponding video clips, trained to output timestamps rather than video. As a result, when presented with an fMRI scan, the model necessarily had already seen the video clip from which it was derived.

Deep Image Reconstruction

The researchers aimed to determine whether they could categorize or reconstruct images from fMRI recordings[3]. They used a multi-layered deep neural network (DNN) approach. A DNN is an artificial neural network built with layers of neurons, also known as interconnected nodes. Each layer in a DNN hierarchically processes information, allowing the network to learn complex patterns and representations from the input data. The key characteristic of a DNN is its depth, meaning it has many hidden layers between the input and output layers. The network is able to learn increasingly abstract features as it processes the data from one hidden layer to another. The process of training a DNN involves repeatedly adjusting the weights and biases of the neurons to reduce the error between the predicted output and the actual output.

The DNN was trained on fMRI recordings of 3 individuals with normal or corrected-to-normal vision over a period of 10 months. The participants consisted of a 33-year-old male, a 23-year-old male, and a 23-year-old female. The participants were shown a series of images and asked to imagine objects based on cue words, while their brain activity was recorded using fMRI.

The strengths of their approach included the use of a deep generative network (DGN) to prioritize the structure in reconstructing natural images, resulting in more realistic reconstructions. A DGN is a combination of a deep neural network and a generative adversarial network (GAN)[3]. A DNN, as mentioned above, is a type of artificial neural network with multiple layers of neurons. Each layer processes information hierarchically, allowing the network to learn complex patterns and representations from the input data, potentially analogous to human visual processing. A GAN is a type of artificial intelligence model that can create new data samples that resemble a given training dataset. GANs consist of two neural networks, a generator, and a discriminator, which are trained together in a competitive process. If we combine these two terms, a "Deep Generative

Network" refers to a neural network architecture that is both deep (having multiple layers) and capable of generating new data samples, using GAN-like structures.

The researchers indeed found similarities between the hierarchical representations in the DNN process and the brain processes, justifying the use of hierarchical structures in their model.

A weakness of their approach was that they only trained the DNN on natural images, limiting its learning of other types of images, such as symbols. They evaluated the performance of their approach by comparing the pixel-wise accuracy, human recognition, color, luminance, and structure of the reconstructed images with the original ones.

The results showed that on natural images, using a DNN without a GAN-like component, the pixel-wise accuracy was marginally higher, though the human recognition was marginally worse (both significant, 76.1% vs. 79.7%, and 97% vs. 96%). For images of human-constructed structures and objects, the pixel-wise accuracy was 70.5%, and for alphabet images, it was 95.6%. Human recognition accuracy for artificial images was 91.0% and for alphabet images, it was 99.6%, demonstrating that the algorithm was able to reproduce alphabet images so well that humans almost always were able to recognize them.

Notably, the researchers found that shape played a more significant role than color in people's ability to recognize objects, a detail reminiscent of the above-established preference of fMRI to represent shape over color.

To test how the "deepness" of their neural network model affects the results, the researchers combined multiple DNN layers and conducted objective and subjective tests. The objective test was pixel-wise accuracy. The subjective test involved having an independent rater be given an original image and a pair of reconstructed images who was then instructed to choose the reconstructed image that looked more similar to the original image. Through their results, they concluded that combining multiple levels of visual feature DNNs decreased pixel accuracy but increased human perceptual reconstruction quality. The researchers also observed that the accuracy of DNN feature decoding strongly influenced the quality of the reconstructed images, where better DNN feature decoding resulted in better reconstructed image quality. DNN feature decoding, also known as feature decoding or feature reconstruction, refers to the process of predicting or reconstructing the DNN features from brain activity data, typically measured using fMRI.

Generic Object Decoding

The researchers aimed to determine whether they could categorize images based on fMRI recordings[6]. They employed 13 candidates of visual feature types/layers developed from four models, including a Convolutional Neural Network (CNN), a Hierarchical Model and X (HMAX)[11], a Global Image Structure Tensor (GIST)[12], and a Scale-invariant Feature Transform with the Bag of Features (SIFT + BoF), to analyze the fMRI data. As mentioned above, CNN is a type of deep neural network designed to process visual data, especially images. The study tested one CNN algorithm with each layer labeled as CNN 1, CNN 2, CNN 3, CNN 4, CNN 5, CNN 6, CNN 7, and CNN 8. HMAX, also known as the Hierarchical Model and X, is a model inspired by the hierarchical organization of the visual cortex in the brain. It consists of multiple stages that mimic different processing levels in the brain's visual system, using Gabor filters and max-pooling to create feature maps at different scales and orientations[11]. The study tested three HMAX models labeled HMAX 1, HMAX 2, and HMAX 3. GIST, on the other hand, is a low-level image descriptor that summarizes the global spatial layout of a scene. It captures the dominant spatial properties, such as rough layout, orientation, and spatial frequency, of an image. GIST is computationally efficient and has been used for scene categorization and image retrieval tasks[12]. SIFT+BoF is a technique commonly used in image recognition and object classification tasks. It combines the Scale-Invariant Feature Transform (SIFT) method, which detects key points and describes them based on local gradients, with the Bag of Features (BoF) approach. BoF creates a histogram of visual word occurrences from SIFT features[6]. Both GIST and the SIFT+BoF only had one model tested. In comparison, because CNNs are able to automatically learn hierarchical features and outperform other methods in many tasks, they have become the dominant approach in computer vision. HMAX was influential as a biologically inspired model, but it has been largely surpassed by CNNs. GIST and SIFT+BoF are useful for certain image processing tasks but are limited compared to the representation power of CNNs.

The study involved 5 participants: 1 female and 4 males. The participants were shown a series of images and asked to imagine objects based on cue words while an fMRI was used to record their brain activity.

The researchers called their new approach based on visual features "generic object decoding." They emulated the approach of the brain, using a hierarchical, top-down approach in an attempt to demonstrate that decoders trained with stimuli can be utilized to decode visual attributes of imagined objects. They found similarities between their approach and the brain's functioning. For example, certain visual feature decoders exhibit a hierarchical structure similar to the human visual system. Others have complex layers that create an output that is statistically similar to visual cortical activity in the brain. However, one weakness of their approach was that 20 out of 1000 categories in the training set were also included in the 50 test set, potentially affecting the results presenting an accuracy rate that is higher than the actual value.

To assess the performance of their approach, the researchers calculated correlation coefficients between the rank of the pre-

dicted object category and the semantic distance for all feature types/layers. On a graph with a correlation coefficient on the y-axis and time from image onset, CNN 8 had the highest coefficient at its peak and CNN 1 had the lowest coefficient. Also, on seen images, they presented a graph showing the percentages of correct categorizations, with CNN 5 achieving the highest accuracy. Similarly, they presented another graph on imagined images, with CNN 6 achieving the highest accuracy. The study found that both seen and imagined objects were recognized with statistical significance during analysis.

Generative Adversarial Network

The researchers utilized a Deep Convolutional Generative Adversarial Network (DCGAN) approach, and the same publicly available dataset as GOD to investigate using the brain activity from fMRI recordings to reconstruct visual stimuli viewed by three human participants[8].

The researchers chose to use a DCGAN approach because they wanted to explore the concept of adversarial training models for image reconstruction. However, unlike the hierarchical models of more recent papers, the authors did not offer any biological motivation for the use of that model. To assess the performance of their approach, the researchers had human subjects evaluate the correctness of the reconstructed images. Given that humans can identify images even with slight blurring, integrating human input during testing becomes crucial for gauging the algorithm's performance. The primary objective of the algorithm is to replicate an image based on fMRI recordings and showcase it to other humans. However, when dealing with images that are highly similar, a potential bias emerges, as varying recognition abilities among individuals may introduce conflicts during the testing process.

The study conducted a behavioral perceptual experiment using Amazon Mechanical Turk as a platform for collecting human labeling and uncomplex behavioral scientific data. This choice was motivated by the advantages of the platform over traditional university subject pools, as highlighted in previous literature. While on the platform, workers were tasked with distinguishing between an original image and a randomly selected reconstruction from the same validation set. Each choice constituted one Human Intelligence Task (HIT) and was compensated with $0.01. To ensure the quality of responses and prevent fraudulent activity and bias, workers needed to possess a Masters status and maintain an approval rate of 95% or higher on the platform to qualify for the tasks. The experiment was repeated ten times for each image in the validation sets, with a different randomly chosen reconstruction paired for each iteration. Measures were in place to prevent workers from seeing the same image-reconstruction pair twice to avoid gaming the task.

When training and testing, the researchers used their DCGAN on two datasets: the BRAINS dataset and a natural images dataset. The BRAINS dataset consisted of 360 images of six handwritten characters: B, R, A, I, N, and S, While, the natural images dataset included square colored images from 150 categories from the ImageNet database. On the BRAINS dataset, the correctness rate determined through human rating was 54%. The reconstructions for natural images were considered to be sufficiently accurate and were seen to have been focusing on preserving contrast differences and structural features. Sufficiently accurate images are those in which the model's ability to reproduce structural elements, including the position and curvature of lines was demonstrated. Some reconstructions lost luminance information, while the presence of horizontal landscape lines was maintained in certain cases, but not consistently.

The study did not involve categorizing the images; instead, they evaluated the accuracy of the reconstruction on a pixel-wise and human-wise basis. For two sets of natural images, natural grayscale images with a circular mask and natural object photos, the human participants were able to recognize a reconstruction of the original image 67.2% and 66.4% of the time, respectively.

Linear Combination

Deviating from the common trend of employing deep neural networks, researchers in this particular study aimed to reconstruct images from fMRI recordings using a linear combination approach, involving multi voxel patterns of fMRI signals and multiscale visual representations[5]. In the study, four human subjects viewed a series of random images consisting of geometric shapes, alphabet letters, and arbitrary figures while an MRI recorded their brain activity, and then the linear combination model attempted to reconstruct the images from the fMRI recording. The study involved four human subjects who participated in random image sessions and figure image sessions with shapes or alphabet letters.

They chose the linear combination approach because conventional retinotopy or the inverse of the receptive field model, the standard approach at the time, was not as effective in predicting local contrast in an image. Conventional retinotopy refers to the mapping of visual stimuli onto specific regions of the retina. The "inverse of the receptive field model" typically involves reversing the process of receptive field mapping, which is a technique used to identify the specific visual stimuli that activate a neuron in the brain. The reason the researchers chose a model that does not necessarily correspond to the brain's processes, was due to the ineffectiveness of the two approaches mentioned above.

To assess the performance of their approach, the researchers measured the accuracy of image recognition. The identification of images was quantified by determining the smallest mean square difference between the fMRI activity pattern and the reconstructed image. The results showed that the approach worked way better than chance, achieving nearly 100% correct identifi-

cation with 100 image candidates and over 10% performance with sets of $10^7.4$ through $10^10.8$ possible images. They also achieved over 95% accuracy with a 3x3 patch area in the foveal region for reconstruction on both random patterns and figure images when a previous study Thirion et al. (2006) only reached between 41% and 71% accuracy in a similar task.

They evaluated the reconstruction accuracy on a pixel-wise and human-wise basis. The research revealed that the multi-scale reconstructed contrast pattern, a pattern that takes into account contrasts of different sizes or resolutions within the image, resulted in the lowest reconstruction error, and the reconstruction based only on single-scale 1x1 image bases resulted in the highest reconstruction error. Additionally, the V1 voxel had the least reconstruction error, while the V3 voxel had the highest reconstruction error. This distinction suggests that higher visual areas, such as V3, contain less reliable information for reconstructing visual images than lower visual areas, such as V1. The increase in eccentricity (degrees from the center of the visual field) led to an increase in reconstruction error, regardless of the reconstructed contrast pattern. Because eccentricity profiles vary based on the scale type, the researchers deduced that multiscale models seem to effectively identify dependable scales at different eccentricities. The decoder was trained on non-shuffled data and the researchers found that the testing performance with shuffled data was significantly lower than with the original non-shuffled data. This difference indicates that the multivoxel pattern decoder efficiently utilizes voxel correlation to attain a strong performance in decoding.

Discussion

In the realm of fMRI imaging, the use of Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Deep Convolutional Generative Adversarial Networks (DCGANs), and linear combinations have emerged as effective techniques for fMRI analysis. Each study approaches the feasibility of reconstructing and categorizing images from fMRI data through their unique ideas and has achieved its goal. However, these endeavors, while promising, fall short of consistent success in a few key aspects. The majority of reconstructed images frequently demonstrate imperfections, and the categorization results often lack accuracy.

The ability to reconstruct visual images from fMRI data holds immense promise in revealing the neural basis of human perception. By capturing the brain's response to external visual stimuli, image reconstruction allows us to infer the patterns of neural activity associated with specific images. This provides a direct window into the inner workings of the visual system, shedding light on how the brain processes and represents visual information. Such insights have broad implications, ranging from advancing our understanding of basic perceptual mechanisms to deciphering the foundations of mental imagery and memory recall.

The simpler problem of categorizing images based on fMRI recordings offers a unique perspective on the brain's ability to classify and differentiate visual stimuli. This approach involves training machine learning algorithms to associate specific brain activity patterns with different object categories, providing a glimpse into the neural mechanisms of object recognition. The knowledge gained from image categorization contributes to our understanding of how the brain organizes and processes complex visual information.

Beyond Vision

All the above studies focused only on decoding visual information. In the future, researchers could delve into uncharted territories such as detecting human emotions and mental states, as well as deciphering human thoughts and even translating abstract concepts like ideas or words. Notably, the distinction between reconstructing images and unveiling thoughts underscores the intriguing potential and intricate challenges that lie ahead. Reconstructing images involves extracting visual information from brain activity and focuses on identifying the patterns of neural activity associated with the perception of the visual stimuli. This neural activity has been well-studied for the past century and researchers know where it is located: in visual cortices. The challenge is to establish a direct link between specific brain activity patterns and the corresponding visual content, which will benefit fields like neuroimaging, augmented reality, and assistive technologies. Unveiling thoughts, on the other hand, delves into deeper layers of neural activity spread throughout the entire cortex. Worse, stimuli like emotions, intentions, memories, or even linguistic content are not as easily standardizable as visual input. Unlike decoding images, unveiling thoughts requires interpreting complex neural patterns associated with higher cognitive functions. However, the rewards will justify the difficulty: being able to successfully recreate thoughts and advance neurocognitive research and brain-computer interfaces could potentially enable communication with individuals with limited speech capabilities.

The applications of fMRI-based image reconstruction and categorization extend beyond research into clinical practice. Neurologists and psychiatrists could potentially employ these techniques to develop objective and accurate diagnostic tools for a range of brain disorders. By analyzing the unique neural signatures associated with specific conditions, clinicians could make more informed and personalized treatment decisions [13]. Moreover, real-time image reconstruction may enable developments in mental therapy, allowing patients to actively participate in reshaping their brain activity for therapeutic purposes [14].

AI can learn from fMRI decoding

The fusion of fMRI-derived image data with artificial intelligence (AI) and machine learning algorithms can also have a transformative impact on the development of intelligent systems. By training AI models on neural activity patterns associated with image categories, we can create AI systems that simulate human-like perception and categorization abilities. Such systems could enhance image recognition, autonomous navigation, and human-computer interaction, ultimately leading to the development of more sophisticated AI applications.

Applications of AI to Psychological Tools

The development of psychology has expanded the scope and objects of research, creating favorable conditions for the rapid integration of AI into the field. This integration has led to the development of various AI products, including facial expression-based emotion recognition systems, intelligent medical image grading, and suicide early warning systems. These AI applications contribute to the advancement of machine learning with fMRI analysis in psychology research, enhancing efficiency and shortening research cycles [15].

Affective computing, an interdisciplinary field, combines computational science with physiology, psychology, and cognitive science. It explores emotions in human-human and human-computer interactions, guiding the design of AI systems with emotion recognition and feedback capabilities. This enables machines to understand human emotional expressions and appropriately respond, fostering emotional interactions between humans and computers [16].

In 2018, a novel automatic depression detection algorithm integrated speech and facial expression analysis. Signal enhancement was applied to depressed speech, and features such as fundamental frequency, resonance peaks, energy, and Mel-Frequency Cepstral Coefficients (MFCC) were extracted. The algorithm, combining speech and facial expression recognition models, achieved a recognition rate of 81.14% using the Adaboost algorithm based on backpropagation neural networks. This demonstrates the potential of AI in providing objective diagnostic criteria for psychological health, particularly in depression diagnosis [17].

The realm of music emotion involves understanding the functional connections between sensory, emotional, and cognitive brain areas. Musical emotions, regulated by limbic and paralimbic brain structures, have implications for AI development. Further research in music generation, education, and medical treatment is anticipated. Dapeng Li and Xiaoguang Liu have integrated incremental music teaching methods into therapy, combining contextual teaching and AI attention theory to create a targeted assisted treatment system. This approach considers patients' individual factors to enhance brain neuron activity and access pathological information, thereby promoting autoimmunity and subsequent treatment [18].

Decoding with more than only fMRI

One possibility to improve neural activity decoding lies in the combination of fMRI data with other neural data streams, a venture that could unveil deeper layers of cognitive processes. The convergence of fMRI data with Electroencephalography (EEG) data emerges as a particularly intriguing prospect. Unlike fMRI, EEG has a high temporal resolution, at the cost of much lower spatial resolution, The integration of EEG's temporal precision with advanced machine learning algorithms could lead to real-time insights into the brain's intricate perceptual mechanisms. By decoding brain responses as they unfold, researchers could uncover the real-time dynamics of neural perception, potentially shedding light on the swift transformations that govern our understanding of the visual world.

The combination of functional magnetic resonance imaging and electroencephalography in studies has already proven particularly effective in areas such as epilepsy, sleep research, and neurofeedback studies [19].

The localization of epileptic generators is a crucial aspect of understanding and treating epilepsy. Identifying the precise region of epileptic foci presents a challenge, but simultaneous EEG-fMRI has emerged as a powerful tool to address this difficulty. Electrical brain activity associated with epilepsy, such as interictal epileptiform discharges (IEDs/spikes) or seizures, can be measured using scalp EEG. These discharges lead to increased metabolism and blood flow, which can be detected by the Blood Oxygen Level Dependent (BOLD) signal in fMRI. Analyzing the time course of these discharges from EEG data enables the identification of brain regions generating the IEDs, offering insights into the localized and sometimes widespread responses associated with the epileptic activity.

EEG-fMRI has been instrumental in revealing specific patterns of synchronized decreased brain activity during sleep. Studies have shown that a distinct network is associated with regulating sleep vigilance levels and arousability, reflecting sleep instability. Additionally, EEG-fMRI has identified active brain regions during sleep spindles in non-rapid eye movement sleep. These findings contribute to a deeper understanding of the neural mechanisms underlying different sleep states.

In the realm of neurofeedback, which opens new therapeutic possibilities in psychiatry and neurology, simultaneous real-time fMRI and EEG neurofeedback has proven beneficial. This approach provides patients with real-time feedback from both modalities, allowing for the simultaneous regulation of both haemodynamic and electrophysiological brain activities. By encouraging patients to learn self-regulation of disordered brain regions, this technique holds promise for therapeutic interventions in disorders affecting brain function.

Clinical Applications

A study focused on the development of intelligent systems utilizing machine learning algorithms to bridge communication between humans and data, particularly targeting patients with conditions such as completely locked-in syndrome (CLIS), brain damage, or motor neuron disease (ALS)[20]. The researchers used an Automated Sensor and Signal Processing Selection (ASPS) approach for feature extraction from EEG signals, aiming to identify the most suitable Sensory Characteristic Features (SCFs) to detect human thoughts and imaginations. Artificial Neural Networks (ANN) were also employed to validate the results.

Their findings demonstrated the efficacy of EEG signals in capturing imaginative information for communication purposes. The ASPS approach is highlighted as a powerful feature extraction method, capable of recognizing various imaginations. Unlike previous studies that primarily used highly sensitive features for detecting machinery faults, this research emphasizes a combination of high and low sensitive features, reducing system cost by utilizing three sensors and a limited number of features. The study is designed as a bespoke experiment for individual subjects, with successful testing on two subjects from training by one of the subjects' data, suggesting the potential for generalization to a broader population.

The ASPS approach incorporates time-domain and frequency-domain (FFT) analyses, integrating multiple signal processing techniques to reduce experimental work, time, and cost. The verification using two types of ANN (LVQ and FFNN) attains accuracy rates between 80% and 100% for recognizing five imaginations, with success rates ranging from 87.5% to 100% for four imaginations. Notably, 100% accuracy is achieved when recognizing three imaginations, suggesting the potential for enabling complex communication commands through EEG signals. The study concludes that recording EEG signals during a combination of relaxation and mental tasks allows the extraction of sensitive delta features, and the ASPS approach efficiently identifies unique and sensitive features that can significantly impact Brain-Computer Interface (BCI) systems for classifying brain signals. This advancement holds promise for enabling intricate communication of feelings and needs through thought.

Dangers and Difficulties of AI

AI algorithms, while holding immense potential, are susceptible to various shortcomings, posing challenges to their widespread applicability[21]. One significant issue is the limited generalizability of AI systems, hindering their reliability beyond the training domain and impeding clinical applicability for most medical data types. Examples include an algorithm mistaking wolves for dogs based on the background (snow or grass) in images. In healthcare, similar issues arise, such as an algorithm classifying a skin lesion as malignant due to the presence of a ruler in the image, rather than the lesion's characteristics. Features like surgical skin markings and urgent scan labels have also been found to influence algorithmic predictions, emphasizing the need to understand the features learned by neural networks for generalization across diverse healthcare settings.

Achieving robust generalization is particularly challenging due to technical disparities between different sites, encompassing variations in equipment, coding definitions, electronic health record (EHR) systems, and laboratory equipment and assays. Additionally, local clinical and administrative practices introduce further complexities.

The brittleness of AI models is another concern, with the potential for blind spots that may lead to suboptimal or erroneous decisions. These blind spots can stem from technical differences between sites and may result in the model being easily fooled or making poor decisions. This lack of robustness poses a significant challenge to the deployment of AI algorithms in real-world clinical settings.

Moreover, the issue of biases in machine learning models raises ethical concerns. Blind spots in AI systems can reflect societal biases, and there is a risk of unintended or unknown inaccuracies, particularly in minority subgroups. The historical biases present in the training data can be inadvertently amplified by the AI model, potentially leading to disparities in healthcare outcomes. Discriminatory bias is intertwined with the generalizability challenge, as blind spots in ML can reflect societal biases, leading to unintended inaccuracies in minority subgroups. Studies show that current AI systems can disproportionately affect disadvantaged groups based on race, gender, and socioeconomic background. For instance, mortality prediction algorithms may exhibit varying accuracy by ethnicity, and algorithms for classifying skin lesions may underperform on patients with skin of color due to biased training datasets.

Algorithmic unfairness is categorized into model bias, model variance, and outcome noise. Model bias results from models favoring majority groups, model variance stems from inadequate data from minorities, and outcome noise is influenced by unobserved variables interacting with model predictions. Addressing these issues requires a heightened awareness among clinicians, empowering them to critically participate in system design and development. ML algorithms should be designed with a global perspective, and clinical validation should involve representative populations. Comprehensive performance analysis, including subgroup analyses based on age, ethnicity, sex, sociodemographic stratum, and location, is essential. Prospective pilots within healthcare systems are recommended to understand product characteristics and identify potential pitfalls before practical deployment. Careful evaluation is crucial, especially when the AI system detects a different spectrum of diseases than current clinical practice, necessitating an assessment of benefits and harms.

Another critical challenge is the limited availability of healthcare data for machine learning. Access to diverse and representative datasets is essential for training AI algorithms effectively. However, many healthcare datasets are not readily accessible, hindering the development of robust and generalizable models.

Addressing these challenges requires careful consideration of technical, ethical, and data-related aspects to ensure the responsible and effective deployment of AI in healthcare. Efforts to improve data sharing, mitigate biases, and enhance the generalizability of AI models are crucial for unlocking the full potential of artificial intelligence in medical applications.

References

1 *What is Machine Learning? Defination, Types, Applications, and more. Great Learning Blog: Free Resources what Matters to shape your Career! Published January 4*, `https://www.mygreatlearning.com/blog/what-is-machine-learning/`, Accessed August 18, 2023.

2 *Published March*, **28**, year.

3 G. Shen, T. Horikawa, K. Majima and Y. Kamitani, *PLOS Computational Biology*, **15**, year.

4 J. Tang, A. LeBel, S. Jain and A. Huth, *Nat Neurosci*, **26**, 858–866.

5 Y. Miyawaki, H. Uchida and O. Yamashita, *Neuron*, **60**, 915–929.

6 T. Horikawa and Y. Kamitani, *Nat Commun*, **8**, year.

7 S. Nishimoto, A. Vu, T. Naselaris, Y. Benjamini, B. Yu and J. Gallant, *Curr Biol*, **21**, 1641–1646.

8 K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk and M. Gerven, *NeuroImage*, **181**, 775–785.

9 J. Brownlee, *MachineLearningMastery.com. Published November*, **7**, year.

10 J. Taylor and Y. Xu, *Representation of Color, Form, and their Conjunction across the Human Ventral Visual Pathway*, Published online August 31,.

11 M. Riesenhuber and T. Poggio, *Nat Neurosci*, **2**, 1019–1025.

12 A. Jain, A. Phanishayee, J. Mars, L. Tang and P. G. Gist, 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA). IEEE, p. 776–789.

13 J. Baker, D. Dillon and L. Patrick, *Proceedings of the National Academy of Sciences*, **116**, 9050–9059.

14 B. Fredborg, K. Champagne-Jorgensen, A. Desroches and S. Smith, *Consciousness and Cognition*, **87**, year.

15 M. Tahan, *Neuropsychopharmacol Hung*, **21**, 119–126.

16 J. Zhao, M. Wu, L. Zhou, X. Wang and J. Jia, *Frontiers in Neuroscience*, **16**, year.

17 J. Zhao, M. Zhang, C. He and K. Zuo, *Frontiers in Psychology*, **10**, year.

18 D. Li and X. Liu, *Occup Ther Int*, **2022**, year.

19 T. Warbrick, *Sensors*, **22**, year.

20 S. Majumdar, A. Al-Habaibeh, A. Omurtag, B. Shakmak and M. Asrar, *Neuroscience Informatics*, **3**, year.

21 C. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado and D. King, *BMC Medicine*, **17**, year.

Appendix

Since CNNs are among the most broadly applicable and widely used of the methods discussed here, I will end my exploration with a comprehensive overview of CNNs.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of deep learning algorithm designed to process and analyze visual data, such as images and videos [3]. They are inspired by the structure and functioning of the human visual cortex. A key feature of CNNs is their ability to automatically learn and extract hierarchical patterns and features from input images, enabling them to recognize and classify objects, scenes, and patterns with high accuracy.

CNNs consist of several layers, including convolutional, pooling, and fully connected layers. In the convolutional layers, small filters or kernels are applied to the input image, convolving across the entire image to detect relevant features, edges, and textures. This process helps CNNs identify unique patterns and shapes that are crucial for accurate image recognition.

Pooling layers are used to reduce the size of the feature maps obtained from the convolutional layers. This downsampling process helps reduce the computational complexity of the network and prevents overfitting by retaining only the most essential information. The fully connected layers are then used to interpret the extracted features and make predictions based on the available classes or categories. The CNN's ability to learn complex patterns and hierarchies of features allows it to excel in tasks such as object detection, image classification, image generation, and image segmentation.

CNNs are inherently translation invariant, meaning that they can recognize patterns and objects regardless of their position in the image. This property allows CNNs to identify objects even if they are shifted, making them ideal for tasks like object detection and image classification. CNNs can also be pre-trained on large-scale image datasets such as those from ImageNet, learning general features that apply to various visual recognition tasks. These pre-trained models can then be adjusted or transferred to specific tasks with limited labeled data, enabling faster and more efficient training. Additionally, for tasks involving large images or video streams, CNNs are more memory-efficient compared to fully connected neural networks. CNNs focus on local feature extraction and do not require the same level of memory for fully connected layers as other architectures.

One notable application of CNNs is in image reconstruction from fMRI data. This concept involves translating brain activity captured through fMRI scans into meaningful visual representations. The process entails training a CNN to predict and generate images based on the provided fMRI signals. The idea is to 'decode' brain activity patterns into recognizable images. CNNs are suited for this task due to their hierarchical feature extraction capabilities. They can learn intricate patterns and features from training images and then apply this knowledge to reconstruct visual content from fMRI data. The network architecture's ability to capture complex relationships between different image elements is leveraged in this process.

CNNs are also proficient at image categorization, a task where they classify images into predefined categories or labels. This process involves training the network to recognize specific patterns of brain activity associated with different objects or scenes. CNNs excel in image categorization due to their ability to learn distinctive features from images during the training phase. These features enable the network to differentiate between various objects or scenes and make accurate categorization predictions.